Further information is included in your conference pack:
- Lunch map
The bus stop for the banquet (Sodoh) and social tour venues (Kyoto University Library & Kyoto University Museum) are also there. Regardless of whether you join the tour or not, all conference participants will be admitted as free visitors to Kyoto University Museum during the conference.
- Kyoto International Manga Museum info
1 invitation ticket to the Kyoto International Manga Museum (Opening hours: 10:00 – 18:00, admission until 17:30. Closed on Wednesdays) and its flyer.
- Wi-Fi info
We provide ID and password for each participants. For more information, see the instruction in your conference bag. If you have an eduroam account, eduroam hotspot is also available.

Welcome to iPRES 2017

As the lead organizing co-chair, I am delighted to host and welcome delegates to Kyoto for the 14th International Conference on Digital Preservation (iPRES) on September 25-29, 2017, under the theme of “Keeping Cultural Diversity for the Future in the Digital Space – from Pop-Culture to Scholarly Information.”

Kyoto is the former capital of Japan with more than thousand years’ of history and possesses rich resources in terms of traditional culture and pop-culture. It also has strong relationships with Asian countries through common research and business links. In keeping with previous years, iPRES2017 aims to contribute to the promotion of research and digital preservation practices for the global community. In particular, iPRES2017 focuses on pop-culture and Asian communities. Workshops and tutorials will provide opportunities for participants to share information, knowledge and best practices, and explore opportunities for collaboration.

Last but not least, I hope iPRES2017 will bear fruit in terms of future research. On behalf of the organizing committee, I express my sincere thanks and wishes to the organizers and participants of the Conference. I also hope you will take the opportunity to enjoy Kyoto and carry back good memories of your stay here.

Shoichiro HARA,
Co-Chair (Lead) of iPRES2017 Organizing Committee
Prof. and Deputy Director, CSEAS, Kyoto University, Japan

Conference Organization
Organizing Committee
Co-Chair(Lead): Shoichiro HARA
Prof. and Deputy Director, CSEAS, Kyoto University, Japan
Co-Chair: Shigeo SUGIMOTO
Prof. Faculty of Library, Information and Media Studies, University of Tsukuba, Japan
Chair: Makoto GOTO
Assoc. Prof., National Museum of Japanese History, Japan

Program Committee
Program Co-Chairs:
- Klaus REICHERTD (Freiburg University, Germany)
- Pang NATALIE (Ass. Prof., Nanyang Technological University, Singapore)
- Umair KARADKARA (Ass. Prof., University of Texas at Austin, USA)
- Tatsuki SEKINOPROF., The Research Institute for Humanity and Nature, Japan

Japanese Program:
- Taizo YAMADA
Assoc. Prof., Historiographical Institute, University of Tokyo, Japan
- Makoto GOTO
Assoc., Prof. National Museum of Japanese History, Japan

Local Committee:
Chair:
- Akiko KAMEDA
Assoc. Prof., CSEAS, Kyoto University, Japan

Members:
- Mikiko ONO
Assoc. Prof., CSEAS, Kyoto University, Japan
- Eriko AMANOJURA
Kyoto University Research Administration Office, Kyoto University, Japan
- Toshiro KAMIYA (URA, Kyoto University Research Administration Office, Kyoto University, Japan)
- Kazuko NISHISupporting Staff, CSEAS, Kyoto University, Japan
- Kichi MINO
- Akiko SAITO
Shizuka IRE, Hiroko MATSUDA, Junko KAWASHIMA, Junko NAKASHI, Koyo NAKAMURA, Kazuto TOMODA, Kumiko KAWAMIYA, Kyoko KAMATA, Makoto TAKAHASHI Masako AKEDO, Motoaki KONO, Naoko MAEN, Norio SAITO, Tomoko KAWAI, Tsubasa OHSA, Yuzuru INAGAMI, Yuka ITO, Yumiko TSCHIBURA

Publications Committee
Co-Chairs:
- Taiko NAKAYAMA (Ass. Prof., CSEAS, Kyoto University, Japan)
- Mario LOPEZ
Assoc. Prof., CSEAS, Kyoto University, Japan

As the lead organizing co-chair, I am delighted to host and welcome delegates to Kyoto for the 14th International Conference on Digital Preservation (iPRES) on September 25-29, 2017, under the theme of “Keeping Cultural Diversity for the Future in the Digital Space – from Pop-Culture to Scholarly Information.”

Kyoto is the former capital of Japan with more than thousand years’ of history and possesses rich resources in terms of traditional culture and pop-culture. It also has strong relationships with Asian countries through common research and business links. In keeping with previous years, iPRES2017 aims to contribute to the promotion of research and digital preservation practices for the global community. In particular, iPRES2017 focuses on pop-culture and Asian communities. Workshops and tutorials will provide opportunities for participants to share information, knowledge and best practices, and explore opportunities for collaboration.

Last but not least, I hope iPRES2017 will bear fruit in terms of future research. On behalf of the organizing committee, I express my sincere thanks and wishes to the organizers and participants of the Conference. I also hope you will take the opportunity to enjoy Kyoto and carry back good memories of your stay here.

Shoichiro HARA,
Co-Chair (Lead) of iPRES2017 Organizing Committee
Prof. and Deputy Director, CSEAS, Kyoto University, Japan

Conference Organization
Organizing Committee
Co-Chair(Lead): Shoichiro HARA
Prof. and Deputy Director, CSEAS, Kyoto University, Japan
Co-Chair: Shigeo SUGIMOTO
Prof. Faculty of Library, Information and Media Studies, University of Tsukuba, Japan
Chair: Makoto GOTO
Assoc. Prof., National Museum of Japanese History, Japan

Program Committee
Program Co-Chairs:
- Klaus REICHERTD (Freiburg University, Germany)
- Pang NATALIE (Ass. Prof., Nanyang Technological University, Singapore)
- Umair KARADKARA (Ass. Prof., University of Texas at Austin, USA)
- Tatsuki SEKINOPROF., The Research Institute for Humanity and Nature, Japan

Japanese Program:
- Taizo YAMADA
Assoc. Prof., Historiographical Institute, University of Tokyo, Japan
- Makoto GOTO
Assoc., Prof. National Museum of Japanese History, Japan

Local Committee:
Chair:
- Akiko KAMEDA
Assoc. Prof., CSEAS, Kyoto University, Japan

Members:
- Mikiko ONO
Assoc. Prof., CSEAS, Kyoto University, Japan
- Eriko AMANOJURA
Kyoto University Research Administration Office, Kyoto University, Japan
- Toshiro KAMIYA (URA, Kyoto University Research Administration Office, Kyoto University, Japan)
- Kazuko NISHISupporting Staff, CSEAS, Kyoto University, Japan
- Kichi MINO
- Akiko SAITO
Shizuka IRE, Hiroko MATSUDA, Junko KAWASHIMA, Junko NAKASHI, Koyo NAKAMURA, Kazuto TOMODA, Kumiko KAWAMIYA, Kyoko KAMATA, Makoto TAKAHASHI Masako AKEDO, Motoaki KONO, Naoko MAEN, Norio SAITO, Tomoko KAWAI, Tsubasa OHSA, Yuzuru INAGAMI, Yuka ITO, Yumiko TSCHIBURA

Publications Committee
Co-Chairs:
- Taiko NAKAYAMA (Ass. Prof., CSEAS, Kyoto University, Japan)
- Mario LOPEZ
Assoc. Prof., CSEAS, Kyoto University, Japan

As the lead organizing co-chair, I am delighted to host and welcome delegates to Kyoto for the 14th International Conference on Digital Preservation (iPRES) on September 25-29, 2017, under the theme of “Keeping Cultural Diversity for the Future in the Digital Space – from Pop-Culture to Scholarly Information.”

Kyoto is the former capital of Japan with more than thousand years’ of history and possesses rich resources in terms of traditional culture and pop-culture. It also has strong relationships with Asian countries through common research and business links. In keeping with previous years, iPRES2017 aims to contribute to the promotion of research and digital preservation practices for the global community. In particular, iPRES2017 focuses on pop-culture and Asian communities. Workshops and tutorials will provide opportunities for participants to share information, knowledge and best practices, and explore opportunities for collaboration.

Last but not least, I hope iPRES2017 will bear fruit in terms of future research. On behalf of the organizing committee, I express my sincere thanks and wishes to the organizers and participants of the Conference. I also hope you will take the opportunity to enjoy Kyoto and carry back good memories of your stay here.

Shoichiro HARA,
Co-Chair (Lead) of iPRES2017 Organizing Committee
Prof. and Deputy Director, CSEAS, Kyoto University, Japan
13:00-14:00 Main Hall  
日本語語学講座1  
杉本誠男

ディジタルリソースの長期保存に関する概要

杉本誠男

ディジタルリソースの長期保存は困難であるが取り組まなければならない問題として広く理解されている。ここではディジタルリソースの長期保存に関する基本的な理解を得ることを目的として、ディジタルリソースの長期保存の考え方、ディジタルリソースの長期保存の標準モデルであるOAI-PMH(Archive Information System)等を紹介し、技術的な面からディジタルリソースの長期保存を視察する。

13:00-18:00 Conference Room 1F  
METS Editorial Board Annual Face-to-Face Meeting  
Betty POST and Tom HABING

METSの構文と使用方法についてのプレゼンテーション

13:00-14:00 Main Hall  
日本語語学講座2  
柳田明世

社会調査用語データの収集、保存、二次分析について

柳田明世

情報素材論文掲載者にとらえ社会調査は1940年代のアメリカ合衆国で始め、その後国からデータの共有、保存、二次分析は研究者集団にとって重要な課題であった。このニュートリシティでは、データの保存と二次分析という観点から、社会調査の歴史を基盤した、日本の現状を東京大学

15:10-16:40 Main Hall  
デジタルリソースの長期保存に関する講演

杉本誠男

15:10-16:40 Pre-conference Workshop of Asian Session 2nd Workshop on Academic Asset Preservation and Sharing in Southeast Asia  
Organizer: Shoichiro HARA

Pre-conference Workshop of Asian Session  
2nd Workshop on Academic Asset Preservation and Sharing in Southeast Asia

Organizer: Shoichiro HARA

This is a session to make arrangements for the Asian Session on 25th. It is planned as a semi-closed session, but observers are welcomed.
## 26 (TUE)

### Main Hall

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 10:00  | 10:00-12:00 Digital Curation of Historical and Cultural Resources in Japan(2)  
                     戦歴資料デジタル記録として何を記録すべきか——日本とアジアと世界——  
Source: National Museum of Japanese History |
| 11:00  | Organizer: Makoto GOTO (National Museum of Japanese History)  
Presenter: John ERTL, Yoshiko SHIMADZU, Shigeki MORO |
| 11:30  | Lunch Break |
| 12:00  | 13:00-13:30 Opening |
| 13:30  | 13:30-14:30 Keynote (1)  
Ingrid DILLO (Data Archiving and Networking Services, Netherlands) |
| 14:00  | FAIR Data in Trustworthy Data Repositories  
Ingrid DILLO (Data Archiving and Networking Services, Netherlands) |
| 14:30  | 14:40-16:40 Asian Session  
Reports and Discussion  
Moderator: Natalie PANG |
| 15:00  | 16:00  
16:30  
17:00  
17:30  
18:00  
18:30  
19:00  |
| 10:00-12:00 | Main Hall (simultaneous translation service available) |
| 14:40-16:40 | Main Hall |

**Asian Session**  
For Papers and Discussion  
Moderator: Natalie PANG (National University of Singapore)  
Invited Speakers: Shujii KAMITSUNA (National Diet Library, Japan), Sophy Shu-Juin CHEN (Academia Sinica, Taiwan), Lee Kee Sing (National Library Board, Singapore), Wararak PATTANAKIJ (Chulalongkorn University, Thailand), Chito ANGELES (University of the Philippines Diliman, Philippines)  
Countries in the Asia-Pacific region are very diverse in terms of culture and language environments. Long-term management, keeping and use of digital resources is a common concern for many of these countries which are producing more and more digital resources. However, reports on digital preservation activities are lacking, especially from the countries in East and South-East Asia. This session is aimed to share up-to-date information about developments in digital archives and digital preservation in East and South-East Asia and to discuss issues on digital preservation in this region with the audience from other parts of the world.  
This session will present five talks by invited speakers from Japan, Taiwan, Philippines, Thailand and Singapore about digital preservation in these countries and/or countries. We will then invite voluntary reports from other Asian countries followed by general discussions with the audience.

### Foyer

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30</td>
<td>Poster Display</td>
</tr>
<tr>
<td>11:00</td>
<td>Lunch Break</td>
</tr>
</tbody>
</table>
| 12:00  | 12:30  
13:00  
13:30  
14:00  
14:30  
15:00  
15:30  
16:00  
16:30  
17:00  
17:30  
18:00  
18:30  
19:00  |
| 14:40-16:40 | Main Hall |
| 16:40-17:30 | Main Hall |
| 17:30-19:00 | Welcome Reception and Poster Session |

**Keynote (1)**  
Ingrid DILLO (Data Archiving and Networking Services, Netherlands)  
National and international funders are increasingly likely to mandate open data and data management policies that call for the long-term storage and accessibility of data. Open data and data sharing can only become a success if we put the concept of trust central stage. The certification of digital repositories is an important means to provide this trust to the different stakeholders involved. In this keynote I will talk about data sharing, repository certification and the concept of FAIR data.
### Main Hall

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30</td>
<td>Registration/Welcome</td>
</tr>
<tr>
<td>8:50</td>
<td>Metadata &amp; Linked Data</td>
</tr>
<tr>
<td>9:00</td>
<td>Metadata &amp; Linked Data</td>
</tr>
<tr>
<td>9:30</td>
<td>Metadata &amp; Linked Data</td>
</tr>
<tr>
<td>10:00</td>
<td>8:50-10:20 Metadata &amp; Linked Data</td>
</tr>
<tr>
<td>10:30</td>
<td>9:00-10:00 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>10:50</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>11:00</td>
<td>9:30-10:00 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>11:30</td>
<td>9:00-10:00 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>12:00</td>
<td>10:00-10:30 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>12:30</td>
<td>10:30-11:00 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>13:00</td>
<td>10:30-11:00 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>13:30</td>
<td>11:00-11:30 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>14:00</td>
<td>11:30-12:00 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>14:30</td>
<td>12:00-12:30 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>15:00</td>
<td>12:30-13:00 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>15:30</td>
<td>13:00-13:30 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>16:00</td>
<td>13:30-14:00 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>16:30</td>
<td>14:00-14:30 Emulation &amp; Software Preservation</td>
</tr>
<tr>
<td>17:00</td>
<td>14:30-15:00 Emulation &amp; Software Preservation</td>
</tr>
</tbody>
</table>

### Conference Room 5F

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30</td>
<td>Keynote (2)</td>
</tr>
<tr>
<td>8:50</td>
<td>Digital Dunhuang: A Standard for Digital Preservation</td>
</tr>
<tr>
<td>9:00</td>
<td>Digital Dunhuang: A Standard for Digital Preservation</td>
</tr>
<tr>
<td>9:30</td>
<td>Digital Dunhuang: A Standard for Digital Preservation</td>
</tr>
<tr>
<td>10:00</td>
<td>Digital Dunhuang: A Standard for Digital Preservation</td>
</tr>
<tr>
<td>10:30</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>11:00</td>
<td>PDF Preservation</td>
</tr>
<tr>
<td>11:30</td>
<td>PDF Preservation</td>
</tr>
<tr>
<td>12:00</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>12:30</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>13:00</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>13:30</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>14:00</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>14:30</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>15:00</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>15:30</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>16:00</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>16:30</td>
<td>Lunch Break</td>
</tr>
<tr>
<td>17:00</td>
<td>Lunch Break</td>
</tr>
</tbody>
</table>

### Social Events

- 16:00-17:00 Social Events
- 17:00-18:00 Banquet venue opens, 18:00- Banquet
<table>
<thead>
<tr>
<th>Time</th>
<th>Main Hall</th>
<th>Conference Room 5F</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8:50-9:50</td>
<td>Preservation Systems (1)</td>
<td></td>
</tr>
<tr>
<td>9:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:50-10:20</td>
<td>Content Analysis</td>
<td>Operational Pragmatism in Digital Preservation: establishing context-aware minimum viable baselines</td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:00-11:30</td>
<td>Preservation Systems (2)</td>
<td></td>
</tr>
<tr>
<td>11:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:00-14:00</td>
<td>Keynote (3)</td>
<td>Established Context-aware Minimum Viable Baselines Organiser: Somaya LANGLEY</td>
</tr>
<tr>
<td>14:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:10-15:10</td>
<td>Closing</td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Main Hall

8:50-9:50 **Preservation Systems (1)**
- Angela DAPPERT and Adam FARQUHAR: Permanent Record Persistent Digital Preservation: A Roadmap
- Bertrand CARON, Jordan De La ROUSSAYE, Thomas LEDOUX, and Stéphane BEECHT: Life and Death of an Information Package: Implementing the Lifecycle in a Multipurpose Preservation System

9:50-10:20 **Content Analysis**
- Naoto YOSHIDA and Shin-ichi TADAKI: Semi-automated Generation of Linked Data from Unstructured Bibliographic Data for Japanese Historical Rare Books (S)
- Christopher LEE and Kam WOODS: Diverse Digital Collections Meet Diverse Users: Applying Natural Language Processing to Born-Digital Primary Sources (S)

9:50-12:00 **Operational Pragmatism in Digital Preservation**
Establishing context-aware minimum viable baselines
Panel: Somaya LANGLEY, Anthea SELES, Andrew HAYNE, Dinah KATIE, Jones LUKOSE, and Bertrand CARON.
Undertaking active digital preservation, holistically and thoroughly, requires substantial infrastructure and resources. National archives and libraries across the Western world have established, or are working towards maturity in digital preservation (often underpinned by legislative requirements). On the other hand, smaller archives and companies situated outside of memory institution contexts, as well as organisations in non-Western and developing countries, are struggling with the basics of managing their digital materials. This panel continues the debate within the digital preservation community, critiquing the development of digital preservation practices typically from within positions of privilege. Bringing together individuals from diverse backgrounds, the aim is to establish a variety of "bare minimum" baselines for digital preservation efforts, while tailoring these to local contexts.

10:30-11:30 **Preservation Systems (2)**
- Emi ZIESEA: OAIS and Distributed Digital Preservation in Practice
- Helen RICKS-YU: Superb Stewardship of Digital Assets: Developing a strategy for Digital Archiving and Preservation at the University of Notre Dame

11:30-12:00 **Certification**
- Barbara SIEMAN and Kees WATERMAN: How the Dutch prepared for certification

13:00-14:00 **Endeavors of Digital Game Preservation in Japan: A Case of Ritsumeikan Game Archive Project**
Akinori NAKAMURA (College of Image Arts and Sciences, Ritsumeikan University, Japan)
In 1998, one of the first academic institutions, which focuses on pursuing the appropriate ways of the video game preservation, the Game Archive Project has been established. The sentiment toward the digital games being products rather than “viable cultural artifcats”, the efforts have started slowly. In our continuous efforts to enlighten academic as well as the professional community for this cause, however, the importance of the preservation activities has been embraced in both domestic and international communities, leading to being a part of the national project for the creating of the Media Art Database by the Agency of Cultural Affairs, Japan. The present paper attempt to introduce an overview of our efforts.

14:10-15:10 **Closing**
### 28 (THU) evening

<table>
<thead>
<tr>
<th>Time</th>
<th>Main Hall</th>
<th>Conference Room 5F</th>
<th>Meeting Room</th>
<th>Conference Room 1F</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:30</td>
<td>15:30-17:00 Pop-Culture Session</td>
<td></td>
<td></td>
<td>15:30-17:00 WDS Collaborative Session</td>
</tr>
<tr>
<td></td>
<td>Archiving and Utilization of Japanese Pop Cultures Materials</td>
<td></td>
<td></td>
<td>Research Lifecycle based Research Data Management Requirements and its Alignment with Institutional, Domestic and International Contexts</td>
</tr>
<tr>
<td></td>
<td>- Kaichiro MORIKAWA (School of Global Japanese Studies, Meiji University): &quot;Issues in archiving manga, anime, games and related cultures&quot;</td>
<td></td>
<td></td>
<td>Chair: Shoji KAJITA, Speakers: Shoji KAJITA, Kazudsuna YAMAJI, Heidi J. IMKER</td>
</tr>
<tr>
<td></td>
<td>- Tadahiro SAIKA (International Manga Research Center, Kyoto Seika University / Kyoto International Manga Museum): &quot;Archiving and Utilization of Manga Materials: in the case of Kyoto International Manga Museum&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>15:30-18:30 Workshop 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Digital Preservation Workshop: Exploring Preservation Storage Criteria and Distributed Digital Preservation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Andrea GOETHALS, Nancy McGOVERN, Jane MANDELBAUM, Sibyl SCHAEPER, Gail TRUMAN and Eld ZIERAU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td>15:30-18:30 Workshop 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A Dutch approach in constructing a network of nationwide facilities for digital preservation together</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joost van der NAT, Marcol RAJ and Eefke SMIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:00</td>
<td>15:30-17:00 WDS Collaborative Session</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:30</td>
<td>15:30-17:00 WDS Collaborative Session</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:00</td>
<td>15:30-17:00 WDS Collaborative Session</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td>Workshops and tutorials can have further instructions online. See iPRES website for more information, <a href="https://ipres2017.jp/programme/">https://ipres2017.jp/programme/</a></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 29 (FRI)

<table>
<thead>
<tr>
<th>Time</th>
<th>Main Hall</th>
<th>Conference Room 5F</th>
<th>Meeting Room</th>
<th>Conference Room 1F</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00</td>
<td>9:00-12:00 Tutorial 1</td>
<td></td>
<td></td>
<td>9:00-12:00 Workshop 4</td>
</tr>
<tr>
<td>9:30</td>
<td>9:00-12:00 Tutorial 1</td>
<td></td>
<td></td>
<td>Innovative approach for project viability: from a diversity of business models to harmonized and scalable national services Pierre-Yves BURGI, Eliane BLUMER and André JELICIC</td>
</tr>
<tr>
<td>10:00</td>
<td>9:00-12:00 Tutorial 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td>9:00-12:00 Tutorial 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:00</td>
<td>9:00-12:00 Tutorial 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:30</td>
<td>9:00-12:00 Tutorial 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00</td>
<td>Lunch Break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:30</td>
<td>13:00-16:00 Tutorial 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:00</td>
<td>13:00-16:00 Tutorial 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:30</td>
<td>13:00-16:00 Tutorial 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>13:00-16:00 Tutorial 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:30</td>
<td>13:00-16:00 Tutorial 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td>13:00-16:00 Tutorial 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:30</td>
<td>13:00-16:00 Tutorial 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>13:00-16:00 Tutorial 5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Abstracts

Keynote Speeches
Full and Short Papers
Posters and Demos
Workshops
Tutorials
Organized Sessions
Posters/Demos Lightning Talk
Abstracts

Keynote Speeches
FAIR Data in Trustworthy Data Repositories

Ingrid Dillo
Data Archiving and Networked Services (DANS)
The Netherlands
ingrid.dillo@dans.knaw.nl

KEYWORDS
Certification, data sharing, FAIR, CoreTrustSeal, open data, repository, Data Seal of Approval, World Data System

1 OPEN DATA AND DATA SHARING
Research funding in recent years often comes with the condition to make some of the resulting data openly available. Most researchers appreciate the benefits of sharing research data, but on an individual basis they may be reluctant to share their own data. Why are some researchers hesitant to share? And what are the most important motivations of researchers who do share their data?1

2 THE CONCEPT OF TRUST AND REPOSITORY CERTIFICATION
National and international funders are increasingly likely to mandate open data and data management policies that call for the long-term storage and accessibility of data. Open data and data sharing can only become a success if we put the concept of trust central stage. The certification of digital repositories is an important means to provide this trust to the different stakeholders involved.

If we want to be able to share data, we need to store them in a trustworthy digital repository. Data created and used by scientists should be managed, curated, and archived in such a way to preserve the initial investment in collecting them. Researchers must be certain that data held in archives remain useful and meaningful into the future. Funding authorities increasingly require continued access to data produced by the projects they fund, and have made this an important element in Data Management Plans. Indeed, some funders now stipulate that the data they fund must be deposited in a trustworthy repository. Sustainability of repositories raises a number of challenging issues in different areas: organizational, technical, financial, legal, etc. Certification can be an important contribution to ensuring the reliability and durability of digital repositories and hence the potential for sharing data over a long period of time. By becoming certified, repositories can demonstrate to both their users and their funders that an independent authority has evaluated them and endorsed their trustworthiness.

Within the tiered framework of certification standards that has developed over the last decade core level certification has been embraced by a large number of repositories around the globe. Within the framework of the Research Data Alliance the ICSU World Data System (WDS) and the Data Seal of Approval (DSA) have developed a unified catalogue of requirements.2 The group built on inherent complementarity between the criteria previously established by the two organizations to harmonize unified and universal requirements reflecting the core characteristics of trustworthy data repositories. The first new CoreTrustSeals (CTS) have already been acquired.

We would like to mention that the DMM presents several advantages with respect to OOMMF for calculating the spectrum of magnetic eigenmodes for the following reasons: a) There is no need to excite the system by any magnetic field pulse, b) A single calculation allows to determine the frequencies and eigenvectors of all spin-wave modes of any symmetry, c) The spectrum is computed directly in the frequency domain, d) The mode degeneracy is successfully solved, e) The spatial profiles of the spin-wave modes are directly determined as eigenvectors and, finally, f) The differential scattering cross-section can be calculated accurately from the eigenvectors associated to each spin-wave mode. This is a clear indication that both the Py and Co sub-elements are in a single domain state where Py and Co magnetizations are all oriented with their magnetic moment along the chain and field direction. At point $\beta (H = 372$ Oe) of the hysteresis loop, where the plateau is observed in the $M-H$ loop, the dark and bright spots of the Py dots are reversed with respect to those of Co, accounting for an antiparallel relative alignment of magnetization.

3 FAIR DATA IN TRUSTWORTHY DIGITALREPOSITORIES
The condition to make data resulting from publicly funded research openly available has the effect that more and more data are rapidly becoming available. Therefore, there also is a growing demand for quality criteria for research datasets. The CTS

requirements and the FAIR principles get as close as possible to giving quality criteria for research data. They do not do this by trying to make value judgements about the content of datasets, but rather by qualifying the fitness for data reuse in an impartial and measurable way.

In 2014 the FAIR Guiding Principles (Findable, Accessible, Interoperable and Reusable) were formulated. The well-chosen FAIR acronym is highly attractive. In a relatively short term, the FAIR data principles have been adopted by many stakeholder groups, including research funders.

The FAIR principles are remarkably similar to the underlying principles of DSA (2005): the data can be found on the Internet, are accessible (clear rights and licenses), in a usable format, reliable and are identified in a unique and persistent way so that they can be referred to. Essentially, CTS presents quality criteria for digital repositories, whereas FAIR targets individual datasets.

Bringing the ideas of the CTS and FAIR together offers an operationalization that can be implemented in any certified Trustworthy Digital Repository.

---

3 https://www.force11.org/group/fairgroup/fairprinciples
Digital Dunhuang: A Standard for Digital Preservation

Peter Zhou
University of California, Berkeley

ABSTRACT
The Digital Dunhuang project is enabling long-term preservation of cultural heritage of inestimable value, while providing a platform for sharing all digital assets generated in the act of preservation. This presentation will examine the major aspects of digital asset management and digital preservation implemented in this ongoing project.

KEYWORDS
Digital preservation, digital asset management, Digital Dunhuang, the preservation of world heritage

1 INTRODUCTION
The Mogao Caves at Dunhuang, a designated UNESCO World Heritage site, are a splendid treasure house of art from Ancient China. [1–2] The Dunhuang Academy has been devoted to the protection of the Mogao Caves since its founding in 1944 (originally called the Dunhuang Art Research Institute). The protection measures include slowing down the degeneration of murals with a variety of protecting techniques, as well as looking for digital tools to preserve the caves since 1990s. The Academy is currently undertaking the Digital Dunhuang project with cooperation from home and abroad. Work is being done to digitize three-dimensional structures of murals, sculptures and all the 492 caves, and also to construct a digital database of the Mogao Caves resources. The size and scope of the project is quite significant. Some exemplary statistics include:[3]

- Over 100 caves already photographed
- Over 14,000 square meters of murals captured
- More than 70 terabytes of data
- Over 87 caves in complete QuickTime Virtual Reality
- Over 20 years of climate monitoring data for 87 caves

The project has called on the expertise of a number of institutions over the years, including The Dunhuang Academy, Zhejiang University, Wuhan University, the Getty Conservation Institute, and Northwestern University. It is designed to incorporate a wide range of data types, including archeological and conservation data and files, documentation for a large number of artifacts, Dunhuang manuscripts, and scholarly publications.

2 FROM DIGITAL ASSET MANAGEMENT TO DIGITAL PRESERVATION
With the support of The Mellon Foundation, the Dunhuang Academy has been exploring building a permanent repository of all digital assets of Digital Dunhuang. It has also made persistent efforts to photograph the caves while conducting archeological study and conservation work. The Academy has made huge progress and accumulated large amounts of data in both analog and digital formats. The consensus seems to be: if the information gathered from the Mogao Caves is to be permanently preserved for future generations, the only way to ensure this is to integrate all the content that has been created in the past, is being created now, and will be created in the future into a large digital repository. [4] This digital repository can facilitate perpetual preservation, effective digital asset management operations, and easy access in a systematic way through the use of high technology delivery mechanisms. This presentation discusses major aspects of the design and development of this repository system. The ultimate goal is to store the totality as well as minutia of the Dunhuang Caves’ digital content, preserve those digital surrogates, reconstructions, files, and digital assets perpetually, and provide a research platform for the study of mural art, history, geography, religion, economics, politics, ethnology, language, literature, science and technology in ancient China and Central Asia.
There is a sense of urgency. More than half a million visitors come here every year. The historical and cultural significance of the caves, as well as their vulnerability— their age, instability of the terrain, and fluctuations of humidity caused by increased human presence—all add to the sense of urgency. The digital capture and restoration of Dunhuang will most likely be the last records of today’s Dunhuang.

2.1 Functional Components [5]

DAM: Facilitates asset creation, cataloging; image, video, and text management and delivery and version control. Tracks digital preservation actions. Pushes metadata and content to the Digital Dunhuang platform. Manages master high resolution files and original documents.

DIGITAL PRESERVATION: Managed digital preservation actions include creating checksums, validating files, and extracting technical metadata upon ingest; monitoring file format obsolescence; migrating file formats as well as content; tracking and copying files to LTO tapes.

PUBLISHING AND SHARING: Surrogates of master assets managed in the DAM will be pushed for external delivery.

2.2 Content Categories

Stitched/composite cave images, raw cave images, cave QTVRs, historical photos, videos, digital restorations, manuscripts from Cave 17, artifacts (approximately 10,000 objects), reproductions (copies) of images in caves, (digitized) microfilm of manuscripts, interactive panoramic of caves; research created by members of the Dunhuang Academy, scholarly publications; previously published bibliographies, indices, glossaries, and finding aids; conservation data and materials, archaeological reports and drawings, CAD drawings, and 3D laser point cloud data.

2.3 File Formats

TIFF, JPEG, JPEG2000 (still image), PSD, BMP, PSB (Photoshop large file format), CR2 (Canon raw format), DCR (Kodak raw format), DNG (Adobe/universal raw format), other RAW camera formats (list), CDR (Corel Draw), CAD, PTX (original 3D cloud points), DGN (Microstation Design File), PDF, CAJ, MOV (QTVR), MPEG2/35 Mbps (AVI wrapper), HD video files (format TBD), DPG (Nintenndo video file format), Word, Excel, txt, MPEG4, H.264, FLV (Flash).

2.4 Cross-linking

All content (images, documents, etc.) must be displayed with metadata. For example, an initial search result might display thumbnails with basic data. The user can then select to see a large 3D image and fuller data.

3 RESULTS AND DISCUSSION

Digital Dunhuang is an enterprise project. It enables us to explore ways to design and develop sophisticated databases, and perpetually preserve massive amounts of data. By extension, Digital Dunhuang has also shed light on what we need to do to preserve digital assets in libraries and museums around the globe.

ACKNOWLEDGMENTS

The work discussed was sponsored by the Dunhuang Academy, with the support from the Mellon Foundation. This presentation draws on a joint report by Linda Tadic and the author on the functional requirements of Digital Dunhuang in 2012; and on a panel discussion at the DH 2015 conference in Sydney, at which the author discussed this project with other panelists - Wang Xudong and Wu Jian of the Dunhuang Academy, and J. Stephen Downie of the University of Illinois. Their contributions are greatly acknowledged.
REFERENCES


ABSTRACT

In 1998, one of the first academic institutions, which focuses on pursuing the appropriate ways of the video game preservation, the Game Archive Project has been established. The sentiment toward the digital games being products rather than “viable cultural artifact”, the efforts have started slowly. In our continuous efforts to enlighten academic as well as the professional community for this cause, however, the importance of the preservation activities has been embraced in both domestic and international communities, leading to being a part of the national project for the creating of the Media Art Database D by the Agency of Cultural Affairs, Japan. The present paper attempt to introduce an overview of our efforts.

KEYWORDS
Digital Game Preservation, Database-Design, Emulation

1 ORIGIN
Ritsumeikan Game Archive Project (or GAP in abbreviation, henceforth, GAP), was established in April 1998. Our endeavours can be categorized into three forms of preservations, which are namely, 1) physical preservation, accompanying with digital data archive which registered information sufficient to identify the preserved materials, 2) preservation in the form of emulation, and finally 3) digital moving images of the people actually playing the preserved object[2]. Thus, the next phase of our institutions was to examine each of mentioned endeavors

2 GENERAL RESULTS FROM EACH PRESERVATION EFFORT

2.1 Physical Preservation

Physical Preservation was initiated simultaneously with the establishment of GAP in 1998. We visited several studios to initiate the actual preservation of game titles. One of the studios, SEGA understood our intention and started donating products published by SEGA from 1998 to the end of 2003 when the major corporate re-organization took place. Other than SEGA, various game titles from multiple platforms had been donated regularly, slowly expanding the number of collections at Game Archive Project.

2.2 Preservation through the Development of Emulator

Development of Emulation Box (hence forth “the box”) for the Famicom initiated in 2002 with the realization that physical objects would decay in the long run. The overall design of the device is shown in the Figure 1. The box was designed to emulate precisely as the Famicom, it would behave both in digital and the analogue as the game data run from a generic computer. The system composed of a server, the game data transfer system, and emulator of the Famicom’s memory management unit which is compatible to over 10 types of Read Only Memory architectures as well as the Famicom Disk System architecture, which allow the device to be compatible to all of the titles released for the Famicom. Only original components used for the device is used was the game
controller interface as this analogue portion of the device is impossible to emulate and thus the device was connected to the actual controller so that software can be operated precisely as they were originally intended. For the testing purpose, two titles, namely, Donkey Kong and Mario Brothers were stored in the generic computer and play tested with the device. The limitation, however, also became apparent as the legal process was found to be a lot more complex for what we anticipated, forcing GAP from pursuing the option for the video game preservation efforts.

### 2.3 Preservation through the Video Image

![Figure 2: Simultaneous Recording of Game Play Images and Pressing of the Input Interface](image)

Regarding video games as cultural artifacts, recording the visual image of the gameplay naturally became a vital part in the preservation activities. Our approaches, however, were not merely recording the playing images, but also simultaneously recording the timing of pressing of buttons for each game play so that spectators will be able to view how they were playing and also precisely how they used the interface upon playing. This was possible by showing both playing images and signal patterns from the pressing of the buttons as shown in the Figure 2. The studies further revealed that there were differences in the play styles between novice players and those of avid players, further showing the importance of the preserving the game play images in this manner.

### 3 Expansion of the Efforts

#### 3.1 Developing Metadata of Digital Game

These efforts lead GAP to work under the Agency of Cultural Affairs for constructing a digital game section of Media Arts Database since April 2012. The data have been cross-referenced before the registration, allowing other researchers to use these data as a source of reference. Number of game titles which have been registered in the database reached 44683 titles by March 2017[3].

The team has been working on modifying the former systems to create with entity-relationship model, considering a special property that enables archivists for cataloging and designing of metadata for the database. The scrutinizing the data allow us on examining the differences in reception of Japanese titles in Japan and the west[4] as well as the changing on the naming patterns of game titles[5].

#### 3.2 Expansion of Physical Preservation

As our efforts has lead to extend our network not only to academic institutions, but also to libraries and museum of which the video game has been objects of collections and exhibitions, making us further realize the importance of preserving physical objects not only for the Japanese community but also for collaborating with the international community of the video game preservations. By March 2017, there are 6211 as shown in the Table 1

<table>
<thead>
<tr>
<th>Table 1: Number of the Titles Collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>PlayStation</td>
</tr>
<tr>
<td>SEGA Saturn</td>
</tr>
<tr>
<td>Super Famicom</td>
</tr>
<tr>
<td>PC Engine</td>
</tr>
<tr>
<td>Family Computer</td>
</tr>
<tr>
<td>PlayStation 2</td>
</tr>
<tr>
<td>Dream Cast</td>
</tr>
<tr>
<td>PlayStation Portable</td>
</tr>
<tr>
<td>Nintendo DS</td>
</tr>
<tr>
<td>SEGA Mega Drive</td>
</tr>
<tr>
<td>Gameboy Advanced</td>
</tr>
<tr>
<td>Nintendo GameCube</td>
</tr>
<tr>
<td>Game Boy</td>
</tr>
<tr>
<td>Others</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

### 4 Future Perspectives of the Game Archive Project

As the video game preservations has been increasingly acknowledged by the public, the demand for further security on this subject has become ever more important. For recent years, GAP has been extending our network both domestically as well as internationally with hope of an organizing international consortium of digital game preservation. By uniting the endeavors together with other institutions around the globe, we will strive to continue in our efforts to preserve this important cultural artifact in our era.

#### REFERENCES


[2] 細井浩一. 2006. 「情報機器技術のための映像情報の特徴と活用に向けて」. 立命館大学 COEニュースレター ISSUE.04. DOI: http://www.arc.ritsumei.ac.jp/art_coe/al_4_03.html. The succeeding section, namely, 2.1-2.3 is a summary of the same article.


Abstracts

Full and Short Papers
Modeling the Domain of Digital Preservation in Wikidata

Katherine Thornton∗
Euan Cochrane
katherine.thornton@yale.edu
euan.cochrane@yale.edu
Yale University

Thomas Ledoux
Bertrand Caron
Bibliothèque nationale de France
Paris, France

Carl Wilson
Open Preservation Foundation
UK

ABSTRACT
Members of the digital preservation community collate and capture metadata to describe file formats, software, operating systems and hardware, and use it to inform and drive digital preservation processes. In this work we describe how the infrastructure of Wikidata meets the requirements for a technical registry of metadata related to computer software and computing environments. Collaboratively creating this metadata, and making it available as linked open data, will reduce the amount of redundant work digital preservation professionals do in order to describe resources. Having machine-readable, linked open data that describes the digital preservation domain will also allow us to reuse this data in our software applications and information systems, reducing the overhead when building new tools. Furthermore the Wikidata social and technical infrastructure will enable the long term continued access to the data digital preservation practitioners collate and capture.

Wikidata is a project of the Wikimedia Foundation (WMF), and is created through commons-based peer production [3]. Simply put, Wikidata is a knowledge base of structured data that anyone can edit [40]. The infrastructure of Wikidata is created using free software, and is designated to the public domain. All content in Wikidata is licensed so that others may freely reuse the data. Volunteer editors, coordinating their own work, add data to Wikidata. Through this analysis we demonstrate how the infrastructure of Wikidata provides distinct advantages to the cultural heritage domain that proprietary knowledge bases do not provide.

ACM Reference format:
DOI: 10.475/123_4

1 INTRODUCTION
Metadata about software, file formats and computing environments is necessary for the identification and management of these entities. Machine-readable metadata about software, file formats and computing environments allows digital preservation practitioners to then automate programmatic interactions with these entities.

∗The corresponding author

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

†iPres’17, Kyoto, Japan
© 2017 Copyright held by the owner/author(s). 123-4567-24-567/08/06...$15.00
DOI: 10.475/123_4

Figure 1: A screenshot of the Wikidata item for TIFF version 6.0 and related properties.

For example, if we ingest a file into the a digital preservation system, we could configure the system to use a tool such as DROID [2] or Sigfreid [20] to automatically identify the file formats of resources at the time of ingest into the digital preservation system. Based on that data, we could then present the user with a list of available software in a collection that could be used to interact with files of that type. This is an example of machine-actionable interaction that is possible with a digital preservation system using DROID in combination with Wikidata data to determine the software options.

A technical registry [6] is a data store of descriptions of file formats, software used to create or interact with files, configured hardware environments, operating systems and sustainability factors. In Figure 1, there is a screenshot of a Wikidata item and the properties used to describe the item. The metadata for file formats in Wikidata includes information about who developed the format, how it is related to other formats, what the file extensions are, and the media type(s).

2 FRAGMENTED AND INCOMPLETE
Many researchers and practitioners in the field of digital preservation have identified the fragmented nature of technical registries [21]. When registry data is stored in multiple systems, communication also takes place in distributed settings. In response to the fragmented landscape, several groups have presented plans for how
to centralize and unify technical registry information [22, 37, 38]. Both CDFR and UDFR have concluded work. The sustainability of a technical registry is important set of factors to consider. The creation of a sustainability plan for the infrastructure of the registry is a crucial component to consider. The solution proposed by the New Zealand team is forthcoming, and infrastructural development described as the subsequent phase of work [22]. In this paper, we explore an alternative to creating infrastructure for a repository of technical registry data through use of existing, independently supported, infrastructure provided by the knowledge base of structured data, Wikidata.

Wikidata went live in late 2012 [39]. The infrastructure of Wikidata is collaboratively built via commons-based peer production [3, 4, 25]. Commons-based peer production is the name given to open collaboration systems where users are creating content that will become part of the public domain. This means that all of the work products of the community are free to be reused by others. The peer production aspect refers to how users coordinate work amongst themselves. Wikidata is edited by volunteers from all over the world in more than 350 languages [14].

The MediaWiki software [1] and WikiBase software [41] are the primary technical components of the knowledge base itself. In the domain of computational systems, the concept of infrastructure is used to describe technologies that support information systems. Theorists of infrastructure, Star and Ruhleder, note that infrastructure is often invisible, and because of this, many people take it for granted [35]. By referring to infrastructure as ‘invisible’ these authors highlight the fact that infrastructure is often purposefully designed to be available only to those who are building or repairing it. For example, the infrastructure of the search algorithms used by Google are not made visible to users of the search engine. The database structure of Amazon.com is not made visible to visitors of the website. In contrast, Wikidata’s infrastructure is open for inspection because the technical components of the system are free software [33] and the source code for the software is shared publicly. Using Wikidata as the technical registry of metadata for the domain of digital preservation equates to using the infrastructure of the Wikidata system to store digital preservation metadata.

3 Wikidata: A Knowledge Base of Structured Data

The knowledge base of structured data, Wikidata, combines a data model with structured data. Editors add content and provide source information for structured data [17]. Wikidata contains data about entities structured in a way that is machine-readable as well as human-readable [14]. As the data management platform of all Wikimedia Foundation projects, the data is free and open for reuse within all WMF projects, and also is freely available for reuse outside of WMF projects [14].

In Figure 2, we see a screenshot of the Wikidata page for entity Q42332 Portable Document Format. Each item is allotted a page in Wikidata and has a unique identifier, with prefix Q plus a string of numbers, ex. Q42332, which is assigned to the item Portable Document Format. Wikidata consists of two entity types: items and properties. In Figure 2 the four properties Commons category, BNCF Thesaurus, image, and topic’s main category are used to assert statements about the item Q114678. In Figure 2 we see four statements about Portable Document Format. Each of these statements expressed a property of the item Q114678. Each item contains a list of claims in the form of triples. The subject of the triple is the Wikidata item to which the claim refers, the predicate is a Wikidata property, and the object is a date, string, quantity, URL, an external identifier or another Wikidata item. Claims can be made more precise through the use of qualifiers. These qualifiers indicate the contexts in which the claim is valid. Claims can be annotated through the inclusion of references. A claim and its references are considered to be a statement. A description of the Wikidata data model can be found in [42].

The knowledge base, Wikidata, contains many millions of such items, many of these items are already described by a set of statements. The Wikidata pages for items allow users to view and enter data [14]. These pages are open for editing by anonymous IP, or by registering for an account as an editor. Volunteer editors contribute to the project in coordination with one another and all content in the system is designated as belonging within the public domain. The content in Wikidata spans general and specialized domains, and is relevant for many application areas [14].

3.1 Collaboratively creating data models

One way that the structure of Wikidata is extended is through the creation of properties to represent specific relationships between items. We participated in efforts to extend Wikidata to more precisely represent information from the domain of digital preservation. We identified pages on the Wikidata wiki where the community was actively describing computing. We found a active group engaged in Wikidata’s WikiProject Informatics 2. On the pages of this wiki project we communicated with other Wikidatans to discuss the models for representation of file formats, software, computing environments and hardware environments. We created items and proposed properties to describe different aspects of computing. We gave feedback on ideas proposed by other editors, and we gathered feedback about our ideas from other editors. Much of the discussion within WikiProject Informatics is based around the creation of new items and new properties for Wikidata. Items can be created by any editor of Wikidata. Editors use the create new item link that is displayed on the sidebar of every Wikidata page, and the newly-created item will be added to the knowledge base. Additional steps are involved in the creation of new properties. Properties must first be proposed using a specific template. Once a property proposal template is applied to the property proposals page, a discussion is opened and other editors are invited to comment on the proposal. Property proposal discussions stay open for a minimum of seven days. Editors vote in support of or opposition to the property, and once there is sufficient support an editor with the Property Creator privilege will create the new property and it will become part of Wikidata.

An overview of the properties that are currently in use in Wiki-data with relevance for software is provided in Table 1.

---

1Each of these properties has a identifier assigned to it. The property identifiers all begin with prefix P plus a string of numbers.
2https://www.wikidata.org/wiki/Wikidata_WikiProject_Informatics
3https://www.wikidata.org/wiki/Special:NewItem
4https://www.wikidata.org/wiki/Wikidata:Property_creators
The most common property used to connect software items to file format items is the use of the properties P1072 readable file format and P1073 writable file format. This allows the relationship between the software and the file format to be machine-readable. Using structured data from Wikidata in an computational system for digital preservation allows us to use this data in a way that helps the computer program determine what the next step in a workflow (such as presenting a list of software choices to a user based on the user’s file format) should be. Statements using P1072 and P1073 are made on the items related to software. In Table 2 the properties related to file format items are listed.

These tables list only a subset of relevant properties. For more exhaustive lists of useful properties, examples of how they are used, please see the wiki project\(^5\). There are many properties that we need to be able to extend the data model for the domain of digital preservation. For example, if we consider how the PREMIS project has modeled their semantic unit environment, we can see that this concept is not yet represented as a Wikidata property \([12]\). People who have expertise in the metadata issues and data modeling issues for the domain of digital preservation who also participate in Wikidata are in a strong position to advocate for data models in Wikidata that are compatible with previous models of the domain such as PREMIS \([11]\) and PRONOM.

Data modeling in Wikidata happens through the use of properties and the use of class-subclass relationships \([5, 32]\). Models for domains within Wikidata are being extended as the project matures. For example, the modeling of the molecular biology domain takes place within Wikidata’s WikiProject for molecular biology \([14, 21]\). Collaboratively modeling the domain requires interaction and communication among participants. Muller-Birn et al. found that the Wikidata interface is a barrier for contributors who’d like to

\(^5\)https://www.wikidata.org/wiki/WikidataWikiProject_Informatics
engage in conceptual modeling, but that it supports instance-level curation more adequately [25].

We participated in collective data modeling activities through WikiProject Informatics6. WikiProjects are sets of wiki pages that groups use to perform focused, collaborative work [44]. One avenue through which we participated in the Wikidata community is the property proposal process. The property proposal process involves using the property proposal template to describe what you would like to add to Wikidata. Figure 3 is a screenshot of a property proposal template from Wikidata.

The template provides the outline structure for the proposal. To complete property proposal, an editor must think through the planned data type, how this property will be combined with other entities in Wikidata, and some examples for others to review.

Once a property has been proposed, other Wikidata editors consider the proposal and provide comments or ask questions about the proposal. The discussion portion of this process helps ensure that multiple editors beyond the editor who created the property proposal reviews the proposal. Reviewers are looking out for redundant properties, duplicate properties, properties that would conflict with other properties, and ensure that all parts are included (such as a formatter URL if the proposed property is an external id). As of March, 2017, there are 3228 properties available for use in Wikidata7. Dozens of properties are created each week8.

Focused engagement with the property proposal workflow in Wikidata allows participants to take active roles in the data modeling activities. The negotiation that occurs among editors through the property proposal workflow helps ensure that properties are designed efficiently.

3.2 Our Wikidata work to date

Members of our team have been proposing properties for external IDs from other databases. For example, we proposed and saw created, P2748 PRONOM file format identifier, P2749 PRONOM software identifier, and P3266 LoC FDD ID. These three properties each have a data type of external id, the property is constructed using the base URL for the external resource. For example, the formatter URL for P3266 LoC FDD ID is shown in Figure 4 9.

We also worked to use these properties to create claims around items representing file formats. By undertaking data curation of the Wikidata entities within the domain of digital preservation, we discovered ways to use properties to model relationships between file formats and other entities. If we are not yet able to express a certain relationship, we make note of properties that do not yet exist which we would like to propose. For example, we are considering proposing a property to express the concept "part of configured software environment" to bring together sets of software used together for a purpose, such as presenting a particular emulated computing environment via a framework such as Emulation as a Service10. A property such as "part of configured software environment" could then be used to connect items for a base operating system, software that has been installed, etc. The data could then be re-used in a system where that information is displayed to a user who needs information about what functionality can be expected from a particular configured environment.

Understanding the data model is very important when attempting to get data out of Wikidata. There are several options for how to get data out of Wikidata. It is possible to get structured data on a per item basis in a selection of data formats11. It is also possible to make a copy of a dump of the data. Another option is to request data through Wikidata's application programming interface12 (API). It is also possible to request data from the Wikidata Query Service, a SPARQL endpoint. The name SPARQL is a recursive acronym for SPARQL Protocol and RDF Query Language [13]. SPARQL queries RDF triple stores in order to identify and return sets of triples that meet the criteria specified in the query. The fact that Wikidata maintains a SPARQL endpoint allows for powerful, flexible queries to be written to get data out of Wikidata [8, 16, 23]. In the following section we will provide some examples of how the way the data is modeled has implications for how we can construct queries to get data out of Wikidata.

4 OPEN MODELING CHALLENGES

Through participation in the collaborative data modeling activities in Wikidata, we have observed several open questions that are relevant for the community in general. We consider these to be open questions because we have observed different segments of the community responding to these challenges using a variety of strategies, but agreement about how to harmonize the strategies is not yet resolved. We will illustrate these discussions with examples from Wikidata and examples of queries written for the Wikidata Query Service.

4.1 The Bonny and Clyde problem

One issue that arises when trying to use Wikidata as a format registry comes from its inception: originally the Wikidata entities were derived from Wikipedia pages. At the beginning of the Wikidata project there existed simple matches between pages on Wikipedias and Wikidata. Wikidata served as a hub to allow the alignment between the different language versions of a given page. Nearly all Wikipedea articles in all language versions have associated Wikidata items [14]. The notability criteria 13 indicate that this match (having at least one valid sitelink) as the major acceptability factor for the existence of a item. However, from a format registry perspective, some notions need to be separated. The main examples come from the different compression algorithms (gzip, bzip2, etc.) for which there is usually a single Wikipedia page describing not only the file format, but also the algorithm and a reference implementation. This leads to a single entry with multiple meanings. As an example, the bzip2 entity (Q273563) was described 14 in 2014 as an instance of file format and free software, and properly linked to 19 different sitelinks. Currently only two possibilities exist to bypass this limitation: (1) either create new entities to reflect the different natures. In the previous example, this implies adding the

---

6 https://www.wikidata.org/wiki/Wikidata_WikiProject_Informatics
7 https://www.wikidata.org/wiki/Special:ListProperties
8 https://www.wikidata.org/wiki/Wikidata:List
9 https://www.wikidata.org/wiki/Wikidata:List_of_properties_visit
10 For a complete list of Wikidata properties visit: https://www.wikidata.org/wiki/Wikidata:List_of_properties
11 http://bw-fla.uni-freiburg.de/
"bzip2 Archive" (Q27866052) entity linked to the software entity by "readable/writable file format" (P1072 and P1073) properties. But this breaks the direct access to the Wikipedia pages describing the file format and goes against the notability criteria which may lead to a future removal of the entity; or (2) take advantage of the multicultural aspect of Wikipedia and find one variation that has already separated the concepts to initiate the right model. An example of such an occurrence is the GZIP program where the Portuguese Wikipedia has a specific GZ file format page leading to the differentiation of the tool (Q283647) and the file format (Q10287816). There remains the issue of generalizing such distinction in the different languages but the "Article Placeholder" plugin might solve this [18]. In any case, in order to extend the model and to be able to describe more precise information, the issue of enhancing the relationships between the entity and the interlinks requires a solution. This relationship is important because the Wikipedia pages are the main way to expose the structured information, primarily through the use of information boxes. The Wikipedia page is often the best way to gather corrections and additions from Wikipedians. Even though no satisfying solution exists currently, the problem is well known as the "Bonnie and Clyde issue" [15] and various attempts are made which should lead to an option that will meet global agreement.

4.2 Properties refined by qualifiers

Properties are used in Wikidata to represent some attribute of an item to form a statement, as depicted in Figure 2. Certain types of statements can be expressed by different properties in combination with a one or more qualifiers. In Figure 5 we see a screenshot of some of the claims for item Q20950365, Wikidata Query Service. We can see that Property 31 instance of is used to claim that Wikidata Query Service is an instance of a SPARQL endpoint.
how the claims have been structured, the use of qualifiers to refine properties is a source of flexibility in how data can be modeled in Wikidata.

Some members of the Wikidata community hold the opinion that P31 should not be used as a qualifier and have stated their rationales on the discussion page for the property. In situations where different subsets of the community decided to model data in different ways there is often a period of time where both strategies are observed and then a larger discussion is held to determine if one strategy is privileged over another. The discussions take place in the property talk namespace and are archived alongside the property proposals so that they may be revisited or reconsidered at a later point, if needed.

This flexibility in how data can be modeled impacts how data can be queried from Wikidata. For example, a query for all items that are described as P31 instances of Q2621192 SPARQL endpoint will return a single result, that of item Q20950365 Wikidata Query Service. In contrast, if we structure the query differently to ask for all items where P31 instances of Q2621192 SPARQL endpoint is used as a qualifier, we get 22 results.

A third option for how to structure the query is a request for any item in which there is a P31 statement in a predicate position, thus returning all 23 results. While it is the case that this third query allows us to gather all 23 results, it highlights the necessity of deep familiarity with the data model, and how different portions of the community are working, in order to write a SPARQL query that will not miss relevant data.

The current state of modeling in Wikidata is that many editors are working independently and people are experimenting with many different options for how domains might be modeled. We have noticed a need for additional tools that could promote awareness among editors of how sub-communities (often affiliated through WikiProjects) are currently working with data models for their domains. An example of such a tool is the navigation box shown in Figure 7. This navigation box was created by members of WikiProject Source Metadata to provide an overview of the properties most relevant to the work of the sub-community. This navigation box provides a glanceable display that anyone wanting to quick gain familiarity with the subset of properties that will be of use while curating data related to bibliographic metadata. Rather than scroll through the list of nearly 4,000 properties currently in use, users can quickly find a short list presented via the navigation box. Many WikiProjects use tables or navigation boxes to present relevant properties, however not all users visit WikiProjects regularly and may not encounter these lists if they are not already familiar with work practices of wikis.

Another promising strategy to address consistent data modeling for a given domain is the creation of a portal on top of Wikidata, such as that of the WikiGenomes project. The WikiGenomes portal provides the relevant set of Wikidata entities so that basic researchers who would like to engage in data curation can interact with Wikidata via the portal interface. The strategy of creating a domain-specific portal is one approach to increasing the consistency of how data is being structured as the portal system design can guide users to structure data according to the current set of best practices as understood by the developers of the WikiGenomes portal.

The open modeling challenges, such as the “Bonny and Clyde” problem, involve mappings between Wikipoias and Wikidata, the divergence in the use of properties or properties mixed with qualifiers, and the ongoing data modeling to extend Wikidata properties in additional domains will require creative solutions. Certain sub-communities are addressing these issues through the development of domain-specific portals which may encourage more consistency in how statements are constructed, at least within the domain. Those interested in contributing to discussions of how else we might address these open challenges are invited to join the Wikidata community.

Contributing to a project in which not only the content (structured data), but also the enabling software (MediaWiki and WikiBase), are open to inspection, revision, or extension. The fact that free software is used to enable this commons-based peer production system means that the infrastructure of Wikidata itself is subject to discussion and improvement. Some members of the Wikidata community actively extend the infrastructure by creating Wikidata-specific software tools that can be used to interact with the system.

4.3 Querying the knowledge base

Mechanisms to get data out of Wikidata are an important component of the infrastructure of Wikidata. The Wikidata Query Service is not only a SPARQL endpoint, but also a set of tools that can generate different visualizations of the data such as timelines, bubblecharts, and maps.

Wikidata uses a distinct data type for properties for external identifiers to other systems or repositories. Data types indicate the kind of values that are appropriate to use in combination with the property. As external identifiers are added to Wikidata, we can begin to see a crosswalk of authorities emerge within the knowledge base. For example, in the domain of digital preservation many people are familiar with PRONOM IDs, known as PUIDs, for
file formats and software. The Library of Congress in the United States also publishes identifiers for file formats, known as Format Description Documents (FDDs). Editors of the “Just Solve the File Formats Problem” wiki24 have created wiki pages for many file formats, and these can be uniquely identified by their URLs. Editors have added statements using properties that have been created for PUIDs, FDDs, and File Formats Wiki identifiers. In combination we now know how these three repositories interrelate, overlap, and where they do not overlap. Coyle noted the value of such mappings and pointed out how vital these “switching stations” would be to the web of linked open data [10]. In Figure 8, we can see a SPARQL query on the Wikidata Query Service. The results of this query are returned as a table. From these results we find that Wikidata item with QID Q278934 has an English label of shapefile, a PUID of x-fmt/235, a Library of Congress Format identifier of fdd000280, and a File Formats Wiki page called Shapefile.

Beyond providing infrastructure for this type of alignment for crosswalking between repositories, another advantage of having digital preservation technical metadata in Wikidata is to allow for queries that are ask questions of this data in combination with other data from Wikidata. For example, we can ask questions about software that leverage data from other parts of the knowledge base. In Figure 9, we see a screenshot of a timeline generated as one of the result format options on the Wikidata Query Service. In this query we are asking for the dates of birth for the developers of free software to be plotted on a timeline. It is possible to run this query against the Wikidata Query Service and results will be returned. This is due to the fact that information about people is also in Wikidata. In contrast, if we attempted to run this query in the context of a repository of technical metadata related to digital preservation, results would only be returned if the repository had been populated with data about the software developers. Inclusion of data about people might be considered outside of the scope of digital preservation. It is also unclear if such a stand-alone

24http://fileformats.archiveteam.org/wiki/Main_Page
25To run this query on the endpoint follow this link http://tinyurl.com/gl3nmq8.
26Query url: http://tinyurl.com/mstqtpg
Figure 9: A SPARQL query on the Wikidata Query Service endpoint illustrating how data from the digital preservation domain can be combined with data about people.

repository for digital preservation would have the resources to maintain a SPARQL endpoint for their data.

Another example is the query seen in Figure 1027. In this query we are asking for the latitude and longitude for the birthplaces of developers of free software. This query demonstrates the power of storing technical metadata in a cross-domain knowledge base, and how data from the domain of digital preservation can be combined with other data to answer a broad range of questions.

Creating SPARQL queries to get data our of Wikidata allows us to present subsets of this data in a flexible manner. Such queries can be harnessed to drive future linked-data powered applications for the digital preservation domain.

5 WHY WIKIDATA?

Some of the challenges of describing the domain of digital preservation are the many divergent perspectives about how to model the domain and how to define the entities within it. In Wikidata, we have the ability to negotiate and refine the items and properties we need for our work in digital preservation, and to experiment with multiple strategies in a live system.

Due to the commons-based peer production of the Wikidata community, this is only the starting point for integration of additional data about the digital preservation domain. As more editors join the Wikidata community, an increasing number of domains are being modeled in Wikidata. New items are being created, new properties are being discussed and created, and more statements are being asserted, many with references that serve as the provenance for the claims [27].

There are several communities already working to model data within specific domains. Examples of well-established data modeling efforts are those of the Su Lab at The Scripps Research Institute and of WikiCite28. A group of researchers from the Su Lab have been leading efforts to model genetic information related to humans and model organisms in Wikidata [15, 23]. The Su Lab team has published documentation of their data modeling activities and created tools for their own work and made those tools available for others to use or adapt. One of the tools that the Su Lab has shared is Wikidata Integrator [43]. Wikidata Integrator is a Python library that supports the creation of Python bots to programmatically edit Wikidata.

Another active community working on data modeling is WikiCite. The participants are working together to model bibliographic relationships, and make bibliographic metadata readily available to editors who add references to Wikidata or other projects such as the different language versions of Wikipedia [36]. This group is made up of participants from many different organizations, many of whom have experience working with bibliographic metadata in several other computational workflows. Members of the WikiCite community have also created specific tools to support interacting with bibliographic metadata and Wikidata. For example, the Scholia project, a web service “creates on-the-fly scholarly profiles for researchers, organizations, journals, publishers, individual scholarly works, and for research topics” [26].

These sub-communities of the larger Wikidata project are examples of how, in a peer production system, the users are constantly updating, correcting, and implementing creative strategies to support collaborative data modeling work.

6 CONCLUSION

Through participating in the Wikidata community, we are exploring the potential for Wikidata to serve as the repository of technical metadata for the international digital preservation community. An advantage to using the Wikidata infrastructure is that this data is structured, queryable and computable [7, 29]. The value of collaborating with people domain expertise in digital preservation in creating the data models for the digital preservation is that their domain expertise is helpful in creating the underlying conceptual relationships of the knowledge base. Contributors to Wikidata create entity-level content, but also create the conceptual structure for the system [28].

Wikidata is multi-lingual knowledge base with support for more than 350 languages[14]. The multi-lingual design of Wikidata means that people from many different language communities are able to read the content of Wikidata in their own languages. For example, in Figure 11 we see a side-by-side comparison of the infoboxes for Adobe Acrobat in English Wikipedia, French Wikipedia, and Japanese Wikipedia. This is an advantage over knowledge bases that only support users of a single language. Due to the way Wikidata was designed, the work of an editor to add content to Wikidata

---

27http://tinyurl.com/mwuxhse
28https://meta.wikimedia.org/wiki/WikiCite
benefits users across all supported languages as that data is available in each of the supported languages.

We see value in working to create this infrastructural component of the web of linked data. We are working together to create this content in a system that was designed to ensure that this structured data will belong to the public domain in order to avoid the situation of needing to purchase access to this data from a for-profit business organization [19].

Making a decision to invest in a project that is building infrastructure for the public domain may not be possible for all domains. In the domain of digital preservation, activities that contribute to public domain infrastructure will allow us all to share the work of maintaining access to this structured data. We are all free to fork this project at any time, all data is downloadable at any time, and we can reuse data from any contributor to Wikidata.

If we are successfully in curating data related to the technical metadata of digital preservation then we will have created a powerful resource that will enable others to reuse this metadata for their own systems and projects. Once we have established which are the readable file formats for a particular version of a particular piece of software and curated that data in Wikidata then, by virtue of the Creative Commons Zero (CC0) license[29], others can reuse that data rather than create it anew themselves.

Imagine a future in which the technical metadata we need for our digital preservation work is freely available to all as linked open data. Please join us as we build this future together.

7 ACKNOWLEDGEMENTS

We thank Wikimedia Deutschland[30] for supporting the Wikidata project. We thank the Wikidata community for sharing thoughts,
feedback, and data modeling work. Specifically, would like to thank all participants of WikiProject Informatics for engaging with us as we work to describe the domain of digital preservation in Wikidata.

REFERENCES


[31] https://www.wikidata.org/wiki/Wikidata:WikiProject_Informatics
Metadata-Driven Approach for Keeping Interpretability of Digital Objects through Formal Provenance Description

Chunqiu Li
Graduate School of Library, Information and Media Studies,
University of Tsukuba
Japan
lichunquua@126.com

Shigeo Sugimoto
Faculty of Library, Information and Media Science,
University of Tsukuba
Japan
sugimoto@slis.tsukuba.ac.jp

ABSTRACT

Metadata about digital objects help users find, understand, use and reuse those objects. Longevity of digital objects is a vital issue for digital preservation, which means that the metadata about digital objects must be maintained as well, so that their content and meaning should be maintained over time. Open Archival Information System (OAIS) defines three metadata components, which have to be maintained with Digital Object – Representation Information of Digital Object, Preservation Description Information (PDI) in an Information Package, and the Content Information given to every Information Package. Provenance of a digital object, which is one of the five categories of PDI, is a crucial record of the history of the object over its lifecycle. Since metadata are exchanged as digital objects on the Web, machine-readable and interoperable provenance description of metadata is required for the long-term maintenance of metadata. This paper presents issues in the longevity of metadata, especially the issue of metadata provenance based on the Singapore Framework for Dublin Core Application Profiles (DCAP), which is well known for metadata interoperability in the networked information environment. The paper first briefly discusses features of metadata as first class objects on the Web. Then, we address potential risks in affecting interpretability of digital objects and issues in the consistent maintenance of metadata. Next, the W3C PROV standard for general provenance description and Resource Description Framework (RDF) for metadata exchange on the Web are adopted as the base models for provenance description of metadata. We developed simple provenance description models for formal provenance description for both structural features and vocabularies of metadata. The models are designed based on Entities and Activities defined by the W3C PROV in correspondence with primitive changes of metadata application profiles and metadata vocabularies, respectively. We also provide formal provenance description examples corresponding to structural changes in a metadata application profile along with semantic changes in the use of its metadata vocabulary. We discuss limitations of our proposed models and review provenance related research. Finally, the main findings of this research are summarized in the conclusions.

KEYWORDS

Metadata, digital object, metadata schema, application profile, metadata vocabulary, provenance, long-term maintenance

1 INTRODUCTION

Long-term accessibility of digital collections requires keeping digital objects usable over time and across communities. Metadata plays an important role in the continued accessibility of digital collections and is used in a wide range of fields, such as computer science, library and information science, archival science, and so forth. Recent developments in research data sharing (e.g., Research Data Alliance 1) and cultural resources aggregation (e.g., Europeana2 and DPLA3) have also increased demands to keep research data and cultural collections alive over time. Maintaining metadata over time is important to keep research data and cultural contents reusable over time.

This paper discusses a metadata-centric study to keep digital objects interpretable, focusing on issues such as long-term maintenance of metadata schemas and metadata vocabularies. Metadata schema defines structural, syntactic and semantic features of metadata and uses metadata vocabulary that is a set of metadata terms to describe metadata record. In the long run, metadata provenance that records revision history of metadata schema and metadata vocabulary should be consistently recorded, since provenance is useful to auditing errors, justifying data authenticity, and identifying invalidated data, etc. In this paper, we developed simple provenance description models for metadata schemas and metadata vocabularies, respectively. The models enable maintainers to formally describe primitive revisions between two consecutive versions of a metadata schema and a metadata vocabulary in a machine-processable form, so that

1 https://www.rd-alliance.org/
2 http://www.europeana.eu/portal/en
3 Digital Public Library of America, see https://dp.la/
their revision history can be consistently managed and effectively traced over time.

This paper provides a brief overview of the main results gained from our earlier research about metadata longevity conducted by the authors [1,2]. The rest of this paper is structured as follows. Section 2 analyzes features of metadata on the Web. Section 3 gives an overview of risks affecting metadata longevity. Section 4 discusses issues of metadata interpretability, long-term maintenance of metadata schema and metadata vocabulary, as well as metadata provenance. Section 5 presents our newly devised models for formal provenance description of metadata, and Section 6 discusses the formal provenance descriptions of metadata application profiles and metadata vocabularies based on our models using examples for illustration. Section 7 presents limitations of our models and briefly reviews related research on provenance. Section 8 summarizes main findings of this study.

2 FEATURES OF METADATA ON THE WEB

Metadata (Greek: meta- + Latin: data “information”) [3] as structured data is generally defined as “data about data”. In library domain, a card catalog and its electronic counterpart are common examples of metadata. Metadata can be an object in databases or in systems. Metadata in the networked information environment has features different from conventional metadata that is primarily designed for use on a single database or a set of databases. An instance of metadata on the Web is no longer an object enclosed in a database, but is a digital object which is transferred from one site to another and shared among those sites. We call such metadata object as a “first class object”. This paper discusses features of metadata on the Web as follows [4,5].

Structural Features: Metadata is typically structured according to a scheme. Structural features of metadata are assertions about data structure, mandatory levels, iteration constraints of description, and so forth. Such assertions represent attributes and values of resource in machine-readable form.

Syntactic Features: Metadata can be serialized in different syntaxes, e.g., HTML, XML, RDF/XML, Turtle, JSON, JSON-LD.

Semantic Features: The elementary semantics of metadata are specified and defined in a metadata vocabulary. The meaning of each metadata term and relationships between terms are identified as the semantic features of metadata. Uniform Resource Identifier (URI) is used as the base scheme to identify a term in the Linked Open Data (LOD) environment.

Metadata interoperability is crucial not only across communities but also over time. Metadata standards are the basis for interoperability of metadata [6]. However, schemes for metadata interoperability over time are still not well developed. In the long term, metadata need to be kept interpretable not only by humans but also by machines. We need to understand the potential risks and develop strategies to keep metadata instances consistently interpretable over time.

3 RISKS IN INTERPRETABILITY OF DIGITAL OBJECTS

Digital objects are preserved as a sequence of bits. It is of importance to ensure that the bits remain intact and correct over time. However, bit preservation alone is not sufficient for the long-term preservation of digital objects. Digital objects should be kept interpretable across the changes in many aspects over time. As OAIS reference model defines three metadata components, Representation Information, Preservation Description Information (PDI) and Content Information, metadata has to be preserved with digital objects. Those metadata may be stored in a database with the preserved digital objects as an archival information package (AIP). This means that metadata schemas and vocabularies used in those metadata have to be maintained over time as well as those AIPs. In the LOD environment, metadata schemas and vocabularies are digital objects, which should be kept usable for long-term as well as the preserved digital objects.

Preservation of digital objects requires preventing damages or loss of digital objects. We need to manage risks of damages or loss in the preservation process of digital objects. Risk management for keeping metadata and their schemas safe is a crucial research issue. In the OAIS reference model, risk management is an essential part of preservation planning [7]. However, OAIS does not discuss risk management for metadata and their schemas exchanged on the Web. Therefore, this paper analyzes the risks in affecting interpretability of digital objects from the viewpoint of metadata. The following risks might cause inconsistency in digital preservation: (1) Metadata schema and metadata vocabulary for a digital object may be unknown, improperly recorded, lost, changed, or obsolete, (2) Metadata schema and metadata vocabulary for a digital object may be improperly maintained and their revision history may not be consistently recorded, (3) Provenance information about the digital objects and their metadata schemas as well as metadata vocabularies may not be consistently recorded in machine-processable form, and (4) Resource identifiers of any of those instances may be inconsistent.

The following functions have to be studied in order to reduce the risks in metadata longevity: (1) preserving the documents of metadata schema and metadata vocabulary, (2) recording and maintaining metadata schema and metadata vocabulary along with their revision history, (3) recording necessary and interoperable provenance of metadata schema and metadata vocabulary, and (4) creating sustainable identification mechanisms and schemes for metadata instances.
4 LONG-TERM MAINTENANCE OF METADATA

This section discusses long-term maintenance of metadata schemas, long-term maintenance of metadata vocabularies, and metadata provenance as key aspects of metadata longevity.

4.1 Metadata Preservation: Keeping Temporal Interoperability of Metadata

Metadata plays crucial roles in digital collections, e.g., finding aids, rights management, etc. Therefore, metadata should be preserved as well as the digital resources in the collection.

PDI of OAIS has five categories which are Provenance, Reference, Fixity, Context, and Access Rights Information. Provenance documents the history of content information, and tells the origin or source of content information, and any changes that may have taken place since it was originated, who has had custody of it since it was originated, providing an audit trial for the content information [8]. Provenance provides the credibility information about a preserved resource to the users in the future as it contributes to evidence supporting authenticity. Provenance describes change history of a digital object over time and can be viewed as a special type of context information [9].

This paper focuses on the formal provenance description of the structural features of metadata and vocabularies defined for metadata. In this paper, we use the terms Metadata Schema and Metadata Vocabulary. We define metadata schema as a description of scheme that defines structural features of metadata and metadata vocabulary as a controlled set of terms defined for metadata. For example, Simple Dublin Core defines a metadata vocabulary composed of 15 metadata terms and very general structural features where any defined element is repeatable and optional. The metadata vocabulary of Simple Dublin Core is known as Dublin Core Metadata Element Set4.

The Singapore Framework for Dublin Core Application Profile (DCAP) defines the components of metadata schemas for an application and related components, such as metadata vocabulary. A DCAP has five components, which are Usage Guidelines, Syntax Guidelines and Data Formats, Functional Requirements, Domain Model, and Description Set Profile [10]. The framework is developed based on the Web standards (e.g., RDF5, RDF/S6) and is used as a basis for our research to discuss issues with the longevity of metadata. This is because the framework separates the structural and semantic features of metadata. In particular, we focus on provenance description of Description Set Profile in this paper.

In the LOD environment, RDF/S and OWL7 are basic schemes to define metadata terms and vocabularies. Each term is defined as a resource identified by URL. We discuss longevity of metadata vocabulary from the viewpoint of provenance description of metadata terms. Term mapping between metadata vocabularies is often done to merge two or more sets of metadata described on different metadata vocabularies. Mapping itself is a crucial data resource to keep a record of the merger for future purposes. However, in this paper, maintaining mapping over time is not discussed although it is a crucial issue for long-term use of metadata.

4.2 Long-term Maintenance of Metadata Schemas

A metadata schema should be preserved as well as metadata instances created from the schema. Metadata schemas are preserved as a document for human readers in a conventional maintenance environment of metadata. The state of the art environment of the Web provides standards and models to formally describe metadata schemas in a machine-processable form, e.g., RDF, OWL and the Singapore Framework for Dublin Core Application Profiles. Metadata schemas are no longer simple document-like objects, but are complex objects transferrable over networks. Therefore, we need to develop technologies to maintain metadata schemas and vocabularies over time for the longevity of metadata.

In this paper, we propose a formal model for provenance description of structural constraints of metadata schema based on the Description Set Profile of Singapore Framework. The changes between two consecutive versions of a metadata schema should be recorded as provenance description of the metadata schema. The advantage of formal provenance description of metadata schemas over conventional change-logs is automated auditing to help find errors and inconsistencies between the different versions of the metadata schema.

4.3 Long-term Maintenance of Metadata Vocabularies

It is recommended to reuse existing and well-known metadata vocabularies to improve semantic interoperability of metadata. This means that application metadata schemas rely on standard metadata vocabularies, in particular maintenance of the definition of metadata terms. For example, changes in descriptions of meanings of a term may affect application metadata schemas that use the term. However, many metadata vocabularies have no clear policy regarding their change documentation [11], and hence monitoring metadata vocabularies is of importance. Maintenance agencies of standard vocabularies should have policies for maintaining the vocabularies for their sustainability.

---

5. http://www.w3.org/TR/rdf11-primer/
7. http://www.w3.org/TR/owl-features/
A metadata vocabulary is comprised of a set of terms and relationships between terms. When a newly defined version of a metadata vocabulary is published, there may be some changes from its previous version, e.g., the meaning of a term can be changed, relationship between terms can be revised, a composite term can be split to single terms, or single terms can be merged into one composite term. The changes to a metadata vocabulary should be also consistently recorded as provenance of the metadata vocabulary.

4.4 Metadata Provenance

Provenance (from the French provenir, “to come from”) is the description of the history of an object. Provenance is used in several fields, such as identifying authorship of art works, justifying trustworthiness of data, reproducibility of scientific research, etc. Provenance of a digital object describes how the digital object came to the current state since its origination.

For long-term preservation of digital objects, we need to record provenance of the digital objects in a digital form, which we call digital provenance in this paper. Digital provenance is included in well-known preservation standards, such as the OAIS reference model and the PREMIS standard for preservation metadata. Digital provenance typically describes agents responsible for the custody and stewardship of digital objects, key events that occur over the course of the digital object’s life cycle, and other information associated into the Web environment, machine-processable, interoperable and traceable provenance is required. The following sections present formal provenance description of metadata proposed by the authors which has its basis in W3C PROV standard and in RDF.

5 FORMAL PROVENANCE DESCRIPTION OF METADATA

Provenance description of metadata can be in various forms, such as natural language, RDF, etc. This study records provenance description of metadata in RDF, which is widely used for metadata exchange on the Web. Formal provenance description in RDF holds advantages over semi-formal and informal provenance description.

5.1 W3C PROV for Metadata Provenance

The Provenance Working Group at W3C has published the PROV family of documents, including the PROV Data Model (PROV-DM), PROV Ontology (PROV-O), and so forth. The working group aims at the interoperable interchange of provenance information in heterogeneous environments such as the Web. PROV-DM is a conceptual data model, which defines a set of concepts and relations to represent provenance [13]. PROV-O defines a set of classes and properties as an OWL2 ontology allowing mapping PROV-DM to RDF [14].

An Entity is a physical, digital, conceptual, or other kinds of thing. An Activity is something that occurs over a period of time and acts upon or with Entities. Activity is used to represent how an Entity comes into existence and how attributes of an Entity change to become a new Entity. W3C PROV defines relationships between Entities, relationships between Entities and Activities, relationships between Activities, and other relation types. To describe metadata provenance, we classified Entities and Activities affecting revisions to metadata application profiles and metadata vocabularies, respectively. The following two sections will briefly address provenance description of metadata application profiles and metadata vocabularies.

5.2 Model for Formal Provenance Description of Metadata Application Profiles

This section shows the DSP-PROV model with functions to describe deletion, addition and revision of structural features of metadata schema. The classified Activities are generalized into three change types – Addition, Deletion and Revision. The classified Entities are Description Set Profile (DSP) and its components, which are named as structural schema instances in this paper.

DSP defines structural constraints of metadata. There are two levels of templates in a DSP, i.e., Description Template (DT) and Statement Template (ST). DTs contain “statement templates that apply to a single kind of description as well as constraints on the described resource”. STs contain “all the constraints on the property, value strings, vocabulary encoding schemes, etc. that apply to a single kind of statement” [10]. Structural Constraints (SCs) contain mandatory levels, iteration constraints and other constraints on properties and property values defined in statement templates.

---

8 http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
Fig. 1 depicts the DSP-PROV model using UML Class diagram: (1) Generalization is represented with a hollow triangle on super-classes (i.e., Entity and Activity), (2) Aggregation is represented with a diamond on containing classes (for example, DSP, RevisionOnDSP), and (3) Association is represented by a line with an arrow that describes the relationship between Entity and Activity. The DSP-PROV model uses the properties from PROV-O when applicable. PROV Validation and PROV Generation respectively represent the deletion and addition of structural schema instances. PROV Derivation, PROV Invalidation, PROV Generation and PROV Usage together describe the revision of structural schema instances. If applicable, DSP-PROV can also describe relationships between Activities in the cases where an Activity used an Entity generated by another Activity.

![Figure 1: Overview of the DSP-PROV model.](image)

Fig. 2 shows the classified Activities to describe structural changes of metadata schema. The naming convention of the Activities in this paper is “Activity Type + On + Abbreviation of structural schema instance”. For instance, Revision Activity that acted upon a DT and led it to a new DT is named as an Activity instance of RevisionOnDT. Fig. 2 also indicates the relationships between classified Activities. The Revision Activity acted upon the containing Entity (e.g., a DSP) has sub-activities – Deletion, Addition and Revision acted upon its contained Entity (e.g., a DT of the DSP). DT changes caused by DeletionOnDT, AdditionOnDT and RevisionOnDT will result in DSP changes caused by RevisionOnDSP. Therefore, RevisionOnDSP has sub-activities, i.e., DeletionOnST, AdditionOnST and RevisionOnST; RevisionOnST has sub-activities, i.e., DeletionOnSC, AdditionOnSC and RevisionOnSC. This paper uses property dcterms:hasPart that is recommended for modeling sub-activities [15].

![Figure 2: Activity relationships in the DSP-PROV model.](image)

5.3 Model for Formal Provenance Description of Metadata Vocabularies

Fig. 3 shows Vocab-PROV model with functions to describe primitive changes of a metadata vocabulary and its metadata terms. The approach for building the Vocab-PROV model is similar to DSP-PROV model. We classified the Activities and Entities to describe the changes of metadata vocabularies. Vocabulary, Term and Term Definition are classified as three subtypes of PROV Entity to describe the provenance of metadata vocabularies. As illustrated above, a Term can be a concept or a class or a property. In the case of a concept, its definition may include its narrower term(s), broader term(s), association/related term(s), and other information. In the case of a class, its definition may include a description of its meaning, a label(s), a URI, super-class(es), sub-class(es), used property(ies), and other information. In the case of a property, its definition may include a description of its meaning, a label(s), a URI, super-property(ies), sub-property(ies), domain, range, expected value and other information. To describe the provenance of metadata vocabularies, Activities acting on the previously classified Entities are categorized into the following types, i.e., Revision, Addition, Deletion and Replacement. The Replacement is defined to describe the change cases such as term split and term merge.
A revision of a vocabulary is caused by a revision of its terms. The revision of a term may be a revision of the term as an instance, or a revision of the documentation for the term. For example, replacement of a single term by a set of terms is a revision of an instance, and replacement of a title text is a revision of the term definition.

For example, replacement of a single term by a set of terms is a revision of an instance, and replacement of a title text is a revision of the term definition.

Fig. 3: Overview of the Vocab-PROV model.

Fig. 4 shows the relationships between the classified Activities to describe provenance description of metadata vocabularies. A RevisionOnVocabulary is comprised of zero or more than one RevisionOnTerm and RevisionOnTermDefinition. Similarly, RevisionOnTerm has sub-activities that are AdditionOnTerm, DeletionOnTerm and ReplacementOnTerm; RevisionOnTermDefinition has sub-activities that are AdditionOnTermDefinition, DeletionOnTermDefinition and ReplacementOnTermDefinition.

Fig. 4: Activity relationships in the Vocab-PROV model.

6 FORMAL PROVENANCE DESCRIPTION EXAMPLES

A metadata application profile usually uses terms defined in existing metadata vocabularies. However, the metadata application profile may use the term meaning, which may be narrowed from the original meaning for better fit of the meaning to the application. The terms included in an existing vocabulary are usually defined within the namespace of the vocabulary without version information. Therefore, in this section, we do not take into account the versions of the terms and we focus on the changes of term meaning defined in the metadata application profiles.

Section 6.1 provides an example of semantic change and structural change that we found from the documents of DPLA MAPs. Section 6.2 and 6.3 discuss the provenance description about the changes using RDF graphs based on our proposed Vocab-PROV model and DSP-PROV model, respectively. In Section 6.4, we briefly present the relationships between the semantic change and structural change in the given change examples.

6.1 Example for Semantic Change along with Structural Change

Digital Public Library of America Metadata Application Profile (DPLA MAP) defines structural constraints of metadata, which include property, usage, obligation, range and others information in tabular form. DPLA MAP uses classes and properties from existing vocabularies, such as EDM, ORE, DC, DCTERMS, DCMITYPE, Geo vocabulary, etc. Three versions of DPLA MAP have been released, i.e., V3, V3.1, V4. DPLA MAP does not provide exact meaning and definition for its classes and properties. The value of the "Usage" column

---

9 https://dp.la/info/developers/map/
provides the kind of information related to meaning and definition of a term (i.e., class and property), which is written as the value of "Term Meaning" in the table titled "comparison between V3.1 and V4" of Fig. 5. DPLA MAP V3.1 provides changes from V3 to V3.1 and DPLA MAP V4 provides changes from V3.1 to V4 in natural language.

Fig. 5 shows change examples from DPLA MAP V3.1 to V4 that includes both structural change in DPLA MAP and semantic change in its metadata vocabulary. Both DPLA MAP V3.1 and DPLA MAP V4 define property edm:object for class ore:Aggregation to describe "object". The meaning of edm:object in DPLA MAP V3.1 and V4 are "Unambiguous URL to the DPLA content preview" and "The URL of a suitable source object in the best resolution available on the website of the Data Provider from which edm:preview could be generated for use in a portal", respectively. Fig.5, Fig.6 and Fig. 8 use a short expression of the two definitions, i.e., "Unambiguous..." and "The URL...".

![Figure 5: Example of semantic change along with structural change.]

6.2 Formal Provenance Description for Semantic Change of Metadata Term

As shown in Fig. 5, the minimum occurrence of edm:object has been changed from “1” to “0”. The change was caused by a revision activity, which is an instance of the class dspprov:RevisionOnSC (note: "dspprov" is the prefix for the classes of the DSP-PROV model [1]). Fig. 7 shows RDF graphs of the provenance description about the structural constraint change.

![Figure 6: Provenance description for the above semantic change using RDF graphs.]

![Figure 7: Provenance description for the above structural change using RDF graphs.]

The minimum occurrence constraint defined in the Statement Template (ST) instance that defines all the structural constraints on the property edm:object is expressed as the literal value of property owl:minQualifiedCardinality. The new Structural Constraint (SC) represented in the lower
dotted-rectangle was derived from the previous constraint in the upper dotted-rectangle. The newly defined minimum occurrence constraint was generated and the previously defined minimum occurrence constraint became invalidated through the same Activity instance of dspprov:RevisionOnSC.

6.4 Connection between Semantic Change with Structural Change

In general, a metadata application profile uses terms from metadata vocabularies. Semantic changes of the terms used in a metadata application profile may be synchronized with structural changes of the metadata application profile. This section shows linkage of semantic change on a term and structural change in a metadata application profile. Figs. 6 and 7 show the provenance description about the semantic change and structural change in the examples given in Section 6.1. As shown in Fig. 8, the connection between Figs. 6 and 7 is the property constraint in the Statement Template (ST), which is expressed as the resource value of owl:onProperty.

Figure 8: Connection between semantic change and structural change.

7 DISCUSSION

In Section 7.1, we briefly discuss the limitation and implication of our proposed provenance models, i.e., DSP-PROV model and Vocab-PROV model. In Section 7.2 and 7.3, we review the current state of provenance description.

7.1 Limitation of the Proposed Models

The DSP-PROV model is designed based on Description Set Profile of Dublin Core Application Profile. However, there are no worldwide standards for the development of metadata application profile as not all metadata application profiles follow the same structure. This means the DSP-PROV model may not be applicable to all existing metadata application profiles.

The Vocab-PROV model generalizes primitive changes in a metadata vocabulary. In practice, the changes to a metadata vocabulary may be complex especially when considering temporal information over a long period. The Vocab-PROV model may not be applicable to complex changes. The maintenance of semantics of terms is a challenging issue and semantic interoperability of terms over time is difficult to achieve.

7.2 Provenance Description Models and Vocabularies

In our earlier research, we conducted survey of existing provenance description models and vocabularies [16]. There are already a wide range of models, ontologies and vocabularies that can be used for provenance description, such as Open Provenance Model (OPM), Open Provenance Model Vocabulary (OPMV), Open Provenance Model OWL Ontology (OPMO), Open Provenance Model (OPM) for Workflows (OPMW), Provenance Vocabulary (PRV), Vocabulary for Data and Dataset provenance (Voidp), Provenance, Authoring and Versioning Ontology (PAV), W7 Model, Provenir Ontology, BBC Provenance Ontology, W3C PROV standard and others. The CIDOC Conceptual Reference Model (CRM) in the museum community has also been extended to model provenance information of digital objects [17].

Provenance description is required in both conventional and Web environment. Recording provenance in a form interpretable by both computers and humans is required. However, existing technologies and standards are not specialized for metadata schema and metadata vocabulary. Specially, models for formal provenance description of metadata are not sufficiently explored. In the Web environment, there is a need to develop models for formal metadata provenance description. It is because that formal provenance description of metadata in machine-readable and interoperable form supports automated and effective metadata maintenance. In this study, we have developed models for provenance description of metadata application profiles and metadata vocabularies, respectively.

7.3 Provenance Description in Different Domains
Provenance related research has been conducted in a wide range of domains, such as museum, library and archive (MLA) community, computer science and e-science, etc. Provenance information can be used to identify authorship, ownership and authentication of objects, e.g., the Council of European Research Libraries (CERL) 10 website and the Getty Provenance Index Databases11 provide search services for provenance information. International Research into the Preservation of Authentic Records in Electronic Systems (InterPARES) 12 project addresses the importance of provenance for keeping trustworthiness of digital records. Provenance in e-science can be used to reproduce research data.

The Bodleian libraries at the University of Oxford devised a data model to represent contextual information of research outputs in the Oxford University Research Archive (ORA) 13, which is a long-term data repository for scholarly research outputs. The model incorporates PROV-DM to describe activity related to research outputs, e.g., creation activity, funding activity, and publication activity. Activity-based description of relationships for a journal article using PROV-0 is given as an example [18].

Capturing the provenance information of electronic records is a concern for archivists. Conventional provenance in the arrangement of archival records arises from their creators, for example, individuals, cooperated bodies or families [19]. The scope of provenance for archival records encompasses to creator history, records history, and custodial history. The archival standards such as, General International Standard Archival Description (ISAD(G)), Encoded Archival Description (EAD), International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAAR(CPF)), and Encoded Archival Context (EAC) define the description elements for provenance information. The recordkeeping metadata standard ISO 23081 provides us reference to capture audit trails in the records management process [20].

Getty Thesaurus of Geographic Names14 adopts W3C PROV to describe revision history of geographic names. W3C PROV is used to document the Activity information about the revision of geographic names, e.g., Activity type (Create, Modiﬁy) and temporal information associated with the Activity.

As introduced above, many communities have paid attention to provenance description, especially the change history and activity related to objects. However, these provenance description elements are designed for specific domain requirements and not generalized for metadata provenance. That is, they cannot be directly applied to describe provenance of metadata application profiles and metadata vocabularies. Therefore, the aim of this research to propose general models for provenance description of metadata is novel.

8 CONCLUSIONS

This paper addresses issues in longevity of metadata, especially temporal interoperability of metadata over time. The main contributions of this paper are: (1) development of the DSP-PROV model for provenance description of metadata application profiles, (2) development of the Vocab-PROV model for provenance description of metadata vocabularies, and (3) provision of examples of formal provenance description, especially considering both the structural constraint changes of metadata schemas along with the semantic changes of metadata vocabularies. Due to the space limitation of this paper, we will not specifically introduce the evaluation and description examples of the two proposed models. The details are presented in our other two papers [1,2].

The provenance description of metadata application profiles and metadata vocabularies together can reveal the revision history of structural features and semantic features of metadata instances, which can help users to interpret metadata instances. These descriptions in machine-processable form on the Web can be traced using Semantic Web technologies.

The main findings of this article are summarized into the following points: (1) Provenance information is crucial component to keep longevity of digital objects, (2) Provenance information should be consistently recorded in machine-processable form on the Web, (3) The devised DSP-PROV model and Vocab-PROV model enable us to keep track structural changes of metadata schemas and primitive changes of metadata terms, respectively, and (4) Formal provenance description holds advantages over provenance description in natural language. For instance, formal provenance description helps consistent maintenance of metadata over time; formal provenance description can be used to find errors in semi-provenance description that is recorded in natural language. In the future, we expect to explore practical services of provenance use cases in memory institutions.

Furthermore, implementation of existing provenance models with metadata standards (e.g., PREMIS dictionary; controlled vocabularies of Library of Congress) is also an applicable approach for provenance description of metadata. Our previous paper used this approach and briefly discussed provenance description of metadata schemas through combining the core of PROV data model with PREMIS data model [18]. In this paper, we adopted W3C PROV and RDF to propose provenance model for metadata longevity from the aspects of metadata application profile longevity and metadata vocabulary longevity.

---

10 https://www.cerl.org/resources/provenance/main
11 http://www.getty.edu/research/tools/provenance/search.html
12 http://www.interpares.org/
13 http://ora.ox.ac.uk/
14 http://www.getty.edu/research/tools/vocabularies/tgn/
Metadata provenance is still a new topic. We defined models for provenance descriptions. The models rely on several infrastructure issues in the practical Web environment, such as longevity of namespaces and URIs, and long-term maintenance of widely used vocabularies. This paper does not discuss these fundamental issues because they need to be discussed in the digital preservation communities and other wider communities due to their importance. We consider that this paper provides several fundamental issues for discussion for the future development of digital preservation.

ACKNOWLEDGMENTS

We express sincere thanks to professor Atsuyuki Morishima, associate professor Tetsuo Sakaguchi, and assistant professor Mitsuharu Nagamori for their contributions to this research through discussions. We thank Mr. Tsunagu Honma and all students at our laboratory for their kind help. We express our special thanks to Dr. Bhuvan Narayan at University of Technology Sydney for her help to elaborate this paper. This work was partially supported by JSPS Kaken Grant-in-Aid for Scientific Research (A) [Grant No.: 16H01754].

REFERENCES


Digital art posterity: building a data model for digital art corpora

Louise Fauduet  
Bibliothèque nationale de France  
France  
louise.fauduet@bnf.fr

Céline Thomas  
Université Paris 8  
France  
celine_thomas@gmx.fr

Chu-Yin Chen  
Université Paris 8  
France  
chu-yin.chen@univ-paris8.fr

Adèle Sicre  
Université Paris 8  
France  
adele.sicrevidalou@gmail.com

Céline Thomas  
Université Paris 8  
France  
celine_thomas@gmx.fr

Chu-Yin Chen  
Université Paris 8  
France  
chu-yin.chen@univ-paris8.fr

Florent Aziosmanoff  
Living Art Lab  
France  
florent.aziosmanoff@gmail.com

ABSTRACT

The research project “Digital art and posterity”, February 2015-February 2018, aims to contribute to perpetuating digital artistic heritage by developing a general descriptive system of digital works of art which provides a way to model the technical and artistic characteristics of the work, its significant properties and its formal principles in text form.

This research is performed in collaboration between three institutions: the National Library of France (Bibliothèque nationale de France, BnF), the INReV laboratory of Paris 8 University and the Living Art Lab. The partners come from different fields — those of conservation, of artistic creation and of academic research — and combine their skills and experiences to devise a conceptual model and a data model true to the specificities of the project’s corpora.

This paper presents the main attributes of the proposed information model and ontology, so that all institutions involved in the collection, preservation and diffusion of digital art can provide input regarding its pertinence and interoperability.

CCS CONCEPTS

• Information systems–Document topic models  
• Information systems–Content analysis and feature selection  
• Information systems–Ontologies  
• Applied computing–Performing arts  
• Applied computing–Media arts  
• Applied computing–Digital libraries and archives

KEYWORDS

Digital Art, Information Model, Data Model, Interactivity, Transdisciplinarity, Versioning

1 INTRODUCTION

In recent years, there have been an increasing number of papers at iPRES on the topics of emulation solutions and digital art preservation [1]-[6]. Both issues are often intertwined, as computer-generated audiovisual artistic creations that are of a composite nature (sound, picture, video, motion picture, etc.) require preservation processes ranging from emulating the original to recreating the code and/or replacing the input and output devices with technologies of another nature. The obsolescence of hardware and software involved in the presentation of digital art is extremely rapid. And yet we need to ensure that future generations will be able to experience these works while maintaining the possibility to interact with the technical and artistic content, in order to affect the course of its performance.

“Art numérique et postérité” (“Digital art and posterity”) is a project based on collaboration between a national library, a university department and a private creation and research partner. The project aims to contribute to the preservation of digital art by working on its description.

We are aware that there is a wide range of definitions of digital art among the different communities of artists, researchers and curators. The relevance of the information model we are developing may be biased by the corpora and the professional environment of the project’s partners, even though the members come from different backgrounds and the collections are diverse. We will include in this paper a presentation of the partners, their work environment and the art used in the project to give an accurate picture of the information model’s immediate applicability.

We will then present the conceptual model designed in the project. The main characteristics of an information model for contemporary digital art are the place given to the artist’s intentions for the presentation of the artwork, as reconstituted by the curators or as expressed in his or her own words, and the importance of accounting for successive versions of an artwork.

We will discuss the challenges associated with the modeling process, and the way building a database and an ontology from the conceptual model impacts the overall description of digital art.

2 RESEARCH AS A MÉNAGE À TROIS

We will first introduce the actors of the project “Digital Art and Posterity”, their contributions and missions. We then will clarify the meaning of “art numérique” as we defined it using the corpora of each collaborator as a representative panel of digital art. We will explain the specificities of digital art in order to
highlight the challenges related to its preservation. Finally, we discuss the issue of the mass treatment of digital artworks.

2.1 A versatile project team

The research project “Digital Art and Posterity” is a project shared between three institutions: the Bibliothèque nationale de France (BnF, the National Library of France), the Images Numériques et Réalité Virtuelle research team (INRéV, Digital Images and Virtual Reality), part of the Arts des Images et Art Contemporain laboratory (AIAC E4010, Images Arts and Contemporary Art) of Paris 8 University, and the Living Art Lab, an independent art studio.

This project was made possible thanks to the Labex Arts-H2H, cluster of excellence in arts and human mediations\(^1\), laureate of the program “Investing for the Future” since 2011\(^2\). All of the Labex’s research projects are built around the principle of transdisciplinarity between different fields in the humanities. This experimental dynamic made it possible for the project partners to meet and to collaborate.

The first partner is the BnF, the project leader, and specifically its Audiovisual department. The BnF has a mandate to collect a vast array of publications. Over the years, it has developed best practices to describe documents ranging from manuscripts to video games. The Audiovisual department has built expertise in the field of digitized and born-digital heritage preservation. In the past few decades, the BnF has been gathering a digital artwork collection through legal deposit. It is mostly composed of documents on a physical medium but also entails web archiving and contents distributed online. One of its objectives in our project is to enrich its collection of digital art. It also aims at bettering access to its collections using more standardized description and access processes.

The second partner is the INRéV research team of Paris 8 University. For 30 years, this team has been developing a unique manner of apprehending digital creation. The laboratory’s aim was and remains to continuously create, innovate and hybridize art, science and technology. This team’s main figures are protagonists of French digital art history. Their artistic practices have given them an insider knowledge of digital techniques and their evolutions. One of the specificities developed by the research team over the years is interactive art with an autonomous and sometimes intelligent behaviour coupled with a reflexion on the creative process. Combined with the undergraduates\(^3\) and graduates’ productions at the Arts et Médiations de l’Image department\(^4\) (ATI, Computer Graphic and Art Department), these artworks constitute a remarkable corpus. INRéV joined the project to start working on its preservation strategy as none is in place at the University.

The third partner is the Living Art Lab, an independent studio which was previously part of Le Cube, a center for digital art creation. Le Cube, through the Living Art Lab, directed and produced interactive art installations, by offering the creators financial resources and IT skills. With experience in fostering and facilitating digital art, the Living Art Lab brings to the project its vision of digital art and its creations. Its goal is to promote and preserve Living Art installations. These form a coherent corpus with a specific approach to interaction which broadens our project’s panel.

In addition to this original trio, we called on other skills to enrich the project by recruiting interns and experts with various backgrounds: contemporary art philosophy, data processing and information science, management of audiovisual archives, computer science, art history, art conservation and digital humanities.

The versatility of the project team makes for a stronger scientific project. We rely on collective intelligence, which leads us to interesting moments of constructive misunderstandings and allows us to move forward to better solutions.

First and foremost, the “art numérique” that the partners want to describe and preserve had to be defined and matched with the corpora.

2.2 What is “art numérique”?

A conventional definition of “digital art” would be a diverse set of artistic creations based on the specificities of the computer language and the IT system. It characterizes artworks produced by computers and that can be rendered with electronic devices. This art genre began to be produced in the early 1960s and has not ceased developing since then through a wide range of artistic expressions. In Québec art médiatique is seen as a whole, and art numérique is understood specifically as online art. In the United States and in the UK, cultural institutions manage time-based media art as an ensemble\(^5\), of which software-based art is a subset. In France, art numérique or digital art is currently the most widely used category\(^6\). Its history encompasses the permanent evolution of digital technologies (for instance, virtual reality or augmented reality) and the developments in connected disciplinary fields (such as artificial intelligence, artificial life, cognitive science or robotics).

While we agree with this definition, it isn’t precise enough to fit the complexity of our corpora. Among the properties we need to express in our description are the characteristics related to the initial publication and production of the artwork, since the corpora of the partners have different publication histories: published works at the BnF, research results at the ATI-INRéV, exhibited objects at the Living Art Lab.

The first essential element of any description of digital art is that computer programming is at its center. In the project, we

---

\(^1\) http://www.labex-arts-h2h.fr
\(^2\) http://www.enseignementsup-recherche.gouv.fr/pid24578/investissements-d-avenir.html
incorporate both interactive, intelligent or behavioral artwork and the computer-animated image, since all these art forms are computer-based. The same artistic impulse governed their creation, as is shown in the ATI-INRéV corpus. They have to be preserved together according to specific strategies, and the computer systems used to create such moving images have to be described in the information packages.

Furthermore, building on earlier research on the role of the viewer in creation [8], digital art explores the ability to create an interactive relationship between the program and its external environment and in particular with the spectator [9]. The BnF’s corpus offers good examples of elementary multimedia interactive artworks with its art CD-ROMs collection, produced in the 1990s. Although these artworks are described in the BnF’s catalog, their bibliographic records are based on a common model for all electronic documents. The existing technical information is enough to render the discs in the reading rooms now, but insufficient to allow long-term preservation. There is, for instance, no record of the basic procedural rules needed to replay the art, although changes in technological context gradually limit our ability to perceive the artwork as intended.

Another expression of the interactive relationship that can be produced by digital art may be defined as “intelligent”, “behavioral” or even “autonomous”. Artworks in this category, more experimental, have been created within the ATI-INRéV and Living Art laboratories since the mid-1980s. They require captors and actuators, and/or virtual or augmented reality systems; their 3D images are generated in real-time by artificial life or intelligence algorithms. From these artworks emerge autonomous or intelligent behaviors that have been defined as a new form of interactivity. Couchot, Tramus and Bret named it “second interactivity” [10]; Aziosmanoff prefers the term of “Living Art” [11], [12], and Chen, using the concepts of complexity and enaction [13], calls it “enactive digital art” [14]. The description of these artworks has to include their behaviors, their self-actualization processes, and the interaction systems.

The most complex description and preservation challenges we have to face come from parts of the corpora heavily reliant on fundamental research on the art-science themes. Research domains such as artificial art/life, artificial art/intelligence, art/enaction, art/emergence have led to very unique and innovative artworks. Examples in our corpora use made-to-measure software (“Anyflo” by Bret [5], “Rodin” by Huitric and Nahas [16]), devices and platforms.

While interaction gives the artworks the appearance of unpredictability and instability due to the individualized experience of the spectator, the real instability of the digital objects comes from their successive versions in different software and hardware. In the thirty-odd years of INRéV’s existence, the artist-researchers have had to either update their artworks or let them die. Whether at the artist’s initiative or at a museum’s or gallery’s request, repairing artworks for a new exhibition requires different strategies: migration, emulation, recreation. Changes to the original art range from slight modifications to complete distortions. Over time, the successive versions of the artwork can diverge from replicas to form something akin to a series based on the same concepts. Describing versioning of complex objects is one of the main challenges of our project.

Along with this challenge emerges another of our issues: how can we make the distinctive nature of digital artworks work with mass processes?

Figure 1: La Funambule virtuelle, Alain BERTHOZ, Michel BRET, Marie-Hélène TRAMUS, 2000. Interactive installation. Interaction with the spectator through position and rotation captors attached to a belt, with an intelligent 3D computer-animated character developing a rebalancing strategy in real-time [15]. Uses a neural network program (Anyflo) [5].

Figure 2: Morphogénèse, Chu-Yin CHEN, 1996 (film; 7’13). Generative artwork, artificial life. Software: Anyflo. Result of researches on the creation of a virtual world (biosphere) within which the virtual creatures self-generate according to genetic algorithms [17][18].

2.3 Dealing with mass

The project partners have very different processes for collecting, describing, preserving and accessing their digital art collections. None of them are currently geared towards presenting digital art in an exhibition context, although it has happened and may happen again.

The Preservation & Art - Media Archaeology Lab (PAMAL7) at the Avignon School of Art has restored digital works of art in a media archeology perspective, to study the impact of the material context of creation on the artworks’ reception [19]. But the partners of our research project have different constraints, and systematic recreation of artworks on software and hardware matching the original is not practical given the volume of collections concerned, and access requests that are relatively unpredictable compared to a planned exhibition project, for instance.

The National Library of France has existing data and metadata repositories, and access systems. Any solution tailored for digital art must be included in these generic systems that combine data on all types of materials, and specific information might lose visibility in the process. There is a catalog and a data repository using semantic web technologies8. There are two digital archives, the Scalable Preservation and Archiving Repository (SPAR) [20] and the audiovisual archive. The latter is the current repository for digital art content, but has limited metadata querying capabilities. It is combined with a display system for the audiovisual research reading room, where emulation processes are only very partially automated. Library assistants deal with requests for materials requiring emulation, and have the tools and skills to provide access to relatively standardized content such as CD-ROMs from the 1990s. Any other level of installation and emulation currently requires the intervention of a technician or engineer on a case by case basis.

In contrast, INRéV has no preexisting processes, repositories or tools for preserving its digital art collections, even though it values the works created by its faculty and students highly, both as art and as teaching devices.

Working on an information model for digital art is a mutually beneficial project. It can improve the quality of preservation at the BnF as the information requirements are implemented throughout the digital preservation process with the input of digital art specialists. It can also help the Paris 8 University and the Living Art Lab develop new research on the properties of digital art, and help them evaluate whether the BnF can be a partner in their preservation needs.

The resulting data model has to be both generic enough to describe any digital artwork despite the variety of techniques and approaches in this field, and specific to allow the descriptive and preservation metadata collected to be used for preservation planning and prompt choices of access solutions.

3 THE CONCEPTUAL MODEL

The data model we have to build can only be constructed after we have set a conceptual structure. To describe it here, we will introduce our working methodology: we drew up an inventory of the digital artworks from our corpora, identified the information necessary for their preservation, and collected as much data as was judged useful. The structure of the conceptual model that organizes the gathered information was then developed. It was constructed according to the digital artwork’s lifecycle and respects its fluctuating nature from genesis to exhibition.

3.1 Awakening sleeping beauties

“Once upon a time, our artworks were forgotten, asleep in the midst of a wild forest. Going through the brambles, we managed to discover marvelous hidden collections; and now the awakening begins…”

The project’s team first needed to identify the “sleeping beauties” of our corpora.

The history of creation at the ATI-INRéV has been documented through exhibitions and the memory of this institution has been transmitted by the artist-teachers. However, an inventory of the artworks created by the professors and students has never been made.

The Living Art Lab has always been concerned with documenting the works it fostered. The artworks have been described on the Cube’s website at the time of their exhibition9. The Living Art Lab’s method includes a focus on collecting artists’ and IT engineers’ testimonies10. Nevertheless, there is no digital archive of the collection.

As for the BnF, there was no typology or keywords in their catalog that marked documents as digital art.

We inventoried over 300 works: around 200 from the BnF, 120 from the ATI-INRéV, 70 from the Living Art Lab. We established a limited number of categories: 3D computer-animated films; generative artworks; interactive, intelligent or behavioral installations; virtual reality and augmented reality artworks; gaming creations; interactive multimedia artworks.

In order to evaluate the information and data required for the identification of the artworks, we compared the descriptive records of institutions outside the project, such as the ones created on Navigart and reachable on Videomuseum’s website11, and we interviewed curators. Simultaneously, we had to design a test sample which was technically, historically and aesthetically representative of our corpora, but also narrow enough to allow us to analyze it thoroughly and to specify the information necessary for the preservation of digital artworks.

This identification of data performed, we had to awake our beauties; not with a kiss, but with documentation and data.
collection. The lack of existing information has led us, as others before, to try to capture information at the source: from the living artists or contributors. We could rely on previous research from the museum community, such as the “The Variable Media Questionnaire” of the Variable Media Network [12], the “Questionnaire for New Media Works” of the DOCAM Research Alliance (Documentation and conservaton of the media arts heritage) [13], or the recommendations of Matters in Media Art [14]. Through several case studies, the Digital Art Conservation project [15], initiated in 2010 by the ZKM (Zentrum für Kunst und Medientechnologie Karlsruhe), was able to apply concrete preservation strategies for a wide range of digital works based on artist interviews [22]. Discussions with researchers and curators in Québec and New York have confirmed this approach to be common practice among heritage institutions.

Technical information about the artwork seemed the aspect the most often absent from records. We began by sending a technical questionnaire on digital artworks to the artists, producers and contributors to gather information on how the work and its system operate (required IT environment, peripheral devices and interfaces, mode of interaction and mode of installation, etc.). We then interviewed them in person, as they were interacting with their artworks; this strategy has been well received so far. We record both the play-through of the emulated artwork on the computer and the artist interview on video. Our partners at the ATI-INRÉV have also introduced the method known as the “entretien d’explicitation” (“clarifying interview technique”) [23] to the project. The entretien d’explicitation is an introspective interview technique designed to lead the interviewee to access previously unconscious memories of the creation process and past actions. We aim at retrieving data on both the objective steps of creation and of the implicit knowledge related to the act.

Digital artists are aware of the ephemeral nature of their art. Many have chosen to reassess and recreate their works in order to continue to exhibit or perform them, but nevertheless proceed in different ways. For instance, some of them are forced to slow down technological improvements when they reactivate their works; this strategy has been well received so far. We record both the play-through of the emulated artwork on the computer and the artist interview on video. Our partners at the ATI-INRÉV have also introduced the method known as the “entretien d’explicitation” (“clarifying interview technique”) [23] to the project. The entretien d’explicitation is an introspective interview technique designed to lead the interviewee to access previously unconscious memories of the creation process and past actions. We aim at retrieving data on both the objective steps of creation and of the implicit knowledge related to the act.

Whether the artist is the only legitimate source for the description of the works and for preservation strategies choices remains to be determined. Part of our efforts in building an information model for digital art has certainly been dedicated to balancing the artist’s word and objective data. While this is not a breakthrough discovery, we aim at translating this line of reasoning into concrete rendering of the data model. It will also be a valuable tool to determine what preservation quality level the BnF is ready to grant to digital artworks, and whether certain types of digital art are more suited to the (relative) mass processes of a library.

3.2 Genesis

The descriptive system we wish to devise includes the concepts and key principles of digital art together with preservation information related to the digital objects. It captures the artwork from its creation to its diffusion, in various possible contexts: exhibitions, classes, reading rooms and so on. We designed our information model around two states of the artwork: the artwork during the creative process, or “artwork-to-be”, and the artwork we encounter as a spectator. We present here the “creation process” section of the model, followed by the “reception process” section.

The knowledge and experience of the project partners dealing in artistic creation have given us an opportunity to explore the creative process in detail. Our aim is to preserve and give access to the craftsmanship within the artwork. The first part of our system thus describes the artwork in its creative process: it begins by the “conceptual artwork”, where we develop the genesis of the artwork, and then continues to the “realized artwork”, where we detail the production process.

12 http://variablemediaquestionnaire.net
13 http://www.docam.ca/fr/outils/questionnaire-de-conservation.html
14 http://mattersinmediaart.org/acquiring-time-based-media-art.html
16 http://artnumeriqueposterite.labex-arts-b2h.fr/fr/content/programme

Figure 3: Simplified conceptual model: creative process
We need to describe the abstract idea of the artwork. This “conceptual artwork” emerges from the individuality of the author(s), their cultural and technical backgrounds, their artistic and scientific references and influences, their intention, etc. In this entity, we highlight the intellectual approach of the artist: the concept that the artworks are intended to convey and the starting point of the artistic project.

This abstract idea is materialized through digital tools and media; it evolves according to the reality it encounters. We named this information entity “realized artwork”. For instance, it encompasses the processes involved in the collaboration between the artists and the programmers, technicians, mathematicians, etc. Their contribution has to be understood not as merely technically proficient but also as creative: they too belong to the genesis of the artwork.

Another meaningful element of the creative process is the context of creation: the date, place, event or occasion of execution, the financial model and the production circuit it relied on.

The heart of the piece is the computer program — which is not necessarily written by the artist. We naturally include information entities dedicated to the artwork’s internal and external tools, whether already existing or produced specifically for the work, and the way they were used. “Internal system” describes all the software, algorithms and their evolutions. “External system” covers any type of hardware included in the artwork.

Parts of the corpora comprise art generated according to complex theoretical models, which forms a real challenge to defining a general descriptive system. For instance, artificial intelligence is based on algorithmic techniques such as neural networks; and artificial life uses genetic algorithms. Beyond considerations of aesthetics or techniques, the conceptual model needs to accommodate complex relationships between entities. We choose to build a network model where information is ranked from the general to the specific. Our information model has to be open and flexible in order to conform to any digital artwork type.

### 3.3 Performed, exhibited, rendered: the many lives of the artwork

We designed our information model around two states of the artwork: the artwork during the creative process, or “artwork-to-be”, and the artwork we encounter as a spectator. This second part of the model, describing the reception process, further distinguishes between the “performed artwork” and the “experienced artwork”.

![Figure 4: Simplified information model: reception process](image-url)

Once the artwork has been materialized, it is presented to the public. This is the best-known state of the piece, which we call the “performed artwork”. Many of the works in our corpora are based on the concepts of movement, real-time execution and interaction. Their appearance changes continuously and also adapts to the public with whom the work interacts; the works exist under a certain form at a certain time. Thus they are not just artefacts but relationships, connecting the virtual environment and the real world.

We describe the IT environment and system required to run the artwork. We also describe the material environment which supports it: the hardware, components and peripheral devices and their networked system. A significant number of digital art installations also require other non-digital equipment and materials which have to be specified. Documentation such as installation plans have to be integrated into our model. The model allows for description of the scenography: space requirements and restrictions (dimensions and characteristics), lightning (light source, luminous atmosphere…), noise level, etc.

The description of the “performed artwork” also needs, in order for our model to be pertinent, to be completed by information on the system that allows the artwork’s relationship to the spectator and participant, and/or the environment. The “system” entity is designed to record the precise role of each interface. The input interfaces characterize the interaction mode: the work can react to a presence, a precise or generic gesture, a position in space, a sound, and so on, and is not restricted to one spectator. The output interfaces describe the perceived reaction of the artwork: the moving image on a display, the sounds from speakers, etc.
The interaction protocol has to be precisely described to keep replay and recreation options open. It can be graded on different levels of complexity that go from the simple action-reaction to an actual relationship between two equal objects with their own behavior: the human and the intelligent digital system [24].

Our model links the performed artwork to an entity that represents the performance event; an exhibition, a festival, a conference, and so on. There is a close relationship between the circumstances of the event and the characteristics of the performed artwork: its hardware changes, its backdrop and setting are modified. In addition, it evolves over time, whether it is partly or entirely emulated or recreated to conform to existing contemporary technology.

To complement the description of the conditions of the artwork’s performance, we created an entity dedicated to the reception of the piece by the spectator: “experienced artwork”. Since the artwork modifies itself anytime it meets an audience, it only exists because someone “read” it, “played” it or with it, “used” it: because someone took part in its full realization. The number of possible experiences is limitless; the implementation of this information entity could include a few representative testimonies.

Textual information on the exhibition materials may be all that documents the artwork if and when it enters the BnF’s collections: the library is not well suited to preserving made-for-measure hardware, or even series of working contemporary computers.

By analyzing the existence of the artwork from its conception to its concrete temporal instances, we can grasp the essence of the piece and thus ensure that its description will be relatively stable through time and with different technologies. The information model is built on the description of the different stages of the artwork’s life: conceptual, realized, performed and experienced. It is both rhizomic and exponential; it is built to allow for the description of the artwork’s consecutive and/or parallel versions. One conceptual artwork can generate an unlimited number of realized artworks, which can generate an unlimited number of performed artworks, which will generate an unlimited number of experienced artworks.

4 IMPLEMENTING THE INFORMATION MODEL

The conceptual model was our first step towards machine-actionable data models. We discuss here the adaptation of our conceptual model into a database and an ontology, and how the specificities of digital art influenced our choices.

4.1 Data models

Our first aim in building a database model and an ontology was to test the conceptual model through different types of implementation. On one hand, this process helped us consolidate our model and define our concepts; on the other hand, it raised new questions. The relational database and the ontology had to be developed in parallel, and they had to encapsulate all the different scenarios offered by a digital work of art that is based on the unpredictable. Ontologies and databases both have their limits, but the most difficult challenge came with the database: to prevent any recurrence or duplication, and moreover to avoid empty tables and keys in the database.

The digital artwork’s lifecycle creates a problem of terminology, in this field of art history where computer technologies are preeminent. It is essential to distinguish the hardware and software used for the creation of the artwork, those used for the initial performance during an exhibition or a festival, those used for later versions of the work, and finally those we may use for purposes of emulation. In the ontology, we get the opportunity to render this distinction by applying the FRBR model. We can link technical metadata to the entities Work, Expression, Manifestation and Item as needed.

In the relational database, we choose to distinguish the hardware and the software as specific tables, and then to separate them from the tables that describe the implementation of hardware and software used by the artwork (“technical set up”) or individual performances of artworks (“event”). This simplifies the
description of versions in the database: using junction tables, software and hardware can be ascribed to a specific setup and/or event. By detailing technical metadata at several points of the artwork's lifecycle — conception, performance and emulation —, we create more opportunities for future research in the fields of art history and computer sciences.

However this level of detail, and the tests we performed in populating the database, gave rise to a new issue: defining the core of the digital artwork. All the artworks in our corpora rely on interactivity, and this cannot be expressed only in terms of hardware and software, even if it is strongly related to both. The table “System” aims at filling this void, by describing the intention governing the conception of the program, the cognitive system within which the artist operated, the actions and reactions of the audience, etc. We believe future researchers will value a standardized description of the type of interactivity to search for artworks in the corpora, even as free text describing the experience of the work remains irreplaceable in certain cases.

![System table](image)

**Figure 7: System table.**

In our ontology, information relating to interactivity type and conceptual system is linked to the “Item” entity, as it is applicable to the preservation context of a specific instance of the artwork.

Implementing a database has led to retrospective changes in the conceptual model: we now distinguish between conception, as an intellectual activity, and production. Our initial “conception” entity was too vague, encapsulating the intellectual and the technical aspects of the created work. We had to make sure information items such as programs, algorithms, software, hardware etc. were not encapsulated in, and thus tied to, the table describing the genesis of the artwork, when we need to use the same information types to describe technical setups or emulation instances. Besides, there would have simply been too many elements in the one table to make it manageable.

One last example of the way our database and our ontology affected the logic of our information or conceptual model is how it allows us to see its flaws. We have difficulties modeling aesthetic currents and notions: the history of digital art is still being made, and at the moment, it is impossible for us to establish definite styles or movements and to fill fields of the sort in a database or an ontology. This underscores the need of agreement in the wider community on several key definitions in digital art, through a thesaurus for example.

### 4.2 What choices for our ontology?

We have additional objectives in working on an ontology derived from our conceptual model: making our information model interoperable, making our information machine-actionable, and evaluating the proportion of data that can be standardized. From this evaluation, we wish to assess how this ratio of free text to controlled data affects the quality of preservation.

Achieving interoperability of the data model depends in part on an accurate assessment of the importance given to the free-text description of interaction and behavior of the artwork. We want the ontology to be precise enough to describe the project’s corpora while opening it up to the possibility of being used in other contexts. Interaction plays an important part in our description, as it is the common denominator to our three collections. But we include in our model the possibility of artworks without interactivity, with Pierre Morelli’s table “Complexity Levels”. The first level is indeed: “The object is unresponsive and without any activity. It is totally inert”.

Manageability and interoperability also depend on choices relating to integrating existing ontologies. We use a fair number of preexisting classes and properties, some widely used and describing general concepts, some specific to information science and digital preservation. In the first category are SKOS and FOAF. In the second are FRBR-o0 [25], Consumer Electronics Ontology and the PREMIS ontology. Used in major research projects such as Biblissima [19] and Doremus, FRBR-o0 responds to our versioning issue, using the Work and Expression classes. As for the PREMIS ontology, it is designed to address “metadata that a digital archive needs to know for preserving objects” [27], which is at the center of our project. We are particularly interested in “premis:Hardware”, describing the IT equipment, “premis:Environment”, as the environmental context

---

19 http://www.biblissima-condorcet.fr/
20 http://www.doremus.org/
and its documentation, both essentials to make digital objects usable and in “premis:ObjectCharacteristics” to express information about the format, size, fixity or else the creating application of our different files composing the artworks.

![Figure 8: Proposed additional classes and properties.](image)

We choose to limit the creation of classes and properties to those absolutely necessary to express the particularities of digital art with an interactive dimension. The terms “Behavior”, “AutonomyLevel” or “InteractionMode” represent these two wishes. They express how an artwork behaves itself and two components of this behavior: its capability of autonomous action (or reaction) and the way it interacts with the audience. In other words, we want to transpose the process and functioning of interaction: how many people it can connect with, what triggers the interaction (breath, heat, movement) etc.

However, some of the data experts we are working with are reevaluating the priority given to class and property reuse. They contend that, except for RDF vocabularies widely used (such as Dublin Core, SKOS, FOAF or even PREMIS), the data needs would be better served by dedicated classes and properties, rather than using other vocabularies in a “Frankenmonster ontology”, as a matter of mastering our data. Indeed, it would be almost safer to manage our own classes and properties rather than using small ontologies that, even if they already exist in the linked data world, are not maintained on a regular basis. Furthermore, we are not aware of any institution that has used the FRBR-oo or CIDOC-CRM ontologies in its processes.

Implementing the conceptual model as a database and an ontology raises the question of the coexistence of controlled vocabularies and full text fields: what degree of standardization is efficient? This depends on several factors: at the art community level, are concepts sufficiently stable to create a precise thesaurus? At the preservation community level, are existing data models adapted to the digital art field? Is there a stable source for normalized information on formats, software and hardware? At the institutional level, are there enough staff to implement the description in a variety of controlled fields? What fields will be useful to index and propose to the users as a request entry point?

The technical and artistic fields in our model are designed with the aim of populating them with controlled vocabularies, but some will have to be full text. Indeed, it does not seem feasible or advisable to express the creation process, or the user’s experience, as a set vocabulary, even if concepts have begun to emerge. We will also have to add documentation (images, videos, testimonies, etc.) to the information package in order to satisfy the digital art historians. A better understanding of the proportion of full text to documents to controlled data in our descriptions will emerge in the third year of the project, as we scale up our tests on the corpora. The right proportion will be the one which satisfies the artist’s, the researcher’s and the archivist’s needs.

### 4.3 Our to-do list

We are aware that our three-year research project will not lead us to a complete data model implementation at the end of 2017. Sorting the partners’ corpora, gathering data and preparing our test corpus has been more time consuming than anticipated. However, we do expect to revise the information model and the ontology within this year, using the input we get from additional artists’ testimonies as well as recommendations from fellow art historians, librarians, and experts from the digital preservation community.

We will try to experiment a mapping from the conceptual model to the BnF’s catalog format, INTERMARC. Description formats and tools are currently being revised to facilitate the adoption of practices compatible with FRBR models, which is an opportunity for our project, as our model has a work / expression / manifestation / item structure. We will also scale up the tests of the data model on the project’s corpora, using the test set which has been defined in the second year of our research.

The closing conference for “Art numérique et postérité” in December will give us a chance to share our results with the art and information science communities. We intend to show artworks from the corpora through virtualization and emulation, and discuss the delta between the extent of our standardized metadata and the aspects of the work that can only be rendered by free text descriptions.

### 5 CONCLUSIONS

We are aware that the mix of artists, art historians and librarians in the project is both an asset and a liability. Therefore, we are tracking another type of result, which we hope to foster: the partners have experienced vast acculturation to one another’s preoccupations. While the members of INRéV had no knowledge of international best practices in digital preservation at the beginning of the project, they are now discussed in the curriculum of the ATI Department. The importance of gathering information on the artwork as it is being made is presented to the masters students who are studying the technologies of digital art creation.

Conversely, the library staff had little knowledge of the significant properties of interactive digital art, which made defining the information packages and evaluating the quality of emulation and virtualization difficult. A first training session with the library assistants in charge of facilitating access to the documents in the audiovisual research reading room has allowed us to consolidate the procedure for giving access to the emulated
artists’ CD-ROMs collection. Others will follow as more corpora are ingested in the BnF’s archive.

In the future, we will have to monitor our information and data models to ensure that they remain applicable as digital art evolves. We have built the models to encompass a large array of digital art creations, and have tried to articulate interaction systems, hardware and software in a way that allows for fast-evolving technologies. Yet we are keenly aware that preservation is a continuous process and that we have to balance the requirements for a stable model with the multifaceted and ephemeral nature of digital art.

ACKNOWLEDGMENTS

We would like to thank the team who made this project possible: Élodie Bertrand, Marie Saladin, Jean-Philippe Humbiot, Bertrand Caron, Sébastien Peyrard, Pierre Choiffé, Françoise Leresche at the BnF; Marie-Hélène Tramus, Vincent Meyrueis, Edmond Couchoth and Michel Bret at Paris 8 University. Our interns brought new skills to our research: Catherine Champenois, Juliette Fattal, Tigran Ghaplanyan, Catherine Helmer, Yu-Ci Huang, Gaëtan Lemaître, Victor Martin, Julie Milhiet, Nola N’Diaye, Laetitia Perez.

We received invaluable help from the Arts-H2H Labex team and our external collaborators: Pauline Cellard, Peter Stirling, Mehdi Bourgeois, Jean-Luc Soret. Finally we would like to thank the artists and researchers, who are the very heart of our project.

REFERENCES


Collections As Data: Preservation to Access to Use to Impact

Short Paper

J. Mears
Library of Congress
United States
jame@loc.gov

A. Potter
Library of Congress
United States
abpo@loc.gov

K. Zwaard
Library of Congress
United States
kzwa@loc.gov

ABSTRACT
The Library of Congress has many millions of objects it has either digitized or acquired in born-digital formats. The first priority is to preserve, process, and provide access to digital collections via loc.gov. Most of these collections are free and available to anyone around the world, and their existence online represent the culmination of a huge amount of time, effort, and expertise. The Library of Congress has embarked upon an effort to maximize the use of our digital collections, promote innovation at the institution, and grow national capacity for cultural memory. This paper will describe how developing and executing a program around promoting collections as data has enabled the expansion of internal and external partners, deepened the exploration and value proposition of digital collections, engaged a broader community, and provided new skills to librarians who are working with collections as data. The programs that will be outlined will include the development of a digital scholarship lab which seeks to provide support for computational engagement with collections, testing different training models for digital skill building, and launching Innovator-in-Residence programs that showcase how data analysis techniques can change digital scholarship and digital curation landscapes. These partnerships and pilot projects build on the digitization and preservation efforts of the Library by broadening the scope of engagement and providing compelling use cases that amplify the impact of digital collections.

KEYWORDS
Data Analysis, Digital Scholarship, Outreach, Education & Training, Impact of Preservation

1 LAUNCHING NATIONAL DIGITAL INITIATIVES
The National Digital Initiatives division was founded in October 2016 and operates in the National and International Outreach service unit under National Programs. The core staff came from the retired National Digital Information Infrastructure and Preservation program after a realignment of the Library of Congress service units. The founding principle of NDI is to honor the knowledge, skills, and effort of the Library of Congress staff who build our digital library. Much of this is invisible work. The acquisition, transfer, processing, and preservation of petabytes of digital collections is central to the Library’s mission and purpose and represent an incredible asset that belongs to the American public. NDI also inherits the network building and community engagement success of the NDIIPP program.

NDI’s mission is to broaden awareness of the Library’s innovation and use of its digital resources through outreach and external partnerships. We seek to work closely with other service units and divisions to reach new audiences, spark innovative projects, and work together on shared problems. To execute this strategy, three program areas were defined: 1) Expand the use of Library of Congress digital resources, 2) Incubate, encourage, and promote digital innovation, and 3) Grow national capacity for cultural memory.

2 ENABLING NEW USES OF THE DIGITAL COLLECTION

2.1 Collections as Data
The Library of Congress and other libraries have been serving digital collections online for over a decade. However, the model for which these resources are accessed largely recreates an analog experience of reading a page at a time or browsing through a collection of photographs. With modern computing power and the emergence of data-analysis tools, these digital collections can be explored more deeply and reveal more connections. In order to take advantage of computation, collections must be made available in a form that can be recognized by computers, a transformation not unlike moving from creating a card catalog to the MARC record schema. Obviously, the collection must be digitized but often more than a just a digitized image is needed to make full use of the object. Optical character recognition that reflects underlying text or metadata that describe the content in a structured way, like using ISO date and formats, latitude and longitude coordinates, or tags according to a taxonomy, are data
that widely available tools can use to create visualizations and perform analysis.

For users that want to work with a large number of digital files and download them in bulk rather than one at a time, The Library of Congress provides processed bulk-data-download derivatives for two collections. The MARC Distribution Service provides all of The Library of Congress catalog records (https://www.loc.gov/cds/products/marcDist.php) and the National Digital Newspaper Program provides the raw OCR text of newspaper pages from the Chronicling America project (http://chroniclingamerica.loc.gov/ocr/). There are additional collections for which there is an approximation of bulk data access for anyone with the technical capacity for scripting and the aptitude to understand and crawl an API or a site. Affixing the text "?fo=json" to the end of many URLs in https://loc.gov results in a JSON API that can get some researchers started (see http://www.loc.gov/pictures/?fo=json). XML sitemaps are available for, currently, 277 digital collections (https://www.loc.gov/collections/sitemap), providing information such as the item URLs, last modified date, and frequency of updates (e.g. https://www.loc.gov/collections/franklin-pierce-papers/?c=1000&sp=1&fo=sitemap). Similarly, https://id.loc.gov has a number of formats available for crawling a large number of datasets, but the bulk download of data is not generally available.

The World Digital Library (https://api.wdl.org) has a well-documented API which serves a similar purpose for those collections. There are collections which do not have an API, documented or otherwise, that allows crawling or web harvesting. Some very motivated researchers have previously used this kind of interaction to build datasets, or to extract particular curated selections, and to make them available elsewhere.

Locations away from our main website occasionally have capabilities like one of those described above. For example, the Library of Congress participates in the Flickr Commons (https://www.flickr.com/commons) which comes with some capabilities of using the Flickr API for downloading collections.

These few examples represent a small percentage of the roughly 300 collections and approximately 2 million items, in multiple content formats, available on loc.gov but not optimized for anything other than interactive browsing and various degrees of searchability. One of the goals of NDI is to promote standards and practices around providing access to collections as data so that the full value of computation can be leveraged to bring even more awareness of the knowledge and creativity contained in the world’s libraries.

2.2 Exploring Digital Scholarship

Emerging disciplines — like data science, data journalism and digital humanities that take advantage of new computing tools and infrastructure — provide a model for creating new levels of access to library collections. Visualizing historical events and relationships on maps, with network diagrams and analysis of thousands of texts for the occurrence of words and phrases are a few examples of what’s possible. NDI is actively exploring how to support these and other kinds of interactions with the Library’s vast digital holdings.

Michelle Gallinger and Daniel Chudnov were asked by NDI to study how libraries and other research centers have developed services that use computational analysis, design and engagement to enable new kinds of discovery and outreach. Their report (PDF)[1], was just released. For the report, they interviewed researchers and managers of digital scholarship labs and worked with Library staff on a pilot project that demonstrated how the collections could be used in data analysis. This work resulted in concrete recommendations to the Library on how to approach setting up a Lab at the Library of Congress. These recommendations could also be helpful to other organizations who may be thinking of establishing their own centers for digital scholarship and engagement.

The recommendations from Gallinger and Chudnov would warrant a paper on its own but we will review some of their feature recommendations: First, that the Lab should be service oriented to outside users. Second, that the Lab support users to grow and develop their research. Third, that the Lab be an educational space, and fourth that the Lab enable organizational transformation. As research, scholarship, and resources become more digitally based, the services provided by the Lab will, in the beginning, be groundbreaking but eventually become part of the normal services offered by the Library. It is this transformation of how librarians perform reference services for the digital collections—services that include identifying available data, working with that data to deliver specific a specific corpus, establishing provenance for the data, and helping to answer research questions with the data—that NDI hopes to cultivate in a digital scholars lab.

Concurrently, NDI is co-leading an internal group charged with studying how the Library of Congress can programatically enable digital scholarship for its collections. The group is completing a paper in 2017 that describes the challenges and opportunities inherent in the Library of Congress’ collections and operations. Much of the information in the Collections as Data section of this paper comes from that group’s efforts.

2.3 Engaging Communities

Using Gallinger and Chudnov’s report and recommendations as guidance, an on-site digital exploration center is an opportunity to open the Library’s digital holdings up for enhanced scholarly inquiry. It would also serve as a point of engagement for more general users for topic or visual exploration. It would serve numerous communities:

- Scholars interested in using Library of Congress digital collections as data to create visualizations, explore
digital scholarship, and pursue computationally-assisted research.

- Teachers and students who want to use Library of Congress data in their STEM or humanities classwork and projects.
- Artists, writers and other creatives who want to access our public domain content to remix and reuse the Library’s digital materials in new ways.
- On-site and virtual users who want bulk access to digital collection data for analysis or use in other applications.

In developing coordinating services, the Library will work collaboratively to design orientation and introductory classes on digital scholarship methods and tools to help introduce the concept of using library collections as data. The courses will be in parallel to professional development trainings for Library staff who will be delivering services, especially reference services, in new ways.

2.4 Skill-building

Providing access to digital library collections as data requires new skills for librarians and using digital library collections similarly requires new skills and modes for researchers. NDI is exploring how to offer educational opportunities for both of these groups to gain these new skills. In February 2017, 40 librarians, archivists and data wranglers came to the Library of Congress to learn advanced skills in managing digital collections. The Library hosted the Software Carpentry workshop, inviting staff from the institution, the DC Public Library and other federal libraries for hands-on learning in the programming language Python, the version-control software Git and the command-line interface Bash. Software Carpentry is a volunteer, non-profit organization that provides short, intensive workshops to help researchers automate tasks and manage information. It started with scholars in the physical sciences who found that traditional graduate programs were not preparing them for the challenges of working with data for their research products. Software Carpentry workshops have lately been adapted for social sciences, the humanities and libraries.

Attendees from the workshop immediately saw a way to apply their new skills. “I can see an opportunity to use scripts to improve researchers’ experience in the reading room,” said Kathleen O’Neill, a senior archives specialist in the Manuscript Division. “For those researchers with limited experience with digital collection material, we could provide a library of simple scripts to search, analyze and report on the born-digital collection material.”

Later in the year, NDI partnered with George Mason and George Washington University Libraries to host a program titled “Hack-to-Learn”. There is clear demand for hands-on computational experience yet librarians are often under represented at events like hackathons. So the organizers worked together to develop an inclusive hackathon that utilizes the skills librarians already have and introduces them to low or no-cost computational tools to explore digital library collection data sets. Over two days, 61 librarians and researchers attended, most with no prior programming experience. They were given instruction around MALLETS, a topic modeling tool, Gephi, a network analysis visualization tool, OpenRefine, a data editing tool, and Carto, a mapping tool. According to post-event surveys, attendees confirmed that the event was valuable in orienting them to computational methodologies and new research and processing possibilities for their digital collections.

Developing and promoting a program to develop skills around collections as data has enabled the expansion of internal and external partners, deepened the exploration and value proposition of digital collections, engaged a broader community, and provided new skills to librarians who are working with collections as data.

3 SHOWCASING DATA ANALYSIS TECHNIQUES

3.1 Innovators in Residence

While 3 million people access the collection every month on the loc.gov website, there is untapped potential of using the digital collections in other ways, be it for data analysis to support digital scholarship, by an artist seeking to remix and reuse content to create new art or commerce, or a journalist in need of authoritative data for in-depth reporting.

NDI has launched an Innovator-in-Residency program to support innovative uses of our collections and partnerships with universities, scholarly societies, artists, corporations, and other organizations. A wide variety of programs would be included in a broad residency program:

- A program that would support the work of digital humanities scholars at the Library of Congress, which have been successful at other large libraries.
- A challenge grant program that would run contests for innovative uses of digital collections and make small awards to winners, modeled after Federal Data Challenges.
- A program that would support software developers to come to the Library and build tools and services that make innovative uses of Library collections.
- A fellowship program to support hybrid teams to use Library of Congress collections and data to support information for the American people. A journalist and a social scientist working with a historian working on a project that will be published widely in a newspaper, magazine or book.
- A program to support artist to create art based on the Library’s collection.

In 2017 two multimedia artists were selected to create digital art pieces based on the Library of Congress digital collections. In
addition to creating the works, they will present their art to the public and Library staff in a workshop.

3.2 Crowdsourcing application

NDI piloted the Innovator-in-Residence program at the Library with an opportunity for staff to apply for a short-term assignment. Two staff members were selected. One, Chris Adams, focused on exploring automatic image identification and fleshing out the concept of labs.loc.gov. The other, Tong Wang, created a proof-of-concept crowdsourcing application. Tong wanted to show how we could leverage open source tools created by other members of our community (specifically, the Scribe http://scribeproject.github.io/ transcription framework) and the Chronicling America API (http://chronlingamerica.loc.gov/about/api/) to engage the public with library collections and to enhance the usability of our digital material. The application he created is called 1000 Words and it invites users to identify cartoons, photographs, and other images in historic newspapers and to describe them. This would allow those images to be searched for the first time and for data sets to be created. For example, the Library could offer to researchers a collection World-War-I-era cartoons, which previously would require flipping through thousands of newspaper pages to collect.

3.3 Events

We have hosted several events in an effort to demonstrate what we mean when we talk about collections as data and provoke exploration of our collections computationally for our staff and members of the public.

We invited programmers, artists, entrepreneurs, activists, researchers and librarians finding new ways to connect with digital collections to share best practices and lessons learned during Collections as Data: Stewardship and Use Models to Enhance Access, a public conference held at the Library of Congress in September 2016. From tracking the popularity of bible verses in historic newspapers to mapping everywhere you've ever walked on your smartphone, speakers such as Jer Thorp from the Office for Creative Research, Pinboard founder Maciej Ceglowski, and Director of the Five College Digital Humanities Initiative Marisa Parham discussed the methods, skills, ethical considerations and community participation necessary to find new meaning in digitized and born digital cultural heritage collections. The topic resonated with many. The event’s hashtag #AsData trended #2 on Twitter the day of the event, receiving over 8500 livestream views. The more than 450 people came, representing over 15 different countries from universities, public libraries, galleries, federal institutions, and newspapers such as the Washington Post and New York Times. Tweets and verbal feedback revealed attendees appreciated the holistic, nuanced, cross-disciplinary consideration of data.

The program was so successful that we will host a second Collections as Data Symposium in July, showcasing the impact of computational research through a series of successful case studies across disciplines and practitioners.

To facilitate learning and discovery, we were the host site for Archives Unleashed 2.0 in 2016, a web archive datathon series facilitated by Matthew Weber, Ian Milligan and Jimmy Lin in which teams of researchers used a variety of analytical tools to query web-archive data sets in the hopes of discovering some intriguing insights before their 48-hour deadline is up. Aside from hosting, the Library of Congress participated on the Hackathon teams and by providing collections for the teams to work with.

In an effort to centralize participatory opportunities for local District of Columbia residents, develop best practices for community engagement, and support sharing of information across DC cultural heritage institutions, NDI is co-hosting a Citizen Public History Fair in September with NARA, the Smithsonian Transcription Center, DC Public Library Special Collections, United States Holocaust Memorial Museum, the Folger Shakespeare Library, the Lincoln Theatre and the Anacostia Community Museum. The fair will showcase or demo volunteer project opportunities including digital crowdsourcing initiatives and feature lightning talks from curators and researchers. Following the event, a series of crowdsourcing programs by our institutions will take place throughout the year in an effort to develop a robust community of practice around historical resources.
In 2018, NDI will also serve as co-host of several high-profile meetings at the Library of Congress that support the collections as data vision. These include the 2018 Code4Lib Annual meeting, the 2018 IIIF Annual meeting, a HathiTrust Research Center workshop, and a workshop in conjunction with a grant funded DARIASH-EU project on digital heritage and culture.

4 CONCLUSION
The Library of Congress has millions of digital items available to the American Public via the loc.gov web site. Millions of people from all over the world visit our vast collections online. An additional 1.6 million people visit the Library of Congress Jefferson building each year [2]. The efforts of the National Digital Initiatives to support the practice of providing data level access to digital collections provides opportunities to enhance the preservability of digital collections and increase their usefulness to the communities we serve. When digital collections are leveraged as data, the public spaces, technical environment, training, tools and expert support are all modernized and made more relevant. These partnerships and pilot projects described above build on the digitization and preservation efforts of the Library by broadening the scope of engagement and providing compelling use cases that amplify the impact of digital collections.

REFERENCES


Research Data Management Organiser

A tool to support the planning, implementation and organisation of research data management

Claudia Engelhardt
University of Applied Sciences Potsdam
Faculty of Information Sciences
Kiepenheuerallee 5
14469 Potsdam
Germany
claudia.engelhardt@fh-potsdam.de

Harry Enke
Leibniz Institute for Astrophysics
Potsdam (AIP)
An der Sternwarte 16
14482 Potsdam
Germany
henke@aip.de

Jochen Klar
Leibniz Institute for Astrophysics
Potsdam (AIP)
An der Sternwarte 16
14482 Potsdam
Germany
jklar@aip.de

Jens Ludwig
Staatsbibliothek zu Berlin
Preußischer Kulturbesitz
Potsdamer Straße 33
10102 Berlin
Germany
jens.ludwig@sbb.spk-berlin.de

Heike Neuroth
University of Applied Sciences Potsdam
Faculty of Information Sciences
Kiepenheuerallee 5
14469 Potsdam
Germany
neuroth@fh-potsdam.de

ABSTRACT

The project Research Data Management Organiser (RDMO)\(^1\) develops a tool to support the planning, implementation and organisation of research data management. The multilingual open source tool can be installed locally and adapted to institutional or discipline-specific needs with regards to contents. It provides interfaces to institutional authentication procedures. Key features of the first version, released in April 2017, include:

- the ability to continuously update and augment the information in the course of a project
- access for different stakeholders, such as researchers, project coordinators, the IT department, data managers etc., with customized views
- export formats for different purposes including data management plans (DMPs) according to funder requirements.

Planned future developments include, among other things, features to support actual data management, e.g. tasks (with deadlines and a reminder functionality) that can be linked with dedicated stakeholder responsibilities.

KEYWORDS

research data, data management planning, active data management, tool, organiser

1 INTRODUCTION

The basis for a successful long-term management and provision of digital research data is a thorough research data management throughout the whole project lifetime. The goal is to get FAIR data – data that are findable, accessible, interoperable and reusable [6]. Data management is not solely a responsibility of researchers, but also of research institutions that have to provide the necessary technical infrastructure, consulting and support. The project Research Data Management Organiser (RDMO) developed RDMO as a tool that supports both researchers and institutions in the planning, implementation and organisation of research data management.

RDMO was funded by the German Research Foundation DFG and, in its first phase, ran from November 2015 to April 2017. The paper outlines the starting points, summarises the conceptual and development work done so far, describes the tool’s features and fields of application and gives an outlook on future developments.

2 DATA MANAGEMENT PLANNING – THEORY AND PRACTICE

2.1 Data Management Plans

The foundation for FAIR data is laid very early in the research process, namely in the planning stage. For this reason, data management plans have become a crucial element of data management policies of funding agencies, and subsequently for universities and other research institutions. DMPs vary in extent and detail; typical elements include statements about what data will be created with which methods, applicable policies, plans for sharing and preservation, data curation measures and respective

\(^1\) http://rdmorganiser.github.io/en/
responsibilities, ownership and access conditions, restrictions (e.g. for legal or ethical reasons) and required resources [2, 3]. In recent years, tools to support the creation of DMPs according to funder requirements have been developed, the most well-known and popular of which are DMPonline by the Digital Curation Centre (DCC) and DMPTool by the California Digital Library (CDL).

2.2 Active Data Management
But creating a plan is just the first step – since a plan does not help much if the outlined data management measures are not put into practice. Data management is an ongoing effort which includes adjusting the DMP and the associated activities if necessary. To emphasize this continuous character, the term active data management has been introduced.

The growing awareness of the need for such an active data management shows, for example, in the Research Data Alliance (RDA) Interest Group on "Active Data Management Plans" 

4 the requirement for EU Horizon 2020 projects to update their data management plans in case of significant changes [5] as well as in the joint activities of DMPonline and DMPTool to "reposition DMPs as living documents" [11] and "integrate[ e] them into the broader ecosystem of data management infrastructure" [10]. The same motivation inspired the RDMO project.

2.3 Why another tool?
Both DMPonline and DMPTool, at least currently, have a strong focus on the requirements of research funders. Their main purpose is to discover, edit and fill in DMP templates of funding organisations, mainly from the UK (in the case of DMPonline) and the US (in the case of DMPTool). In most instances, these DMPs will be forgotten after the submission. Although active data management is considered in their future plans, it is not supported by the tools at the moment.

Both tools are centralised web applications. This is associated with the transfer of potentially sensitive information off-site. Also, it offers institutions only limited possibilities for customisation. Furthermore, DMPonline and DMPTool are not re-deployable without investing considerable effort.

All of these points are addressed in the RDMO concept.

3 CONCEPTUAL DESIGN
The design process of RDMO was guided by the following aims. The tool should:

- support the data management throughout the whole project lifetime
- enable users to gather and organise all information necessary for a sustainable data management
- involve all relevant stakeholders
- be locally installable and configurable.

As crucial features required to realise these aims we identified:

- collaborative editing
- specific roles, rights and views for different stakeholders
- input via a structured interview; skipping of redundant or unnecessary questions based on given information
- output of gathered data in various forms, amongst these textual data management plans for different funder requirements
- possibility to adapt the contents (questions as well as answering options)
- tasks and reminders
- easy application and administration in different contexts (e.g. university, research institute, joint research project).

The conceptual design was developed mainly on the basis of three activities: 1) the further development of previous work to design a basic, generic questionnaire, 2) desk research, user tests and interviews on discipline-specific requirements and content, and 3) user stories to model the requirements and the associated functions for different stakeholders.

During the whole process, we were supported by a number of projects and institutions who tested the tool at different stages and gave us valuable feedback and input.

3.1 Basic questionnaire
The major source for the development of the basic questionnaire was the WissGrid-Leitfaden zum Forschungsdatenmanagement [9]. This is a research data management manual and checklist produced by the project WissGrid (2009-2012). In Germany, it has become one of the standard works on research data management. In addition, we also looked at DMPs and similar material of other tools, institutions or funding agencies as well as checklists that one of the project members had developed for the use in consultations with researchers while working as a data manager at a Max Planck Institute. The latter already included tasks associated with certain questions to stress the active element in data management.

The basic questionnaire of RDMO is designed to cover all relevant data management aspects. In particular, it comprises of the following areas:

- general information about the project (including data management requirements from policies, e.g. of funders or the home institution)
- content classification
- technical classification
- data usage
- data storage and security
- collaborative work
- quality assurance

---

2 https://dmponline.dcc.ac.uk/
3 https://dmp-tool.org/
4 https://www.rd-alliance.org/groups/active-data-management-plans.html
5 See section 3.3 for a description of stakeholder roles and requirement.

---
3.2 Discipline-specific requirements

To find out more about the discipline-specific requirements of such a tool, the project investigated two disciplines as an example: the social sciences and astrophysics. In a first step, we undertook desk research on data management practices and requirements in these two fields [the main sources being 4, 7, 8]. The results were then verified in expert interviews with a data manager of TwinLife\(^7\), a 12-year longitudinal twin study on the development of social inequality, and the working group on optical solar physics\(^8\) at Leibniz Institute for Astrophysics Potsdam (AIP). The interviews took place in May and June 2016 and included testing a draft version of the generic RDMO questionnaire. Finally, the results were also discussed in a breakout session of an RDMO workshop held in Potsdam in July 2016, where our previous findings were verified.

The results suggest that – on the mid-range or coarse-grained level of granularity which most DMPs or tools, including our own, cover – the need for the customisation of questions according to specifics of certain disciplines is given only for single questions or topics. In astrophysics, for example, usually no personal data or data protected by copyright is gathered or produced, so that these topic areas can be omitted. For large, quantitative studies in the social sciences, it is common to outsource the data collection and parts of the data preparation to external survey institutes. Therefore, a DMP tailored to the social sciences should address the topic of external survey institutes and the steps in the research process these are responsible for. But as mentioned before, in general this applies only to a few selected areas of the DMP. A discipline-specific customisation of larger parts or the whole DMP would only be useful on a very detailed, fine-grained level. Then, however, the questions become so specific to a narrowly defined set of use cases that this does not seem to be a sensible strategy for the vast majority of user groups.

With respect to the answering options as well as help texts and links for further information based on a DMP or RDM tool, however, our interview partners as well as the workshop participants saw great potential benefit in customised options, e.g. by preparing sets of options that refer to data types, methods, tools, standards, vocabularies etc. commonly used in the discipline in question or suggesting repositories for certain disciplines or data types where possible. This would make it more comfortable for researchers to answer the questions. In addition it would help to standardise the answers which then would make it easier for an institution or IT department to collect information across a number of projects to, for example, assess what infrastructure resources (e.g. storage) need to be provided. However, the design of questionnaires tailored to different disciplines or sub-disciplines is a task that cannot be accomplished by a project like RDMO. It can best be tackled by the scholarly communities in question, and RDMO offers the means to implement this.

3.3 Stakeholders and user stories

To find out more about the requirements of the different data management stakeholders, we collected user stories. Instead of just simply postulating requirements, user stories reflect the perspectives of different actors (in my role as …), describe what activity they want to use the tool for (I want to …) and then indicate the purpose (to get the benefit of…) to put the activity into context [1].

In total, we compiled about 70 user stories for the following stakeholder roles:

- author (most common)
- infrastructure provider (second most common)
- superior (third most common)
- data manager
- guest
- manager
- IT administrator
- developer
- IT support
- reviewer

We cannot go into detail at this point, but shortly present the categories that were aggregated from the user stories, each with one or two examples.

Collaboration

- As author, I want to invite other persons to my DMP as reader or author, so that they can contribute.
- As superior, I want to be able to read and approve DMPs, to fulfil my controlling duties.

Usability, input assistance and templates

- As author, I want to use templates and recommendations from my institution and funding organisation, to know what to focus on to fulfill their requirements.
- As author, I want to have predefined selections of useful, correct and standardised answering options where possible, to save time because then I do not have to think of my own ones first.

Versioning

- As author, I want to access older versions of my DMP, to undo incorrect entries.

---

\(^7\) http://www.twin-life.de/en

Adaptability of questions and answering options

- As infrastructure provider, I want to be able to define selections of useful, correct and standardised answering options, in order to be able to aggregate user inputs for easier analysis.

Logic of questions

- As author, I do not want to answer questions that can be identified as irrelevant on the basis of my previous entries, so that I can concentrate on the questions that are important for my project.
- As infrastructure provider, I want to offer questions and answers in different levels of granularity, so that plans can be made according to the needs, abilities and most useful level of detail for a certain project.

Task administration

- As infrastructure provider, I want to connect certain questions and answers to data management tasks for different actors and roles, so that all stakeholders are aware of their tasks and responsibilities in the process.
- As infrastructure provider, I want to be able to view the answers for relevant questions, in order to plan accordingly (e.g. how much storage space needs to be provided).

Review

- As reviewer, I want to have access to the information in RDMO, to evaluate if the planning and implementation of a project’s data management are / were carried out properly.

APIs and export functions

- As author, I wish that information I or others entered in this or other systems can be migrated, so that I do not have to enter everything from scratch again.
- As author, I want to be able to export my answers in a machine readable form and link them to other systems, to be able to use the information entered in other RDMO installations (at other institutions).

4 TECHNICAL DEVELOPMENT AND FUNCTIONS

RDMO is implemented as a web application based on the Python framework Django\textsuperscript{12} and the JavaScript framework AngularJS\textsuperscript{13}. It is an open source tool licensed under Apache Version 2.0\textsuperscript{14} and is available on GitHub\textsuperscript{15}.

The information in RDMO is organised along projects. It is up to the users to define what project means in their specific context. In most cases, a project in RDMO will represent a ‘real life’ research project, but it might also relate to a subproject, one particular survey or study or a range of studies inside a project, or others.

The RDMO implementation is multilingual, currently available languages are German and English. It is designed to support internationalisation, so that more languages can easily be implemented in the future.

In the project life span of 18 months we achieved the goals set in the project proposal and even realised some additional features. However, not all requirements and desirable features that were identified by the user stories and feedback from testers could be implemented by the end of the project in April 2017.

In the following, we give a brief description of the features of the first version, released at the end of April 2017.

4.1 Input

Information can be entered in a web interface and edited by different stakeholders. It is collected via a structured interview that guides users through all relevant topic areas. Based on the answers given, redundant or irrelevant questions are skipped. Depending on the type of question/answer, different widgets are used (e.g. radio buttons, check boxes, drop-down lists, rulers, free text). Controlled vocabularies or predefined sets of answers are used when available. However, as the basic questionnaire is generic, this is extensible in questionnaires tailored to particular disciplines or fields of study. Snapshots can be made at any time to freeze and document the state of the information about a project at a given point in time.

4.2 Output

There are several kinds of output. On the project level, the information previously entered can be aggregated into textual views (e.g. DMPs to be used for project proposals). These views can be defined as required. There is a to-do list of data management tasks to be performed and a reminder function that notifies the responsible parties of upcoming assignments.

On a departmental or institutional level, information can be aggregated across projects, which is useful for a number of purposes, e.g. to derive the demand regarding infrastructure resources or to get an overview of the types and amount of data produced at the institution.

4.3 Setup and operation

One aim was to design RDMO in a way that makes it easy to install and to customise in various contexts (e.g. universities, research institutes, departments of universities or research institutes, libraries, larger joint research projects, research groups etc.). Accordingly, the tool is flexible in several respects. In order to allow institutional branding, the user interface can be freely customised. The content can also be tailored completely to suit context-specific needs. This applies to the questions and answering options as well as to help texts, templates for the output of DMPs and tasks.

Another important point is the ability to integrate it into the local infrastructure. RDMO can be linked to the institutional

\textsuperscript{11} For the whole chapter see also http://rdmorganiser.github.io/en/software/
\textsuperscript{12} https://www.djangoproject.com/
\textsuperscript{13} https://angularjs.org/
\textsuperscript{14} https://www.apache.org/licenses/LICENSE-2.0
\textsuperscript{15} https://github.com/rdmorganiser/rdmo
authentication and authorisation infrastructure, namely an LDAP system or a Shibboleth federation.

5 OUTLOOK

In addition to the demo version provided by the RDMO project (hosted by Leibniz Institute for Astrophysics Potsdam (AIP) and available at https://rdmo.aip.de), several institutions have set up local instances, among them Karlsruhe Institute of Technology, University of Stuttgart, University of Konstanz, University of Duisburg-Essen, and the Göttingen eResearch Alliance. Several more have expressed interest in doing so.

Further development of RDMO is planned in the future, the second project phase will start in autumn 2017. Some of the aforementioned institutions will closely collaborate with the project in this respect.

The range of functionalities will be extended, e.g. by the possibility to upload relevant documents, such as codebooks, metadata documentation or applicable guidelines, and a commenting functionality. Existing functions will be refined. The focus of these projected activities lies on features supporting the implementation and organisation of the actual data management throughout the project lifetime, e.g. roles, tasks, modules for cost estimation and ingest-process. The interoperability both between different RDMO instances and with external services such as re3data18 or research information systems will be enhanced. Set-up, upkeep and integration in different institutional environments will be made easier with standardised installation, integrated maintenance mechanisms and extended support of authentication and authorization procedures. Exchange and cooperation with different stakeholder groups will be continued and intensified with the objective to build an active user community dedicated to the continuous joint distributed development as the basis for a sustainable future of RDMO. This includes tutorials and workshops for different user groups as well as the extension of our cooperation with existing data management and infrastructure initiatives

REFERENCES


18 http://www.re3data.org/
ABSTRACT
Software is transient: it’s the data that matter. Anyone who works with software understands that applications come and go, and even those that last many years go through numerous upgrades and re-architectures as they try to keep up with the latest technological advances. This flux creates a problem: how do we efficiently move our data through changing software and version upgrades as losslessly as possible? The Fedora community recently confronted this problem, which led to the development of a tool that could have broad utility for data migrations.

Fedora is a flexible, extensible, open source repository platform for managing, preserving, and providing access to digital content. Fedora has gone through several major version upgrades since its initial conception almost 20 years ago, most recently with the move to Fedora 4. Fedora 4 is a complete software re-architecture, which necessitates a data migration from previous versions of Fedora. This experience led the Fedora community to focus on making it as easy as possible to get data into and out of Fedora in standard formats, an effort that culminated in the creation of an import/export utility that standardizes on RDF and the BagIt specification to transfer data between versions of Fedora as well as external preservation systems. This effort coincides with a recently chartered Research Data Alliance working group with a mandate to explore technologies and standards for interoperability between repository platforms and make recommendations on this topic. The group is exploring the Fedora import/export utility as a possible basis for broader interoperability between diverse repository platforms by building on this standard import/export functionality. This paper will provide an overview of the import/export efforts that have taken place so far, and discuss how we might achieve broader interoperability and ease data migrations through the work of the RDA Research Data Repository Interoperability working group.

KEYWORDS
Fedora, repository, open source, migration, data

1 INTRODUCTION
Fedora – the flexible, extensible, durable object repository architecture – is both a concept and a software implementation. The concept of Fedora was first proposed in a research paper by Sandy Payette and Carol Lagoze in 1997. This paper lays out a model for digital objects in an open repository architecture that would later be implemented as software: “A fundamental requirement of an open architecture for digital libraries is a reliable and secure means to store and access digital content. FEDORA is a digital object and repository architecture designed to achieve these requirements, while at the same time providing extensibility and interoperability.” [3] Fedora was envisaged as a flexible and extensible architecture that could interoperate with new and existing systems and services. These ideas were present in the original version of the Fedora software, and they continue to be represented in the current version to this day.

However, like any other software application, Fedora will not be around forever – at least not in its current form. Fedora has already gone through a major re-architecture from version 3 to 4, necessitating a data migration for any existing users. This highlights a problem that memory institutions will continue to face over the years: how to maintain the resources and support that will be required for inevitable data migrations as software applications change or are phased out.

2 TRANSIENT SOFTWARE
If there is one thing all memory institutions should be able to agree on it is this: the data, not the software, are what matter most. Of course, institutions need software to manage, preserve, and provide access to their digital collections, but it is the collections themselves that matter; the software will inevitably need to be replaced, continually, over time. The Fedora community grappled with this basic truth recently as we made the change from Fedora 3 to 4, and determined that an in-place upgrade would not be possible due to deep architectural changes in the codebase. Thus, the community embarked on a quest to migrate their data from Fedora 3 to 4, but this turned out to be more difficult than we originally imagined.
Anyone who has undertaken a data migration will tell you that they are always more difficult than originally planned. There are several reasons for this: different data models and formats from the old system to the new one, different metadata standards, the time it takes to move large amounts of data, and, perhaps most difficult, accounting for all the inconsistencies in the source data. Over the years, through different curators and standards, data inevitably becomes inconsistent; new metadata standards are adopted, custom metadata fields are created when the current standard doesn’t seem to support a given need, names are misspelled, etc. Just figuring out what you actually have and normalizing (to the extent possible) all the data is probably the most time-consuming part of a migration.

3 MIGRATING DATA

Once your data is normalized you need some way to get it out of the source system and into the new system. This can be challenging as different systems store data in different ways, and there are no universal tools for doing this kind of work. This migration use case was one of the primary drivers behind the import/export tool that the Fedora community has developed.

As a community-supported, open source project, Fedora is focused on adopting widely used standards and making data easy to get in and easy to get out in a standardized way. This goes back to the earlier point about the primacy of the data – Fedora will not be around, at least in its current form, forever, so adopters should have a relatively easy way to get their data out in a standard format that doesn’t require Fedora or any proprietary software to access and understand.

New Fedora features are developed by and for the Fedora community; as such, each new feature starts with an expression of interest, followed by meetings where stakeholders gather to discuss potential use cases and functionality. In the case of the import/export tool, a set of use cases were added to the Fedora wiki and vetted by stakeholders. These use cases include the following [1]:

1. Transfer between Fedora and external preservation systems, such as APTrust, MetaArchive, LOCKSS, DPN, Archivematica, etc
2. Package [Export] the content of a single Fedora container and all its descendant resources
3. Transfer between fedora instances or (more generally) from Fedora to an LDP archive
4. Load [Import] the contents of a package into a specified container.
5. Round-tripping resources in Fedora in support of backup/restore
6. Round-tripping resources in Fedora in support of Fedora repository version upgrades
7. Batch loading arbitrary sets of resources from metadata spreadsheet and binaries
8. Import or export containers or binaries using add, overwrite, or delete operations.

Once the use cases were agreed upon, functional requirements were derived and organized by priority. These requirements were then assigned to code sprints and community developers signed up to work on each sprint. This process allowed the community to break the effort down into manageable chunks and produce milestone releases that could be tested and verified by stakeholders. The testing results could then be rolled back into the development process to be worked on during the next sprint. After following this process for several cycles, the development team completed an initial release of the tool [4].

The import/export tool is a command line utility that can be supplied with parameters to both export content from Fedora and import content back into the repository. The default export format is RDF, with serialization options that include JSON-LD, XML, n-triples, N3, Turtle, and plaintext. Resources can also optionally be exported with any binary files they may be associated with to achieve a complete representation of the repository contents. Additionally, the BagIt [2] packaging format is supported for both exports and imports. This format is supported by several long-term digital preservation systems, making it easier to move content from Fedora to these services. Since BagIt is a fairly loose standard, BagIt Profiles [6] are also supported to provide more information on the contents and structure of the Bag.

Once content has been exported from Fedora, it can be reimported to the same repository or a different Fedora instance using the same tool. This supports the migration use cases listed in the previous section, and lays the groundwork for more robust migration scenarios in the future. Additionally, by leveraging BagIt profiles, the import/export utility can account for differences in data modeling between repository instances; the Bag manifest contains relevant details that can be provided to the tool when importing content to the destination repository, which ensures important structural information is not lost between systems.

While this initial work is important in its own right, the real value comes from the foundation it lays for future work. As an external tool, the import/export utility could be made to work with a variety of repository systems without the need to modify the core code of these other systems. The next section lays out the potential for leveraging this opportunity more broadly.

4 ACHIEVING INTEROPERABILITY

The import/export utility was deliberately designed as a separate module; it interacts with Fedora via the REST-API [5], and therefore does not modify any of Fedora’s core code. This design decision opens the door for broader interoperability between repository systems by providing a means to interact with other systems without modifying core code. The main barrier in this case is adoption; developers and maintainers of other repository systems would need to adopt the utility as a means of getting data into and out of their systems.
This need for broad adoption may be addressed by a recently formed Research Data Alliance working group. The Research Data Repository Interoperability working group was struck to establish a standard by which content can be moved between different repository platforms. The group has created a primer document outlining candidate standards and technologies, of which the Fedora import/export utility is one. The group has also considered specifying a generic API for this purpose, but this would present a problem for users of existing platforms who do not have the resources to modify or upgrade their existing installations. Many users customize their repository platforms in ways that make it impossible to upgrade to a new version without significant code rewrites, thus setting the burden for adoption of a generic API very high. The import/export utility presents a much more attainable solution precisely because it does not require core repository code modifications in order to use.

The Research Data Repository Interoperability working group will meet at the 9th plenary meeting of the Research Data Alliance on April 6th, 2017 to finalize the primer document and decide on which technology/standard to pursue. Should the Fedora import/export utility be selected, the group will proceed to work with representatives of other repository platforms (who have already been identified) to determine how to integrate the utility with as many repository platforms as possible. This work will in turn feed into a published specification document that will outline the standards implemented by the utility and how it can be integrated with other platforms that may not be on our initial list. The overall goal of this work is to facilitate data migrations between as many repository platforms as possible, both in order to ease the burden of system migrations and to make it easier to share data objects between different platforms.

4 CONCLUSIONS

Experience has taught us that no software applications last forever; even long-lasting applications go through significant changes and rewrites that often require data migrations from one version to another. By recognizing and acknowledging this fact, we can plan for data migrations by adopting standards and practices that make the data more transparent, allocating resources for migration work, and developing tools to aid in the migration process.

The transition from Fedora 3 to 4 presented an opportunity for the Fedora community to better align with well-recognized standards and develop tooling to support data migrations, at least in the case of moving data between different versions and instances of Fedora. A group of community stakeholders drafted use cases, gathered requirements, and allocated developer resources to designing and building an import/export utility that can be used by anyone wishing to move data between Fedora instances or from Fedora to an external preservation system using the BagIt packaging standard.

While this work represents a significant milestone for the Fedora community, it has even broader implications for repository data migrations more generally. The Research Data Alliance Research Data Repository Interoperability working group will evaluate the import/export utility as a starting point for migrating data between different repository platforms. Should this utility be selected, group members will work to draft a specification document outlining how repositories can adapt the import/export utility to work with their repository platform to get data in and out in standardized formats. As this specification becomes more widely adopted, we will ease the burden of data migrations and make it easier to both upgrade existing repository platforms and migrate to new platforms in the future.

ACKNOWLEDGMENTS

This work would not be possible without the generous support of the Fedora community, particularly the DuraSpace members who allocate funding to the Fedora project on an annual basis. The import/export utility was designed, built, and tested by a diverse group of Fedora community stakeholders, all of whom deserve great thanks for their dedication to the project and our goals.

REFERENCES

Getting Persistent Identifiers Implemented By ‘Cutting In The Middle-Man’

R. van Veenendaal  
National Archives of the Netherlands  
The Hague  
The Netherlands  
remco.van.veenendaal@nationaalarchief.nl

M. Ras  
National Coalition for Digital Preservation  
The Hague  
The Netherlands  
marcel.ras@ncdd.nl

M.C. Dangerfield  
Rotterdam City Archives  
Rotterdam  
The Netherlands  
mc.dangerfield@rotterdam.nl

ABSTRACT
The Persistent Identifier (PID) project of the Dutch Digital Heritage Network (DHN) promotes sustainable access in the cultural heritage domain by stimulating the use of PIDs as references to (digital) cultural heritage objects. PIDs are long-lasting references to objects and offer a solution for link rot, which often results in 404 or Not Found error messages.

Cultural heritage organisations, regardless of size, are often hesitant to implement PIDs in their systems. They lack knowledge of PIDs, are unaware of the capabilities and benefits of PIDs, and fear possible complex and costly implementation processes as well as the maintenance costs for sustained services.

To address these issues, the DHN PID project focussed on:
- Raising awareness of (the importance of) PIDs in cultural heritage organisations
- Increasing knowledge regarding the use of PIDs within cultural heritage
- Supporting the technical implementation of PIDs in cultural heritage collection management systems

At the time of writing, four vendors have implemented PIDs in the collection management systems (CMS) they provide, and a dozen cultural heritage institutions are taking part in this pilot implementation phase. More importantly however, is that through these CMS vendors, PIDs have become more available and affordable as a sustainable long term access solution for hundreds of institutional digital collections and their users. To ensure that the information, training and education the project provided about PIDs can be disseminated and used as widely as possible, the project created (best practice) documentation in Dutch and English, along with a unique PID Guide for learning about PIDs and taking the first steps towards selecting a suitable PID solution, along with a PID implementation Roadmap. This was an effort to encourage the adoption and use of PIDs in the cultural heritage domain, by utilising their existing collection management system vendors.

The concept of ‘cutting in the middle-man’ – cooperating with collection management system vendors and supporting the implementation of PIDs in their products – has proven extremely successful, and the outcomes of this project may help cultural heritage initiatives in other countries to get PIDs implemented in their organisations too.

1 INTRODUCTION
Link rot, links resulting in a 404 or Not Found HTTP error message, is a nuisance to internet users and a threat to the sustainable availability of online cultural heritage and scientific information. Especially when links have been harvested and are (re)used in cultural heritage portals and aggregators like Europeana, broken links can have a significant impact. Museums, libraries, archives and scientific organisations in the Netherlands are joining forces to fight link rot by implementing Persistent Identifiers (PIDs) in their (collection management) systems. PIDs are unique and permanent links to (digital) objects.

The Digital Heritage Network (DHN) established a PID project to raise awareness of PIDs, increase knowledge regarding the use of PIDs, and support PID implementation projects.

2 DIGITAL HERITAGE NETWORK
The DHN is a partnership focused on developing a network of national faculties and services for improving the visibility, usability, and sustainability of digital heritage in the Netherlands. The network was established as an initiative by the Dutch Ministry of Education, Culture and Science and is undertaken by several key institutions and experts relating to the area of digital heritage [1].

The DHN has developed a three-pronged strategy covering Visible, Usable and Sustainable aspects of Digital Heritage. For each of these aspects a work package has been established outlining projects necessary to achieve their core goals. The Sustainable work package is coordinated by the Netherlands Coalition for Digital Preservation (Nationale Coalitie Digitale
3 PERSISTENT IDENTIFIERS

A Persistent Identifier is a unique and permanent link to a (digital) object, e.g. scan, audio file, metadata record, web page, or book. In the context of the internet, PIDs are posed as a solution to the problem of link rot: where web pages become inaccessible when a link or site address is changed. This is especially common when websites link to each other. When site addresses change (because the contents have been moved to a new page, for instance, or because the organisation maintaining the site has changed its domain name), the links no longer point to a valid address, resulting in a 404 or Not Found HTTP status code. With the rise of aggregators like Europeana, and the increasing availability of (linked) open data, link rot has become a serious threat to the sustainable findability of digital cultural heritage. The metadata of digital objects are harvested on a large scale and reused on various websites, and if the original address is then changed, thousands or even millions of links can suddenly break.

4 THE PERSISTENT IDENTIFIER PROJECT

The Persistent Identifier project promotes sustainable access within the cultural heritage domain. Cultural heritage organisations, regardless of size, are often hesitant to implement PIDs. They lack knowledge of what PIDs are, don’t know about the capabilities and benefits of PIDs, and fear a possible complex and costly implementation process as well as the maintenance costs for a sustained service.

To address these issues, the project worked on:

- Raising awareness of (the importance of) PIDs in cultural heritage organisations
- Increasing knowledge regarding the use of PIDs within cultural heritage
- Supporting the technical implementation of PIDs in cultural heritage collection management

The project ran from September 2015 to June 2017, and will continue in a follow-up project in the second phase of the DHN program.

4.1 Raising Awareness

To raise awareness, the PID project first created an inventory of cultural heritage organisations currently using PIDs, and those that did not. Organisations using PIDs were asked to share their (best) practices with us. Organisations not yet using PIDs were invited to workshops and presentations, for which training and education material was created. A PID Roadmap was also written, explaining the most important steps in a PID implementation project: preparation, implementation and publication, and the subsequent management of PIDs.

The PID project made extended efforts to educate a wider audience about the importance of PIDs, by publishing articles in the magazines Archievenblad and Informatieprofessional, also resulting in an online PID dossier, and regular communication through the DHN’s social media channels. Three videos were recorded and published in Dutch and English: ‘What are Persistent Identifiers’, ‘Choosing the most suitable Persistent Identifier’, and ‘Implementing Persistent Identifiers’.

4.2 Increasing Knowledge

Once a cultural heritage organisation is aware of the importance of PIDs, they often notice that they need to learn more about PIDs and the various PID options, and eventually select one (or more) to implement. There are multiple PID options, including Archival Resource Keys (ARKs), Digital Object Identifiers (DOIs), the Handle System, OpenURL, Persistent Uniform Resource Locators (PURLs) or Uniform Resource Names (URNs). Each system has its own particular properties, communities, strengths and weaknesses. But which one is best suited to a particular organisation’s needs? Choosing a PID solution is an important decision with long-term implications. To help cultural heritage organisations learn and think about important PID aspects, and to guide them through their first steps towards selecting a PID solution, the project developed a PID Guide. The Guide was developed in collaboration with PID experts from Delft University of Technology (for DataCite DOI), SURFserva (the Handle System) and the National Library of the Netherlands (for the URN:NBN namespace). DataCite DOI, the Handle System and URN:NBN are the three most used PID solutions in the Netherlands.

4.3 The PID Guide

The PID Guide reuses the idea of Voting Advice Applications and guides participants through 25 statements about PIDs. The participant is presented with 5 options to choose from per statement. These options reflect a 5-point Likert scale. The outcome of this guide points towards a preference for one
or more PID solutions. More importantly, the participant has learned and thought about 25 important PID issues in the process.

The preference(s) calculated by the Guide should not be seen as a definitive choice for a particular PID solution. Other factors not represented in the guide may affect the final decision. An organisation may also be part of a community that has a strong preference for a particular PID solution. Where the Guide does not rule out any PID solution, an organisation could have showstopper criteria for selecting a PID solution, such as cost, level of support or formal standardisation.

PID solutions are rarely restricted to national borders, and while the PID Guide was written with Dutch requirements in mind, the 25 statements and technical setup are transferable internationally. We therefore welcomed reviews of the PID Guide by the International DOI Foundation13 and nestor14. Their feedback helped improve the PID Guide before publication. Additionally, the PID Tutorial held at iPres 2016 [4], resulted in interest from other countries (to translate the PID Guide) and communities to create a more comprehensive PID resource (including, e.g., more PID solutions). The PID Guide has a CC0 license15 and is freely reusable16: www.ncdd.nl/en/pid-wijzer.

At the time of writing, 91 participants have used the PID Guide. Participants are asked to provide information about their type of organisation, job title and collection management system. This gives the project valuable information regarding the types of organisations using the PID Guide, and their (resulting) preferences. After filtering out incomplete entries, the data from 69 participants, including 23 international participants, was used to calculate preferences for PID solutions in the organisation types archive, library, museum, science and other.

The bar chart from figure 1 is based on the average preferences resulting from the PID Guide. For each participant, the resulting preference scores for all (3) PID solutions are used as input. The chart shows a preference for handles in museums (based on 18 participants), a preference for URN:NBN in libraries (20 participants), and a preference for DataCite DOI in research institutions (10 participants). Archives (16 participants) have a less clear preference for a particular PID solution. In the Other category, URN:NBN is preferred. This is however based on a limited set of 5 participants, mostly from commercial organisations. Given the limited number of participants for this survey and the fact that no immediate follow-up research has yet been undertaken, figure 1 is a first sketch of the kind of information that can be gathered from the PID Guide. In time, we hope to receive more data from PID Guide participants, and analyse the PID Guide’s database in greater detail. The results could help focus future PID advocacy and implementation efforts.

### 4.4 Cost

The cost of (implementing) PIDs is an important issue. While some PID solutions are free, i.e. have no annual fee, the combined cost of implementation, maintenance and use is never zero. In the past, the cost of PID solutions has been a major inhibiting factor in PID implementation in cultural heritage organisations. For example during the Continuous Access to Cultural Heritage (CATCH, 2004-2014) and CATCHPLUS (2009-2012) projects, important steps were taken towards implementing PIDs in cultural heritage organisations, but the cost of using PIDs remained an obstacle for widespread uptake17. The timing of this PID project was more fortunate. The establishment of EPIDC18, the foundation of DONA19 and changes in SURFsara’s organisational policies led to affordable Handle System based PID services becoming available for cultural heritage organisations. With URN:NBNs provided for free by the National Library of the Netherlands, and clients of Delft University of Technology’s DataCite DOI service required only to pay an annual fee of €750 per prefix, the basic annual cost for using the major PID solutions in the Netherlands are all below €1,000 and within financial reach of most cultural heritage organisations regardless of size20.

One important cost issue remained: the cost of implementing PIDs within collection management systems. This is where the cutting in the middle-man strategy of the PID project came into play. Instead of requiring individual cultural heritage organisation

---

13 See https://www.doi.org/, visited 2017-03-29.
15 See https://creativecommons.org/publicdomain/zero/1.0/, visited 2017-03-29.
16 We are discussing the transfer of the stewardship of the PID Guide to the Open Preservation Foundation, thus further increasing its international visibility and sustainability.
19 See https://www.dona.net/, visited 2017-03-29.
20 Free DOI services, like Zenodo, figshare and Dataverse were not taken into consideration. Their modus operandi include uploading objects, and most cultural heritage organisations have published or want to publish and manage their collections in their collection management systems.
to finance a tailor-made PID implementation project, the PID project offered to pay for the initial implementation costs to the vendors, under the proviso that a generic and reusable solution was developed. In theory, the implemented PID solution should become available to all users of that (collection management) system.

4.5 Cutting in the middle-man

Via the projects’ various communication channels, workshops and presentations, word spread that support for PID implementation was available. This led to talks with vendors of collection management systems working with Dutch cultural heritage institutions. As a result, at the time of writing, 4 of the 5 key software vendors have started implementing PIDs in their systems, together with a variety of pilot institutions:

- Picturae\(^21\), together with the Noord-Hollands Archief, Archief Eemland, Regionaal archief Zutphen, Groninger archieven and Regionaal archief Tilburg, have made handles available in Memorix software,
- Deventer\(^22\), together with the Centraal Bureau voor de Statistiek, Regionaal Archief Nijmegen and Gemeentearchief Ede, are implementing handles in Atlantis,
- De Ree Archiefsystemen\(^23\), together with Regionaal Archief Rivierenland, are implementing handles in MAIS-MDWS,
- C\&i\(^24\), distributor of The Museum System, are implementing handles with the Wereldcultuuren group, which counts the Tropenmuseum, the Afrika Museum and Museum Volkenkunde as user institutions.

These vendors and their clients discussed possibilities and their institutional preferences, and in all cases, a preference for the Handle System based PID service from SURFsara was chosen. One of the benefits to this solution seems to be that PID registration and maintenance can be handled automatically, via an API. The PID project has also started discussion with the Dutch representatives of the Axiell Group\(^25\) regarding the implementation of PIDs in the Adlib software.

Each vendor has a specific business model and annual fee schedule for the provision of PIDs in their systems. Most see PIDs as a service that should be available in modern collection management systems, and keep their fees relatively close to the minimum tariff set out by SURFsara’s PID service. The aim of the project was to make the provision of PIDs advantageous for all parties, which has been achieved by this cutting in of the middle-man.

4.6 Best Practices

The collection management system vendors that implement PIDs and the cultural heritage organisations that use them, were asked to share their experiences from the implementation projects. Their shared experiences will help encourage other organisations in the future. Their experiences have been gathered in a best practice document. As the template for these documents has the same headings as the PID implementation Roadmap, a consistent feedback loop was created: organisations starting a PID implementation project can use the PID Roadmap, and their best practice documentation helps fill a library of PID implementation project experiences, which helps improve the Roadmap.

5 CONCLUSION

The DHN PID project set out to raise awareness of PIDs, increase knowledge regarding the use of PIDs, and support PID implementation projects. The project’s results show that the project’s approach was successful.

The PID information, training and education material used in workshops and presentations contributed to raising awareness, together with PID Roadmap and best practice documents. The unique PID Guide is still actively used by organisations for learning about PIDs and taking the first steps towards selecting a suitable PID solution. SURFsara’s Handle System-based PID service has become available to cultural heritage organisations at an affordable price, ensuring that all major PID solutions in the Netherlands are now available for less than €1,000 per year. Most importantly, by cutting in the middle-man, PIDs have been implemented in (4) leading collection management systems in the Netherlands, and are available to all their (inter)national customers.

Although the DHN is a national program, the results of the PID project can have an international impact. DataCite DOI, URN:NBN and the Handle System are available internationally. Other countries or communities can also choose to reuse the materials, PID Guide and the concept of cutting in the middle-man.

REFERENCES

Trustworthy and Portable Emulation Platform for Digital Preservation

Zahra Tarkhani  
Computer Science Department  
Indiana University Bloomington  
ztarkhan@indiana.edu

Geoffrey Brown  
Computer Science Department  
Indiana University Bloomington  
geobrown@indiana.edu

Steven Myers  
Computer Science Department  
Indiana University Bloomington  
samyers@indiana.edu

ABSTRACT

We present an architecture for a trustworthy and portable emulation platform designed to protect the confidentiality and integrity of sensitive born-digital content when executed on a fundamentally untrustworthy platform. In evidence, we present a modified GameBoy emulator which is executable on a remote user platform while simultaneously protecting the contents of game ROM files. In a more general application, an archive or a library might use such an emulation architecture to control access to restricted material on more sophisticated computer emulators.

Our solution relies on Intel’s Software Guard Extensions (SGX) technology for implementation of the trusted emulation environment. Access to sensitive data is protected by server controlled encryption keys accessible only within the protected execution environment. This enables secure caching of encrypted data on the untrusted user platform for use by the emulator and hence limits the potential performance issues originated from remote execution over Internet connections.

KEYWORDS

secure digital preservation, trustworthy emulation, malicious operating systems

1 INTRODUCTION

This paper provides a technological solution to a fundamental problem faced by libraries and archives with respect to digital preservation — how to allow patrons remote access to digital materials while limiting the risk of unauthorized copying. The solution we present allows patrons to execute trusted software on an untrusted platform, the example we explore is a game emulator which provides a convenient prototype to consider many fundamental issues.

Examples of born-digital content include commercial and academic publications, such as interactive CDROMs as well as the archived works of writers and statesmen. For example, Emory University provides access to the processed portions of Salman Rushdie’s digital archive, and an interactive simulation of Rushdie’s original computing environment[16]. Access is only possible via an emulator located in the reading room that limits access to the content.

While emulation is the main preservation strategy for legacy applications that require obsolete software for access [10, 28–30, 36, 37, 43], the approach we propose can also be used to secure access born-digital contents that are not being preserved by emulation platforms. For example, consider the case of PDF documents: while PDF encryption is widely supported [44] accessing such documents requires distributing keys; the approach we propose might be used to execute a trusted PDF viewer on an untrusted platform that is able to verify the untrusted platform and protect keys/documents.

One fundamental limitation of our approach is that anything the user is allowed to view or hear could be captured through screenshots or audio recordings. That is, our solution is susceptible to attacks on the output devices, just as all traditional digital-rights management systems are. However, this is an issue with current solutions even within the confines of a library or archive, unless patrons are strictly monitored (e.g. by impounding cell phones). Our solution does enable a new use case for born-digital sharing, even if tight controls are necessary — archives with restricted reading rooms might use our approach to safely share materials with other archives that provide similar physical security. The architecture can also be used to protect the content against unauthorized access from servers that provide web/cloud based emulation services.

Figure 1 illustrates the major components of the type of system we explore: an emulator, executing on the client side, accesses secure storage containing both the emulation system software stack (e.g. a disk image containing an operating system) and any digital content to be accessed. Because we cannot guarantee the security of the client operating system or file storage, we require a trusted emulator that has controlled access to any secured materials. Our proposed architecture is based upon several key assumptions:

• The remote platform is under full control of the user, and the institution does not trust any of its software or hardware components. Therefore, establishing trust between the server and client is essential. Trust establishment allows the server to look at the execution platform and verify
key properties that are required for the trusted application to protect its secrets on the remote platform. For example, in our solution, the server needs to verify that the remote platform uses a real SGX enabled CPU.

- The emulator provided by the source institution runs on the remote platform and relies on its resources. This means that the emulator is managed by the untrusted client-side operating system, which itself may have been subverted by malware.
- All sensitive materials on the server-side storage are encrypted, and there is a secure protocol to transmit the encrypted files and keys from the server to the emulation client after trust establishment.

Delivering sensitive content to an untrustworthy remote platform is very challenging. Solutions that rely only on data encryption and access control mechanisms are vulnerable to several software and hardware attacks. For example, a compromised OS may steal encryption keys and disk image contents from a running emulator. This paper will address the following security problems:

- **Trustworthy Emulator**
  - The emulator core, which has access to keys and unencrypted security sensitive data, is meant to be kept private from the user on the client side. Therefore, such data and keys must be secured by the emulator. The emulator should also support on-the-fly decryption/encryption of data stored on the local file system (see Figure 2).
- **Trust Establishment**
  - The server should be able to authenticate the user and verify his/her platform before exchanging keys and encrypted files (see Arrows (1) and (2) in Figure 2). The server should also be able to ensure that the client’s trusted emulator behaves correctly. That means the client should be able to attest to the fact that all requirements (e.g., running a right version of the emulator) for the emulator to behave as expected (e.g., protecting the restricted data) are satisfied.

Many different methods for solving both of these problems have been proposed [5, 6, 8, 12, 19, 20, 25, 26, 33, 34], but mostly they are based on modified OS, hypervisor, or compiler. Our approach relies only on Intel SGX hardware support that provides protected execution environments called enclaves. The enclaves allow execution of software within them such that others, including the owner or administrator of the machine, cannot peer into or modify—it is as if the code executes in an impenetrable black box. SGX further provides a hardware-based mechanism to attest to the software running in the enclave. All Intel Skylake processors support SGX, and thus it should be widely available in the future. At the time of writing this paper, The firmware (BIOS) and the drivers required for SGX are available for Windows and Linux operating systems. Applications that use SGX are currently most easily developed in the C/C++ languages, as Intel has released development tools for them.

The direct interface through assembly is both documented and possible, suggesting that a large number of development tools will be developed in due time. In section 3.1 we summarize SGX functionality related to our work. Although our solution relies on SGX, one may consider using other trusted execution environment (TEE) solutions such as ARM TrustZone technology [1] to implement our architecture.

As a concrete example of the proposed system architecture, we developed a prototype of our solution using GearBoy, an open source emulator for GameBoy. The GearBoy architecture is described in Section 3.3. While developing this system, our focus was to protect materials on the client side and we assumed a trusted server. However, the proposed techniques
can easily be applied to protecting resources on server-side and dedicated workstations as well.

The remainder of this paper is organized as follows. We begin with a review of related work in Section 2. For background, in Section 3 we explain required security concepts and the basic architecture of GearBoy, the GameBoy emulator. Next, in Section 4 we describe the architecture of our trustworthy emulator which is utilizing trusted hardware based enclaves. Finally, implementation details and results are provided in Section 5.

2 RELATED WORK

We survey related work in two areas: emulation as a digital preservation technique and systems that execute on and protect applications from an untrusted OS.

**Emulation as a digital preservation technique**

Emulation has been successfully tested to provide access to i) obsolete software and hardware systems [4, 7, 27, 32, 35], ii) preserved digital artifacts [3, 22, 31], and iii) console games [11, 23]. However, there are fundamental problems that must be resolved when providing remote access to sensitive preserved content.

One approach to provide remote access to born-digital materials is via a remote desktop or web-based access to emulation environments. The bwFLA project [27] implements a distributed framework and a networked approach, where the emulators run in a well-controlled environment. GRATE [38] provides a web-based remote emulation. The GRATE server is responsible for session management (managing VNC sessions, executing emulators, etc.) and transporting uploaded digital objects into the emulated environments. Although these systems can reduce technical hurdles on the client’s side, they provide limited security and are not practical solutions for highly sensitive collections. For example, it is possible to steal the digital objects from the servers through RDP/VNC vulnerabilities [39].

The KEEP project [4] transfers a complex service stack (it consists of a core application, a software archive, and emulator archive) to the remote user, but doesn’t address the security issues considered in this paper.

**Systems to protect apps from an untrusted OS**

Considerable effort in the research community have been devoted to protecting security-sensitive applications from a compromised OS (or other privileged system code) running on an untrustworthy platform [2, 5, 6, 8, 12, 18–20, 47]. While these works have resolved many of the issues, they have many drawbacks that need to be addressed.

Many of the solutions rely on trusted hypervisors to protect applications from a malicious OS [5, 12, 34, 46, 47]; however, they do not offer any protection against a compromised hypervisor. Our solution does not need a trusted hypervisor but does require SGX support, which has a smaller and more securely deployable code-base in comparison (small code bases developed by high-quality teams, e.g. Intel, are more amenable to the application of development tools and techniques). Among the proposed secure systems, some incur huge performance loss due to the costly encryption/decryption operations [5, 12, 20], or they require significant modifications to legacy applications [19, 20]. Another major approach is isolating applications in a dedicated VM [8, 34], which dramatically increases the size of trusted computing base (TCB) and therefore weakens their security strength. Systems like Haven [2] shields applications from an untrusted OS using a secure library OS, based on Drawbridge [24], which works inside an enclave. A library OS is a lightweight OS that runs in the address space of an application and only supports the minimal OS features which the application depends on. The approach taken by Haven requires building a customized library OS to support legacy emulator/application dependencies and also has a large TCB size.

3 BACKGROUND

In this section, we briefly describe SGX and its features used in developing our trusted emulator. We also review the trust establishment problem and explain how SGX’s attestation capability can be used to solve this problem. Finally, we briefly describe the GameBoy emulator architecture that we modified to build our trusted emulator. It is important to note that while our work builds upon Intel-specific features, it is reasonable to assume that this or equivalent features will be commonplace in the near future.

3.1 Software Guard Extensions (SGX)

The key technology that we exploit in implementing our system is Intel Software Guard Extensions (SGX) [21] that is a new technology added to the Intel Skylake family of processors to aid in trusted computation on remote platforms. SGX provides a set of new instructions and new memory access protections for the Intel x86 computing architecture. Collectively these extensions enable the creation of so-called “enclaves”. These are figurative black boxes in which secure code can be loaded and executed; even the operating system — the ultimate arbiter and controller of the machine— cannot peek into the box to inspect computation or its memory accesses. Through an attestation process, remote entities can confirm that particular code is running in the enclave.

In the most recent application development tools, enclaves are introduced to developers through a metaphor of trusted shared libraries.

Because enclaves isolate sensitive code and associated data from the host environment, they can securely handle secrets...
without fear of release to the host. In particular, enclaves are useful for managing cryptographic keys.

The mechanisms for how SGX protects the integrity and privacy of programs executing within an enclave are quite technical, but at a high-level, the CPU ensures that processes outside of the enclave do have access to read or manipulate the data within an enclave. This is done using physical capabilities provided by CPU (i.e., the CPU’s architecture and microcode), and by ensuring that all memory reads and writes are encrypted and authenticated using cryptographic methods. This is implemented using a secret-key that is created on the CPU, and is only accessible to it—that is, it is not even directly accessible within the enclave to programs operating within it.

An SGX-enabled trustworthy application is formed by two classes of trusted and untrusted components (see Figure 3).

- **Trusted component**
  The application’s trusted code (i.e., the code that has access to and processes the application’s secrets) which works inside an enclave. A trusted application can have several trusted components that reside in one or more enclaves. If necessary, multiple enclaves can arrange to communicate amongst themselves securely.

- **Untrusted component**
  The application’s code that doesn’t process sensitive data and doesn’t need to be protected inside an enclave.

Untrusted code creates and runs the enclave which is placed in the protected memory. When a trusted function is called, only the code running inside the enclave sees the sensitive data in a non-encrypted form, and any external access to the data is denied. An enclave only returns trusted function results to the untrusted component and the enclave data remains in the trusted memory.

### 3.2 Trust Establishment

This section includes technical details about how SGX can be used for trust establishment. Consider the scenario in which a trusted application (in our case an emulator) behaves correctly on a real SGX enabled platform and can protect sensitive data while being processed. When the application is running on an untrusted platform, the application’s provider needs mechanisms to verify the integrity of the remote platform and assure that it has not been tampered with. For example, the provider may need to verify that it is communicating with a real SGX-enabled platform and not with a malicious SGX emulator. It also needs to be able to ensure that the application behaves correctly and as expected, and hence, will be able to protect the critical information. After this verification step, it will be safe for the provider to send its encrypted data/key, via a secure channel (established by traditional public-key cryptographic mechanisms and with different keying data than those needed for the emulator), to the remote platform to be processed by the trusted application.

Providing secrecy of computation is important for our goals, but without a mechanism for ensuring that the code one believes is running in an enclave is the actual code running, secrecy is meaningless. For example, if the enclave can be loaded with malicious code that would “export” secret data outside of an enclave to a waiting malicious machine, then secrecy inside the enclave has achieved little. Unsurprisingly, SGX provides mechanisms to demonstrate that the trusted components of an application have been properly instantiated and are securely running within enclaves on a valid Intel SGX-enabled platform. These attestation mechanisms use two measurement registers provided by SGX to uniquely distinguish each enclave based on its code, data, and author. The measurements are created by using cryptographic hash functions which provide what are essentially unique “fingerprint” values. These values are recorded when the enclave is built, and are finalized before enclave execution starts. The CPU then uses a signing key that is i) unique to the CPU; ii) secret and known to no others; and iii) physically burned into the CPU upon its creation. It performs something similar to a digital signature\(^1\) of the unique fingerprint that was created from the measured code in the enclave [14, 15]. Intuitively, this signature allows an initiating party to verify that the correct code has been loaded into an enclave. The signature is verified by confirming from Intel that the signature is indeed valid, and thus that the loaded code is as declared.

---

\(^1\)Due to the goal of providing privacy, an actual digital signature is not used, but a digital signature is a good first approximation.
A service provider (institutions, libraries, and archives) may utilize attestation to establish secure communication with a remote platform and provision sensitive materials such as encryption keys to the enclave through a secure channel.

The following example contains technical details that non-technical readers may safely skip.

Consider the scenario, illustrated in Figure 4, in which the application contains two trusted modules, Enclave 1 and Enclave 2 that are running on the same platform. For example, in our trusted GameBoy implementation, one enclave is responsible for trust establishment and key exchange between the service provider and the user. The other enclave contains the secured emulator modules that are responsible for protecting ROM files. The two enclaves can verify and authenticate with one another before exchanging information (see Arrow (2) in Figure 4). This process is called local attestation and is different, but related to the previous attestation methods discussed. After the two enclaves verify that their counterpart is trustworthy, they can exchange information on a secure channel, which provides confidentiality and integrity protection. The local attestation and protected channel establishment use the Diffie-Hellman Key Exchange protocol. After local attestation, the two enclaves trust each other, but the server still needs to verify the trustworthiness of the platform/application before exchanging any secrets.

The server only needs to perform remote attestation with one of the enclaves, Enclave 1 in our example (see Arrow (1) in Figure 4). After the attestation phases, each enclave receives secrets via a secure channel. Next, each enclave can securely encrypt (using a hardware-based encryption key) and store its sensitive data outside the enclave (e.g., on a disk) such that the data can be retrieved only by itself or a trusted enclave (sealing phase, see Arrows (3) and (4) in Figure 4).

3.3 Generic GameBoy Emulator Architecture

In this section, we describe a generic GameBoy emulator architecture and how our trusted emulator architecture can be implemented on it.

The key modules we describe here are part of the trusted or untrusted parts of our trusted emulator. Game Boy is a video game console, originally developed and manufactured by Nintendo in 1989 [40]. There are different GameBoy models. Figure 5 shows the essential components of a GameBoy emulator.

The emulator core contains the major emulated modules of a real GameBoy console. The core includes one or more emulated CPU models. Usually, the CPU is based on a subset of the Zilog Z80 microprocessor and has instructions and registers similar to the Intel 8080 and Intel 8085. It has eight 8-bit registers and two 16-bit registers (SP and PC) [17].

The core contains different memory modules such as video RAM or graphics memory. The amount of such memory is very small (e.g., less than 1MB). GameBoy has no operating system. Therefore, it operates by storing all system functions inside a game ROM cartridge which contains the game’s program and data. The cartridge may use memory bank controllers (MBCs) to switch between memory banks and control backup RAM. An emulated cartridge loads a video game’s ROM file and maps it to the emulator memory.

Video and audio controllers generate the video/audio timings and read video/audio data from memory in order to output it to screen. The drivers may depend on cross-platform multimedia libraries like SDL [42] to display video/audio.

The core contains all modules that directly access and process ROM files—the secrets we want to protect—so they should be protected and secured in any emulation. There is
a narrow interface to connect the emulator core to the rest of the emulator. The non-core parts of the emulator implement the graphical user interface and IO operations. These portions don’t need to be secured in our model. Through the interface, the trusted and untrusted components communicate in a way that no secret from the trusted code can leak out to the untrusted part.

4 TRUSTWORTHY EMULATOR ARCHITECTURE

In this section, we explain the following key components of our secure emulation architecture, and how these components operate to provide the security level needed for remote accessing to restricted materials (see Figure 6). We begin with a description of a threat model for our system.

4.1 Threat Model

Our goal is to allow possessors of private digital-born media to be able to share this over an untrusted Internet to a foreign client location so that it can be viewed on an emulator at a foreign site. The emulator will provide the appropriate output (audio and video), and those channels are not secured by our system. The assumption is that the foreign client may have malicious software, and be operated by malicious operators. The goal is that the system keeps any data otherwise private. We do not protect against physical attacks which are able to decompose the physical CPU and retrieve physical information that is embedded in the device or otherwise physically alter the execution of the CPU via some physical attack. We assume that the possessor has some way of authenticating the foreign-site to ensure that they are talking to the correct individual; this is traditionally done through password or public-key infrastructure. We also do not provide security against software attacks running on the machine that makes use of memory-page lookups, or timing and cache side-channels. Such attacks are known to be plausible [45], but they require software which explicitly targets the emulator in question and the software running under emulation. These attacks are expensive, as they must be specifically targeted at the emulator in question, and require more technical sophistication than traditional software exploits on the part of the authors. Therefore, such exemptions in the threat model really should provide substantial improvements in real-world security. Further, countermeasures are actively being developed for these attacks which can be incorporated into our model at a future date.

4.2 Key Components of Emulation Architecture

The key components of our emulation architecture are explained in this section.

Trusted Emulator: The emulator contains a trusted component which is hosted inside a secure enclave and it is responsible for protecting the restricted data from the untrusted platform.

Authentication and Verification (AV): The server needs to setup an AV module on both the client and server to authenticate the user and verify untrusted platform. These AV modules are responsible for trust establishment, and to perform a key exchange between the server and client.

Secure Remote File Access: The server needs to provide access to the encrypted files, i.e. sensitive materials we want to protect, via a secure remote file system.

In this architecture, the restricted content provided by the server is encrypted, and the encrypted files and keys will be provisioned to the client after the establishment of trust. On the client side, only the trusted modules, which are working inside SGX enclaves, have access to the keys and unencrypted data. We now elaborate on each component.

4.3 Trusted Emulator

The trusted emulator should be able to protect the sensitive data on the client platform by processing the data inside secured enclaves. In our architecture, only the trusted component executes inside an enclave and thus has access to restricted content. The trusted component includes the core, the trusted interfaces, and the security module (see Figure 6).

The core, which is responsible for processing ROM files, contains the emulated cartridge, processor, registers, memory, and video/audio controllers (see Figure 5). The core communicates outside the enclave through narrowtrusted interfaces. The untrusted component is responsible for creating the core enclave and passing the encrypted data to the core to be processed, and thus no enclave protection is necessary. It also includes the emulator graphical user interface (GUI).

Since the restricted materials are encrypted by the service provider, we add a security module to the trusted component which is responsible for on-the-fly encryption/decryption of the data inside the enclave, allowing the core to access the decrypted data, but never making it available outside an enclave. The enclave re-encrypts any data that it needs to store on the local file system. It seals keys and encrypted files before shutting the emulator down (destroying the enclave). Sealing is the process of encrypting enclave secrets to disk for long term storage. Encryption of these keys makes use of long-term secret keys embedded in the CPU, created randomly and burned into it at manufacturing time.

The security module gains access to the keys after ensuring the AV enclave is trusted and running in the same SGX enabled platform through a local attestation during the emulator setup phase. In the local attestation process, two
4.3.1 Authentication and Verification. The AV modules (on the server and client side) are responsible for establishing trust between the server and remote platform. The client AV module has a trusted component, called the AV enclave, which is responsible for secure session establishment, attestations and key exchange/store in the remote platform. First, the client AV launches the trusted emulator and ensures the trusted emulator’s enclave is instantiated correctly through a local attestation—refers to two enclaves on the same platform authenticating to each other, (see Arrow (2) in Figure 6). After the local attestation with the emulator’s enclave, the client AV enclave generates a verifiable report of the client identity and the emulator’s enclave. This report is generated and signed by an architectural enclave, called a quoting enclave, and bound to the client platform by the CPU. The server uses this report to verify the platform/emulator on the client side through a remote attestation (see Arrow (1) in Figure 6). In this process, the server verifies the remote attestation report generated by the AV enclave to ensure that is communicating with a real SGX-enabled machine and the user is authorized to access the materials on the specified platform.

Remote attestation requires establishing a secure communication session between the client and server to securely exchange the secrets. The secure session establishment is done via traditional public-key cryptography (analogous to how the SSL handshake includes both authentication and session establishment).

After a successful remote attestation, the server provides access to encrypted digital objects and provisions the keys to the AV enclave. The keys would be securely stored and only accessible by the AV enclave and emulator’s enclave using the SGX sealing mechanism (e.g., sealing by signing identity). Since the keys are securely stored, there is no need to repeat the attestation and secret provisioning steps every time the emulator is restarted on the platform. However, if one wants to insert policies on use, it is certainly possible to change the architecture so that there could be provisioning from the provider server on each initiation, or on a regular time period.

4.3.2 Secure Remote File Access. After a successful remote attestation, the server provides secure access to the encrypted files via mounting of a secure remote file system on the local platform (see Arrow (3) in Figure 6). For example, in our implementation, the trusted emulator uses SSHFS (SSH Filesystem) that is a filesystem to mount and interact with files located on a remote server via the SSH file transfer protocol (SFTP), a network protocol providing secure file transfer and a secure remote file system. SSHFS authenticates and encrypts connections. Thus, only those who should have access to remote directories can mount them. Files are encrypted inside such secure remote file system, and only the enclaves (the AV and emulator’s enclave) have access to the keys and can decrypt the data.

5 IMPLEMENTATION AND RESULTS

We implemented a prototype of the system on a platform with Skylake processor and SGX-enabled BIOS support. We
started with an open-source GameBoy emulator, called Gearboy [9], and added the trusted component and the AV module to it. Designing the trusted GameBoy using SGX requires considering fundamental decisions that affect both (1) the security properties of the system, such as the TCB size and the exposed interface to the enclave’s outside world, and (2) the performance, mostly due to the restrictions of SGX such as unsupported system calls. Therefore, the most important step in developing the system is refactoring the emulator into the trusted and untrusted components and designing the trusted interfaces to connect them. Besides the system design trades-off, refactoring process can also be challenging because of strong dependency between the core and untrusted modules. As Figure 5 demonstrates, the trusted core includes the cartridge, processor, registers, memory, and video/audio controllers. In our work, we assume no secrets are displayed to the user. However, since the Video/Audio drivers have direct access to the emulated processor, memory and registers, we added the Video/Audio controllers and the main frame-buffers to the trusted core to decrease security risks against the core and improve performance by decreasing the interaction between the core and the untrusted world. Having minimal interfaces is also important for step-by-step debugging the system and ensure its correctness. The untrusted component includes the graphical user interface, IO modules, and SDL library which are not hosted inside the enclave. The untrusted part is responsible for creating enclave pages using the ECREATE instruction, loading the trusted code and data into the enclave using the EADD instruction, updating the enclave’s measurement for the attestation phase using the EEXTEND instruction, and initializing the enclave using the EINIT instruction. Then a process can execute the enclave’s code using the EENTER instruction and leave it using the EEEXIT instruction. Because SGX does not support many legacy (frequently insecure) standard C library functions, system-calls, and some of the user-mode instructions [13, 21], one of the main challenges in our work was re-implementing a large part of the core to use the Intel’s provided trusted libraries or providing trusted interfaces for accessing the unsupported system-calls or instructions. For Example, SGX does not support IO operations inside an enclave; so we developed a trusted interface for the basic file operations used by the core. Following the same approach, we provided trusted interfaces for accessing SDL library and Qt framework [41] (that are not trusted) from the core enclave. The emulator depends on SDL and Qt for providing the GUI, displaying video/audio, and keyboard/event handling. To ensure no secret from the trusted core can leak out to the untrusted part, the trusted interfaces are designed in a manner that requires carefully validating all the inputs and outputs of the trusted interfaces within the enclave. To detect unauthorized input from outside into the enclave, we check the correctness of the input based on its type. As a simple example, in the case of numerical input, we can verify its range, length boundaries, and value before it is processed by the trusted code. A similar approach is considered for validating outputs from enclave to the untrusted domain. The emulated cartridge loads encrypted ROMs from the secure remote file system through the trusted IO interface. The security module has access to the keys and is responsible for decrypting ROMs after being loaded by the ROM loader and encrypting them before writing to the local file cache. Therefore ROMs would be unencrypted only inside the enclave and encrypted before storing outside the enclave. The security module uses Intel’s trusted OpenSSL library (topenssl) for cryptographic operations inside the enclave. Refactoring of the emulator to fit with the trusted/untrusted partitioning and provide the needed SGX support increased the lines of code (LOC) by $\approx 15\%$. Less than 10k LOC were added to enclaves, so while an enclave’s trusted storage is limited (128MB at most) it is more than enough for many similar emulators. For performance evaluation, we first compare running time of the original emulator with the trusted emulator, both executing the same (not encrypted) ROM file, without enabling the on-the-fly encryption/decryption support for the trusted emulator. This evaluation shows the running time of our trusted emulator increased by 8.6X. In the second evaluation, we compared the two emulators, but this time we enabled the trusted emulator on-the-fly encryption/decryption support. In this case, the trusted emulator performance decreased by 9.4X. Our analysis shows that the main performance cost is because of the IO and cryptography operations from inside enclaves during the startup and shutdown phases. Finally, we developed the client AV module using the Intel Elliptic Curve Diffie-Hellman (ECDH) key exchange library to establish a trusted channel between the two enclaves (the security module is responsible for the local attestation) and attest the trusted emulator locally. For remote attestation, it uses the Intel’s key exchange and remote attestation library. 6 CONCLUSION This paper introduces a secure, trustworthy and portable emulation architecture for digital preservation to provide a secure remote access to restricted born-digital materials while protecting the confidentiality and integrity the sensitive data. Our solution is built on top of Intel SGX which is the state-of-the-art technology in trustworthy computing. We developed a trusted GameBoy emulator as a proof of concept. However, our solution can be used to build more
general-purpose trustworthy digital preservation platforms with reasonable performance overheads. For example, we have been working on developing a trusted classic MacOS emulator by modifying Basilisk emulator (a popular classic MacOS emulator). The trusted Basilisk is designed to protect confidentiality and integrity of restricted disk images while processing them on an untrusted platform. Similar to the trusted GameBoy architecture, our trusted Basilisk architecture has a trusted core and an untrusted component which are connected by narrow trusted interfaces. The trusted Basilisk core, which works inside an enclave, contains the 68k CPU engine (including the interpreter, CPU/FPU emulator, and memory management), the ROM and resource patches, and some of the drivers (including the VHD disk, CD-ROM, and video drivers). The untrusted component includes the remained modules. As an example, Emory University could use the trusted Basilisk to provide remote access to the processed portions of Salman Rushdie’s digital archive[16], while protecting the digital archive from unauthorized copying. To protect confidentiality and integrity of the restricted disk images, the trusted Basilisk core loads encrypted disk images from a secure file system and decrypts them inside the emulator core enclave and encrypts the data while writing the modified images back to a local cache.

An institution can use such trustworthy emulation platforms to provide a secure remote access to restricted born-digital content while protecting the confidentiality and integrity of the sensitive data while being processed on an untrusted platform.

Our solution can also be used for providing secure remote access to born-digital documents that do not depend on emulation platforms. For example, a service provider can use a similar architecture to develop a trusted PDF viewer to protect confidentiality and integrity of sensitive documents while being processed on an untrusted platform. Basically, PDF is a container that includes information about how to layout and organize contents to display. The core functionality of a trusted PDF viewer that needs to be protected inside and enclave is the process of rendering (encoding image information) the contents of a PDF file into a display format; Usually, PDF viewers depend on PDF rendering libraries or SDKs to render PDF contents. The PDF viewer needs to provide trusted interfaces to access such untrusted libraries from the enclave.

ACKNOWLEDGMENT

This material is based in part upon work supported by the National Science Foundation under Grant Numbers IIS-1529415 and CSN-1111149. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Software Heritage: Why and How to Preserve Software Source Code

Roberto Di Cosmo
Inria and University Paris Diderot
France
roberto@dicosmo.org

Stefano Zacchiroli
University Paris Diderot and Inria
France
zack@irif.fr

ABSTRACT
Software is now a key component present in all aspects of our society. Its preservation has attracted growing attention over the past years within the digital preservation community. We claim that source code—the only representation of software that contains human readable knowledge—is a precious digital object that needs special handling: it must be a first class citizen in the preservation landscape and we need to take action immediately, given the increasingly more frequent incidents that result in permanent losses of source code collections.

In this paper we present Software Heritage, an ambitious initiative to collect, preserve, and share the entire corpus of publicly accessible software source code. We discuss the archival goals of the project, its use cases and role as a participant in the broader digital preservation ecosystem, and detail its key design decisions. We also report on the project road map and the current status of the Software Heritage archive that, as of early 2017, has collected more than 3 billion unique source code files and 700 million commits coming from more than 50 million software development projects.

ACM Reference Format:
https://doi.org/10.1145/nnnnnnn.nnnnnnnn

1 INTRODUCTION
Software is everywhere: it powers industry and fuels innovation, it lies at the heart of the technology we use to communicate, entertain, trade, and exchange, and is becoming a key player in the formation of opinions and political powers. Software is also an essential mediator to access all digital information [1, 5] and is a fundamental pillar of modern scientific research, across all fields and disciplines [38]. In a nutshell, software embodies a rapidly growing part of our cultural, scientific, and technical knowledge.

Looking more closely, though, it is easy to see that the actual knowledge embedded in software is not contained into executable binaries, which are designed to run on specific hardware and software platforms and that often become, once optimized, incomprehensible for human beings. Rather, knowledge is contained in software source code which is, as eloquently stated in the very well crafted definition found in the GPL license [13], "the preferred form [of a program] for making modifications to it [as a developer]."

Yes, software source code is a unique form of knowledge which is designed to be understood by a human being, the developer, and at the same time is easily converted into machine executable form. As a digital object, software source code is also subject to a peculiar workflow: it is routinely evolved to cater for new needs and changed contexts. To really understand software source code, access to its entire development history is an essential condition.

Software source code has also established a new relevant part of our information commons [20]: the software commons—i.e., the body of software that is widely available and can be reused and modified with minimal restrictions. The raise of Free/Open Source Software (FOSS) over the past decades has contributed enormously to nurture this commons [32] and its funding principles postulate source code accessibility.

Authoritative voices have spoken eloquently of the importance of source code: Donald Knuth, a founding father of computer science, wrote at length on the importance of writing and sharing source code as a mean to understand what we want computers to do for us [19]; Len Shustek, board director of the Computer History Museum, argued that "source code provides a view into the mind of the designer" [33]; more recently the importance of source code preservation has been argued for by digital archivists [4, 22].

And yet, little action seem to have been put into long-term source code preservation. Comprehensive archives are available for variety of digital objects, pictures, videos, music, texts, web pages, even binary executables [18]. But source code in its own merits, despite its significance, has not yet been given the status of first class citizen in the digital archive landscape.

In this article, we claim that it is now important and urgent to focus our attention and actions to source code preservation and build a comprehensive archive of all publicly available software source code. We detail the basic principles, current status, and design choices underlying Software Heritage,1 launched last year to fill what we consider a gap in the digital archiving landscape.

This article is structured as follows: Sections 2 through 4 discuss the state of the art of source code preservation and the mission of Software Heritage in context. Sections 5 and 6 detail the design and intended use cases of Software Heritage. Before concluding, Sections 7 through 9 present the data model, architecture, current status, and future roadmap for the project.

1https://www.softwareheritage.org
2 RELATED WORK

In the broad spectrum of digital preservation, many initiatives have taken notice of the importance of software in general, and it would be difficult to provide an exhaustive list, so we mention here a few that show different aspects of what has been addressed before.

A number of these initiatives are concerned with the execution of legacy software in binary form, leveraging various forms of virtualisation technologies: the Internet Archive [18] uses web-based emulators to let visitors run in a browser old legacy games drawn from one of their software collection; the E-ARK [8] and KEEP [9] projects brought together several actors to work on making emulators portable and viable on the long term, while UNESCO Persist [36] tries to provide a host organization for all activities related to preserving the executability of binaries on the long term.

NIST maintains a special collection of binaries for forensic analysis use [26]: while the content of the archive is not accessible to the public, it has produced interesting studies on the properties of different cryptographic hashes they use on their large collection of software binaries use [23].

The raising concern for the sore state of reproducibility of scientific results has spawn interest in preserving software for science and research, and several initiatives now offer storage space for depositing software artefacts: CERN’s Zenodo [39] also provides integration with GitHub allowing to take snapshots of the source code of a software project (without its history).

Finally, the raise of software engineering studies on software repositories have led several researchers to build large collections of source code [17, 25, 35] and more recently databases with event metadata from GitHub [10, 14]; these initiatives have as main goal to provide a research platform, not an archive for preservation.

To the best of our knowledge, in the broader digital preservation landscape the niche of software archival in source code form has not been addressed in its own right before.

3 SOFTWARE SOURCE CODE IS AT RISK

Despite the importance of source code in the development of science, industry, and society at large, it is easy to see that we are collectively not taking care of it properly. In this section we outline the three most evident reasons why this is the case.

The source code diaspora. With the meteoric rise of Free/Open Source Software (FOSS), millions of projects are developed on publicly accessible code hosting platforms [34], such as GitHub, GitLab, SourceForge, Bitbucket, etc., not to mention the myriad of institutional “forges” scattered across the globe, or developers simply offering source code downloads from their web pages. Software also tend to move among code hosting places during its lifetime, following current trends or the changing needs and habits of its developer community.

Once a particular version of a software is released, the question arises of how to distribute it to users. Here too the landscape is quite varied: some developers use the code hosting also for distribution, as most forges allow it. Other communities have their own archives organized by software ecosystems (e.g., CPAN, CRAN, …), and then there are different software distributions (Debian, Fedora, …) and package management systems (npm, pip, OPAM, …), which also retain copies of source code released elsewhere.

It is very difficult to appreciate the extent of the software commons as a whole: we direly need a single entry point—a modern “great library” of source code, if you wish—where one can find and monitor the evolution of all publicly available source code, independently of its development and distribution platforms.

The fragility of source code. We have known for a long time that digital information is fragile: human error, material failure, fire, hacking, can easily destroy valuable digital data, including source code. This is why carrying out regular backups is important.

For users of code hosting platforms this problem may seem a distant one: the burden of “backups” is not theirs, but the platforms’ one. As previously observed [37], though, most of these platforms are tools to enable collaboration and record changes, but do not offer any long term preservation guarantees: digital contents stored there can be altered or deleted over time.

Worse, the entire platform can go away, as we learned the hard way in 2015, when two very popular FOSS development platforms, Gitorious [11] and Google Code [16] announced shutdown. Over 1.5 million projects had to find a new accommodation since, in an extremely short time frame as regards Gitorious. This shows that the task of long term preservation cannot be assumed by entities that do not make it a stated priority: for a while, preservation may be a side effect of other missions, but in the long term it won’t be.

We lack a comprehensive archive which undertakes this task, ensuring that if source code disappears from a given code hosting platform, or if the platform itself disappears altogether, the code will not be lost forever.

A big scientific instrument for software, or lack thereof. With the growing importance of software, it is increasingly more important to provide the means to improve its quality, safety, and security properties. Sadly we lack a research instrument to analyze the whole body of publicly available source code.

To build such a “very large telescope” of source code—in the spirit of mutualized research infrastructures for physicists such as the Very Large Telescope in the Atacama Desert or the Large Hadron Collider in Geneva—we need a place where all information about software projects, their public source code, and their development history is made available in a uniform data model. This will allow to apply a large variety of “big code” techniques to analyze the entire corpus, independently of the origin of each source code artifact, and of the many different technologies currently used for hosting and distributing source code.

4 MISSION AND CHALLENGES

In order to address these three challenges, in June 2016 the Software Heritage project was unveiled, with initial support by India, with the stated goal to collect, organize, preserve, and make easily accessible all publicly available source code, independently of where and how it is being developed or distributed. The aim is to build a common archival infrastructure, supporting multiple use cases and applications (see Section 6), but all exhibiting synergies with long-term safeguard against the risk of permanent loss of source code.

To give an idea of the complexity of the task, let’s just review some of the challenges faced by the initial source code harvesting
phase, ignoring for the moment the many others that arise in subsequent stages. First, we need to identify the code hosting places where source code can be found, ranging from a variety of well known development platforms to raw archives linked from obscure web pages. There is no universal catalog: we need to build one!

Then we need then to discover and support the many different protocols used by code hosting platforms to list their contents, and maintain the archive up to date with the modifications made to projects hosted there. There is no standard, and while we hope to promote a set of best practices for preservation "hygiene", we must now cope with the current lack of uniformity.

We must then be able to crawl development histories as captured by a wide variety of version control systems [28]: Git, Mercurial, Subversion, Darcs, Bazaar, CVS, are just some examples of the tools that need to be supported. Also, there is no grand unifying data model for version control systems: one needs to be built.

To face such challenges, it is important that computer scientists get directly involved: source code is the DNA of their discipline and they must be at the forefront when it comes to designing the infrastructure to preserve it.

5 CORE PRINCIPLES

Building the Software Heritage archive is a complex task, which requires long term commitment. To maximize the chances of success, we based this work on solid foundations, presented in this section as a set of core principles for the project.

Transparency and Free Software. As stated by Rosenthal [31], in order to ensure long term preservation of any kind of information it is necessary to know the inner workings of all tools used to implement and run the archive. That is why Software Heritage will develop and release exclusively Free/Open Source Software (FOSS) components to build its archive—from user-facing services down to the recipes of software configuration management tools used for the operations of project machines. According to FOSS development best practices development is conducted collaboratively on the project forge and communication happens via publicly accessible media.

Replication all the way down. There is a plethora of threats, ranging from technical failures, to mere legal or even economic decisions that might endanger long-term source code preservation. We know that we cannot entirely avoid them. Therefore, instead of

Source code first. Ideally, one might want to archive software source code "in context", with as much information about its broader ecosystem: project websites, issues filed in bug tracking systems, mailing lists, wikis, design notes, as well as executables built for various platforms and the physical machines and network environment on which the software was run, allowing virtualization in the future. In practice, the resources needed for doing all this would be enormous, especially considering the no a priori selection principle, and we need to draw the scope line somewhere.

Software Heritage will archive the entire source code of software projects, together with their full development history as it is captured by state-of-the-art version control systems (or "VCS").

On one side, this choices allow to capture relevant context for future generations of developers—e.g., VCS history includes commit messages, precious information that detail why specific changes have been made to a given software at a given moment—and is precisely what currently nobody else comprehensively archives.

On the other side, a number of other digital preservation initiatives are already addressing some of the other contextual aspects we have mentioned: the Internet Archive [18] is archiving project

No a priori selection. A natural question that arises when building a long term archive is what should be archived in it among the many candidates available. In building Software Heritage we have decided to avoid any a priori selection of software projects, and rather archive them all.

The first reason behind this choice is pragmatic: we have the technical ability to archive every software project available. Source code is usually small in comparison to other digital objects, information dense and expensive to produce, unlike the millions of (cat) pictures and videos exchanged on social media. Additionally, source code is heavily redundant/duplicated, allowing for efficient storage approaches (see Section 7).

Second, software is nowadays massively developed in the open, so we get access to the history of software projects since their very early phase. This is a precious information for understanding how software is born and evolves and we want to preserve it for any "important" project. Unfortunately, when a project is in its infancy it is extremely hard to know whether it will grow into a king or a peasant. Consider PHP: when it was released in 1995 by Rasmus Lerdorf as PHP/FI (Personal Home Page tools, Forms Interpreter), who would have thought that it would have grown into the most popular Web programming language 20 years later?

Hence our approach to archive everything available: important projects will be pointed at by external authorities, emerging from the mass, less relevant ones will drift into oblivion.

Multi-stakeholder and non-profit. Experience shows that a single for profit entity, however powerful, does not provide sufficient durability guarantees in the long term. We believe that for Software Heritage it is essential to build a non profit foundation

that has as its explicit objective the collection, preservation, and sharing of our software commons.

In order to minimise the risk of having a single points of failure at the institutional level, this foundation needs to be supported by various partners from civil society, industry, and governments, and must provide value to all areas which may take advantage of the existence of the archive, ranging from the preservation of cultural heritage to research, from industry to education, (see Section 6).

The foundation should be run transparently according to a well-documented governance, and should be accountable to the public by reporting periodically about its activities.

5 CORE PRINCIPLES

Building the Software Heritage archive is a complex task, which requires long term commitment. To maximize the chances of success, we based this work on solid foundations, presented in this section as a set of core principles for the project.

Transparency and Free Software. As stated by Rosenthal [31], in order to ensure long term preservation of any kind of information it is necessary to know the inner workings of all tools used to implement and run the archive. That is why Software Heritage will develop and release exclusively Free/Open Source Software (FOSS) components to build its archive—from user-facing services down to the recipes of software configuration management tools used for the operations of project machines. According to FOSS development best practices development is conducted collaboratively on the project forge and communication happens via publicly accessible media.

Replication all the way down. There is a plethora of threats, ranging from technical failures, to mere legal or even economic decisions that might endanger long-term source code preservation. We know that we cannot entirely avoid them. Therefore, instead of attempting to create a system without errors, we design a system which tolerates them.

To this end, we will build replication and diversification in the system at all levels: a geographic network of mirrors, implemented using a variety of storage technologies, in various administrative domains, controlled by different institutions, and located in different jurisdictions. Releasing our own code as FOSS is expected to further ease the deployment of mirrors by a variety of actors.

Multi-stakeholder and non-profit. Experience shows that a single for profit entity, however powerful, does not provide sufficient durability guarantees in the long term. We believe that for Software Heritage it is essential to build a non profit foundation
websites, wikis, and web-accessible issue trackers; Gmane [12] is archiving mailing lists; several initiatives aim at preserving software executables, like Olive [21], the Internet Archive, KEEP [9], E-ARK [8], and the PERSIST project [36], just to mention a few.

In a sense, Software Heritage embraces the Unix philosophy [30], focusing on the source code only, where it contribution is most relevant, and will go to great lengths to make sure that the source code artifacts it archives are easy to reference from other digital archives, using state-of-the-art linked data [3] technologies, paving the way to a future “semantic wikipedia” of software.

**Intrinsic identifiers.** The quest for the “right” identifier for digital objects has been raging for quite a while [2, 29, 37], and it has mainly focused on designing digital identifiers for objects that are not necessarily natively digital, like books or articles. Recent software development practices has brought to the limelight the need for intrinsic identifiers of natively digital objects, computed only on the basis of their (byte) content.

Modern version control systems like Git [6] no longer rely on artificial opaque identifiers that need third party information to be related to the software artifacts they designate. They use identifiers that can be computed from the object itself and are tightly connected to it; we call these identifiers intrinsic. The SHA1 cryptographic hash [7] is the most used approach for computing them today. The clear advantage of crypto-hard intrinsic identifiers is that they allow to check that an obtained object is exactly the one that was requested, without having to involve third party authorities.

Intrinsic identifiers also natively support integrity checks—e.g., you can detect alteration of a digital object for which an intrinsic identifiers was previously computed as a mismatch between its (current) content and its (previous) identifier—which is a very good property for any archival system.

Software Heritage will use intrinsic identifier for all archived source code. Pieces of information that are not natively digital, such as author or project names, metadata, or ontologies, non-intrinsic identifier will also be used. But for the long term preservation of the interconnected network of knowledge that is built natively by source code, intrinsic identifiers are best suited.

**Facts and provenance.** Following best archival practices, Software Heritage will store full provenance information, in order to be able to always state what was found where and when.

In addition, in order to become a shared and trusted knowledge base, we push this principle further, and we will store only qualified facts about software. For example, we will not store bare metadata stating that the programming language of a given file is, say, C++, or that its license is GPL3. Instead we will store qualified metadata stating that version 3.1 of the program pygments invoked with a given command line on this particular file reported it as written in C++; or that version 2.6 of the FOSsology license detection tool, ran with a given configuration (also stored), reported the file as being released under the terms of version 3 of the GPL license.

**Minimalism.** We recognize that the task that Software Heritage is undertaking is daunting and has wide ramifications. Hence we focus on building a core infrastructure whose objective is only collecting, organizing, preserving, and sharing source code, while establishing collaborations with any initiative that may add value on top or on the side of this infrastructure.

### 6. APPLICATIONS AND USE CASES

A universal archive of software source code enables a wealth of applications in a variety of areas, broader than preservation for its own sake. Such applications are relevant to the success of the archive itself though, because long term preservation carries significant costs: chances to meet them will be much higher if there are more use cases than just preservation, as the cost may then be shared among a broader public of potential archive users.

**Cultural heritage.** Source code is clearly starting to be recognized as a first class citizen in the area of cultural heritage, as it is a noble form of human production that needs to be preserved, studied, curated, and shared. Source code preservation is also an essential component of a strategy to defend against digital dark age scenarios [1, 5] in which one might lose track of how to make sense of digital data created by software currently in production.

For these reasons Inria has established an agreement with UNESCO on source code preservation, whose main actions will be carried on in the context of Software Heritage.

**Science.** In the long quest for making modern scientific results reproducible, and pass the scientific knowledge over to future generations of researchers, the three main pillars are: scientific articles, that describe the results, the data sets used or produced, and the software that embodies the logic of the data transformation, as shown in Figure 1.

Many initiatives have been taking care of two of these pillars, like OpenAire [27] for articles and Zenodo [39] for data, but for software source code, researchers keep pointing from their articles to disparate locations, if any, where their source code can be found: web pages, development forges, publication annexes, etc. By providing a central archive for all publicly available source code, Software Heritage contributes a significant building block to the edifice of reproducibility in all fields of science.

And there is more: in the specific field of Computer Science, there is a significant added value in providing a central repository where all the history of public software development is made available in a

---


2and more recently for self-selected software releases distributed via GitHub
uniform data model. It enables unprecedented big data analysis both on the code itself and the software development process, unleashing a new potential for Mining Software Repository research [15].

**Industry.** Industry is growing more and more dependant on FOSS components, which are nowadays integrated in all kinds of products, for both technical and economic reasons. This tidal wave of change in IT brought new needs and challenges: ensuring technical compatibility among software components is no longer enough, one also needs to ensure compliance with several software licenses, as well as closely track software supply chain, and bills of materials to identify which specific variants of FOSS components were used in a given product.

Software Heritage makes two key contributions to the IT industry that can be leveraged in software processes. First, Software Heritage intrinsic identifiers can precisely pinpoint specific software versions, independently of the original vendor or intermediate distributor. This *de facto* provides the equivalent of "part numbers" for FOSS components that can be referenced in quality processes and verified for correctness independently from Software Heritage (they are intrinsic, remember?).

Second, Software Heritage will provide an open provenance knowledge base, keeping track of which software component—at various granularities: from project releases down to individual source files—has been found where on the Internet and when. Such a base can be referenced and augmented with other software-related facts, such as license information, and used by software build tools and processes to cope with current development challenges.

The growing support and sponsoring for Software Heritage coming from industry players like Microsoft, Huawei, Nokia, and Intel provides initial evidence that this potential is being understood.

7 **DATA MODEL**

In any archival project the choice of the underlying data model—at the *logical level*, independently from how data is actually stored on physical media—is paramount. The data model adopted by Software Heritage to represent the information that it collects is centered around the notion of *software artifact*, that is at the key component of the Software Heritage archive, and we describe it in what follows. It is important to notice that according to our principles, we must store with every software artifact full information on where it has been found (provenance), that is also captured in our data model, so we start by providing some basic information on the nature of this provenance information.

### 7.1 Source code hosting places

Currently, Software Heritage uses of a curated list of *source code hosting places* to crawl. The most common entries we expect to place in such a list are popular collaborative development forges (e.g., GitHub, Bitbucket), package manager repositories that host source package (e.g., CPAN, npm), and FOSS distributions (e.g., Fedora, FreeBSD). But we may of course allow also more niche entries, such as URLs of personal or institutional project collections not hosted on major forges.

While currently entirely manual, the curation of such a list might easily be semi-automatic, with entries suggested by fellow archivists and/or concerned users that want to notify Software Heritage of the need of archiving specific pieces of endangered source code. This approach is entirely compatible with Web-wide crawling approaches: crawlers capable of detecting the presence of source code might enrich the list. In both cases the list will remain curated, with (semi-automated) review processes that will need to pass before a hosting place starts to be used.

### 7.2 Software artifacts

Once the hosting places are known, they will need to be periodically looked at in order to add to the archive missing software artifacts. Which software artifacts will be found there?

In general, each software distribution mechanism will host multiple releases of a given software at any given time. For VCS, this is the natural behaviour; for software packages, while a single version of a package is just a snapshot of the corresponding software product, one can often retrieve both current and past versions of the package from its distribution site.

By reviewing and generalizing existing VCS and source package formats, we have identified the following recurrent artifacts as commonly found at source code hosting places. They form the basic ingredients of the Software Heritage archive:5

- **file contents** (AKA "blobs") the raw content of (source code) files as a sequence of bytes, *without* file names or any other metadata. File contents are often recurrent, e.g., across different versions of the same software, different directories of the same project, or different projects all together.
- **directories** a list of *named* directory entries, each of which pointing to other artifacts, usually file contents or sub-directories. Directory entries are also associated to arbitrary metadata, which vary with technologies, but usually includes permission bits, modification timestamps, etc.
- **revisions** (AKA "commits") software development within a specific project is essentially a time-indexed series of copies of a single "root" directory that contains the entire project source code. Software evolves when a developer modifies the content of one or more files in that directory and record their changes.
- **origins** code "hosting places" as previously described are usually large platforms that host several unrelated software projects. For software provenance purposes it is important to be more

5 As the terminology varies quite a bit from technology to technology, we provide both the canonical name used in Software Heritage and popular synonyms.
With all the bits of what we want to archive in place, the next as commonly intended are more abstract entities that
precisely source code artifacts are massively duplicated. This is so for
questions. As they result in creating an additional repository copy at
any given point in time, of a software component—in the example
and that can be associated to arbitrary metadata as well as
origins where their source code can be found.

any kind of software origin offers multiple pointers to
the "current" state of a development project. In the case of
VCS this is reflected by branches (e.g., master, development,
but also so called feature branches dedicated to extending
the software in a specific direction); in the case of package
distributions by notions such as suites that correspond to
different maturity levels of individual packages (e.g., stable,
development, etc.).

A "snapshot" of a given software origin records all entry
points found there and where each of them was pointing
at the time. For example, a snapshot object might track the
commit where the master branch was pointing to at any
given time, as well as the most recent release of a given
package in the stable suite of a FOSS distribution.

links together software origins with snapshots. Every time
an origin is consulted a new visit object is created, recording
when (according to Software Heritage clock) the visit
happened and the full snapshot of the state of the software
origin at the time.

7.3 Data structure

With all the bits of what we want to archive in place, the next question is how to organize them, i.e., which logical data structure to adopt for their storage. A key observation for

- code hosting diaspora discussed in Section 3;
- copy/paste (AKA "vendoring") of parts or entire external
  FOSS software components into other software products;
- large overlap between revisions of the same project: usually
  only a very small amount of files/directories are modified by
  a single commit;
- emergence of DVCS (distributed version control systems),
  which natively work by replicating entire repository copies
  around. GitHub-style pull requests are the pinnacle of this,
  as they result in creating an additional repository copy at
each change done by a new developer;

migration from one VCS to another—e.g., migrations from
Subversion to Git, which are really popular these days—
resulting in additional copies, but in a different distribution
format, of the very same development histories.

These trends seem to be neither stopping nor slowing down, and it is reasonable to expect that they will be even more prominent in the future, due to the decreasing costs of storage and bandwidth.

For this reason we argue that any sustainable storage layout
for archiving source code in the very long term should support
deduplication, allowing to pay for the cost of storing source code
artifacts that are encountered more than once—only once. For
storage efficiency, deduplication should be supported for all the
software artifacts we have discussed, namely: file contents, directo-
ries, revisions, releases, snapshots.

Realizing that principle, the Software Heritage archive is con-
ceptually a single (big) Merkle Direct Acyclic Graph [24] (DAG), as
depicted in Figure 2. In such a graph each of the artifacts we have
described—from file contents up to entire snapshots—correspond
to a node. Edges between nodes emerge naturally: directory entries
point to other directories or file contents; revisions point to direc-
tories and previous revisions, releases point to revisions, snapshots
point to revisions and releases. Additionally, each node contains all
metadata that are specific to the node itself rather than to pointed
nodes; e.g., commit messages, timestamps, or file names. Note that
the structure is really a DAG, and not a tree, due to the fact that the
line of revisions nodes might be forked and merged back.

In a Merkle structure each node is identified by an intrinsic
identifier (as per our principles detailed in Section 5) computed as
a cryptographic hash of the node content. In the case of Software
Heritage identifiers are computed taking into the account both
node-specific metadata and the identifiers of child nodes.

Consider the revision node shown in Figure 3. The node points
to a directory, whose identifier starts with ff3cc22... which
has also been archived. That directory contains a full copy, at a
specific point in time, of a software component—in the example
a component that we have developed ourselves for the needs of
Software Heritage. The revision node also points to the preceding

Figure 2: Software Heritage direct acyclic graph data model
The ingestion data flow of Software Heritage is shown in Figure 4.

Ingestion acts like most search engines, periodically crawling a set of "leads" (in our case the curated list of code hosting places discussed in Section 7) for content to archive and further leads. To facilitate software extensibility and collaboration, ingestion is split in two conceptual phases though: listing and loading.

**Listing** takes as input a single hosting place (e.g., GitHub, PyPI, or Debian) and is in charge of enumerating all software origins (individual Git or Subversion repositories, individual package names, etc.) found there at listing time. The details of how to implement listing vary across hosting platforms, and dedicated lister software components need to be implemented for each different type of platform. This means that dedicated listers exist for GitHub or Bitbucket, but that the GitLab lister—GitLab being a platform that can be installed on premises by multiple code hosting providers—can be reused to list the content of any GitLab instance out there.

Listing can be done fully, i.e., collecting the entire list of origins available at a given hosting place at once, or incrementally, listing only the new origins since the last listing. Both listing disciplines are necessary: full listing is needed to be sure that no origin is being overlooked, but it might be unwieldy if done too frequently on large platforms (e.g., GitHub, with more than 55 million Git repositories as of early 2017), hence the need of incremental listing to quickly update the list of origins available at those places.

Also, listing can be performed in either pull or push style. In the former case the archive periodically checks the hosting places to list origins. In the latter code hosting sites, properly configured to work with Software Heritage, contact back the archive at each change in the list of origins. Push looks appealing at first and might minimize the lag between the appearance of a new software origin and its ingestion in Software Heritage. On the other hand push-only listing is prone to the risk of losing notifications that will result in software origins not being considered for archival. For this reason we consider push an optimization to be added on top of pull, in order to reduce lag where applicable.

**8.2 Loading**

Loading is responsible of the actual ingestion in the archive of source code found at known software origins.

Loaders are the software components in charge of extracting source code artifacts from software origins and adding them to the archive. Loaders are specific to the technology used to distribute source code: there will be one loader for each type of version control system (Git, Subversion, Mercurial, etc.) as well as one for each source package format (Debian source packages, source RPMs, tarballs, etc).

Loaders natively deduplicate w.r.t. the entire archive, meaning that any artifact (file content, revision, etc.) encountered at any origin will be added to the archive only if a corresponding node cannot be found in the archive as a whole.

Consider the Git repository used for the development of the Linux kernel, which is fairly big, totaling 2 GB on disks for more than 600,000 revisions and also widely popular with thousands of (slightly different) copies available only on GitHub. At its first encounter ever, the Git loader will load essentially all its file contents, revisions, etc., into the Software Heritage archive. At the next encounter of an identical repository, nothing will be added at all. At the encounter of a slightly different copy, e.g., a repository containing a dozen additional commits not yet integrated in the official release of Linux, only the corresponding revision nodes, as

---

**Figure 3: A revision node in the Software Heritage DAG**

revision node (e4feb051... in the project development history. Finally, the node contains revision-specific metadata, such as the author and committer of the given change, its timestamps, and the message entered by the author at commit time.

The identifier of the revision node itself (64a78321... is computed as a cryptographic hash of a (canonical representation of) all the information shown in Figure 3. A change in any of them—metadata and/or pointed nodes—would result in an entirely different node identifier. All other types of nodes in the Software Heritage archive behave similarly.

The Software Heritage archive inherits useful properties from the underlying Merkle structure. In particular, deduplication is built-in. Any software artifacts encountered in the wild gets added to the archive only if a corresponding node with a matching intrinsic identifier is not already available in the graph—file content, commits, entire directories or project snapshots are all deduplicated incurring storage costs only once.

Furthermore, as a side effect of this data model choice, the entire development history of all the source code archived in Software Heritage—which ambitions to match all published source code in the world—is available as a unified whole, making emergent structures such as code reuse across different projects or software origins, readily available. Further reinforcing the use cases described in Section 6, this object could become a veritable "map of the stars" of our entire software commons.
well as the new file contents and directories pointed by them, will be loaded into the archive.

8.3 Scheduling

Both listing and loading happen periodically on a schedule. The scheduler component of Software Heritage is in charge of keeping track of when the next listing/loading actions need to happen, for each code hosting place (for listers) or software origins (loaders).

While the amount of hosting places to list is not enormous, the amount of individual software origins can easily reach the hundreds of millions given the current size of major code hosting places. Listing/loading from that many Internet sites too frequently would be unwise in terms of resource consumption, and also unwelcome by the maintainers of those sites. This is why we have adopted an adaptive scheduling discipline that strikes a good balance between update lag and resource consumption.

Each run of periodic action, such as listing or loading, can be “fruitful” or not. It is fruitful if and only if it resulted in new information since the last visit. For instance, listing is fruitful if it resulted in the discovery of new software origins; loading is if the overall state of the consulted origin differs from the last observed one. If a scheduled action has been fruitful, the consulted site has seen some activity since the last visit, and we will increase the frequency at which that site will be visited in the future; in the converse case (no activity), visit frequency will be decreased.

Specifically, Software Heritage adopts an exponential backoff strategy, in which the visit period is halved when activity is noticed, and doubled when no activity has been observed. Currently, the fastest a given site will be consulted is twice a day (i.e., every 12 hours) and the slowest is every 64 days. Early experiences with large code hosting sites seem to tell that ≈90% of the repositories hosted there quickly fall to the slowest update frequency (i.e., they don’t see any activity in 2-month time windows), with only the remaining ≈10% seeing more activity than that.

8.4 Archive

At a logical level, the Software Heritage archive corresponds to the Merkle DAG data structure described in Section 7. On disk, the archive is stored using different technologies due to the differences in the size requirements for storing different parts of the graph.

File content nodes require the most storage space as they contain the full content of all archived source code files. They are hence stored in a key-value object storage that uses as keys the intrinsic node identifiers of the Merkle DAG. This allows trivial distribution of the object storage over multiple machines (horizontal scaling) for both performance and redundancy purposes. Also, the key-value access paradigm is very popular among current storage technologies, allowing to easily host (copies of) the bulk of the archive either on premise or on public cloud offerings.

The rest of the graph is stored in a relational database (RDBMS), with roughly one table per type of node. Each table uses as primary key the intrinsic node identifier and can easily be sharded (horizontal scaling again) across multiple servers. Master/slave replication and point-in-time recovery can be used for increased performance and recovery guarantees. There is no profound reason for storing this part of the archive in a RDBMS, but for what is worth our early experiments seem to show that graph database technologies are not yet up to par with the size and kind of graph that Software Heritage already is with its current coverage (see Section 9).

A weakness of deduplication is that it is prone to hash collisions: if two different objects hash to the same identifier there is a risk of storing only one of them while believing to have stored them both. For this reason, where checksums algorithms are no longer
considered strong enough for cryptographic purposes,\textsuperscript{7} we use multiple checksums, with unicity constraints on each of them, to detect collisions before adding a new artifact to the Software Heritage archive. For instance, we do not trust SHA1 checksums alone when adding new file contents to the archive, but also compute SHA256, and “salted” SHA1 checksums (in the style of what Git does). Also, we are in the process of adding BLAKE2 checksums to the mix.

Regarding mirroring, each type of node is associated to a change feed that takes note of all changes performed to the set of those objects in the archive. Conceptually, the archive is append-only, so under normal circumstances each feed will only lists additions of new objects as soon as they get ingested into the archive. Feeds are persistent and the ideal branching point for mirror operators who, after an initial full mirror step, can cheaply remain up to date w.r.t. the main archive.

On top of the object storage, an archiver software component is in charge of both enforcing retention policies and automatically heal object corruption if it ever arises, e.g., due to storage media decay. The archiver keeps track of how many copies of a given file content exist and where each of them is—we currently operate in a two-house mirror of the entire object storage, plus a third copy currently being populated on a public cloud. The archiver is aware of the desired retention policy, e.g., “each file content must exist in at least 3 copies”, and periodically swipe all known objects for adherence to the policy. When fewer copies than desired are known to exist, the archiver asynchronously makes as many additional copies as needed to satisfy the retention policy.

The archiver also periodically checks each copy of all known objects—randomly selecting them at a suitable frequency—and verifies it for integrity by recomputing its intrinsic identifier and comparing it with the known one. In case of mismatch all known copies of the object are checked on-the-fly again; assuming at least one pristine copy is found, it will be used to overwrite corrupted copies, “healing” them automatically.

9 CURRENT STATUS & ROADMAP

The Software Heritage archive grows incrementally over time as long as new listers/loaders get implemented and periodically run to ingest new content.

Listers. In terms of listers, we initially focused on targeting GitHub as it is today by far the largest and most popular code hosting platform. We have hence implemented and put in production a GitHub lister, capable of both full and incremental listing. Additionally, we have recently put in production a similar lister for Bitbucket. Common code among the two has been factored out to an internal lister helper component that can be used to easily implement listers for other code hosting platforms.\textsuperscript{5} Upcoming listers include FusionForge, Debian and Debian-based distributions, as well as a lister for bare bone FTP sites distributing tarballs.

Loaders. Regarding loaders, we initially focused on Git as, again, the most popular VCS today. We have additionally implemented loaders for Subversion, tarballs, and Debian source packages. A Mercurial loader is also in the working.

Archive coverage. Using the above software components we have already been able to assemble what, to the best of our knowledge, is the largest software source code archive in existence.

We have fully archived once, and routinely maintain up-to-date, GitHub into Software Heritage, for more than 50 million Git repositories. GitHub itself has acknowledged Software Heritage role as 3rd-party archive of source code hosted there.\textsuperscript{8}

Additionally we have archived, as one shot but significant in size archival experiments, all releases of each Debian package in between 2005–2015, and all current and historical releases of GNU projects as of August 2015. We have also retrieved full copies of all repositories that were previously available from Gitorious and Google Code, now both gone. At the time of writing the process of ingesting those repositories into Software Heritage is ongoing.

In terms of storage, each copy of the Software Heritage object storage currently occupies \(150\) TB of individually compressed file contents. The average compression rasion is 2x, corresponding to 300 TB of raw source code content. Each copy of the RDBMS used to store the rest of the graph (Postgres) takes \(5\) TB. We currently maintain 3 copies of the object storage and 2 copies of the database, the latter with point-in-time recovery over a 2-week time window.

As a logical graph, the Software Heritage Merkle DAG has \(5\) billion nodes and \(50\) billion edges. We note that more than half of the nodes are (unique) file contents (\(\approx 3\) B) and that there are \(\approx 750\) M revision/commit nodes, collected from \(\approx 55\) M origins.

Features. The following functionalities are currently available for interacting with the Software Heritage archive:

content lookup allows to check whether specific file contents have been archived by Software Heritage or not. Lookup is possible by either uploading the relevant files or by entering their checksum, directly from the Software Heritage homepage.

browsing via API allows developers to navigate through the entire Software Heritage archive as a graph. The API offered to that end is Web-based and permits to lookup individual nodes (revisions, releases, directories, etc.), access all their metadata, follow links to other nodes, and download individual file contents. The API also gives access to visit information, reporting when a given software origin has been visited and what its status was at the time. The API technical documentation\textsuperscript{10} has many concrete examples of how to use it in practice.

The following features are part of the project technical road map and will be rolled out incrementally in the future:

Web browsing equivalent to API browsing, but more convenient for non-developer Web users. The intended user interface will resemble state-of-the art interfaces for browsing the content of individual version control systems, but will be tailored to navigate a much larger archive.

provenance information will offer “reverse lookups” of sort, answering questions such as “give me all the places and timestamps where you have found a given source code artifact”.

\textsuperscript{7}note that this is already a higher bar than being strong enough for archival purposes
\textsuperscript{8}see https://www.softwareheritage.org/?p=9594 for a detailed technical description
\textsuperscript{9}https://help.github.com/articles/about-archiving-content-and-data-on-github/
\textsuperscript{10}https://archive.softwareheritage.org/api/
This is the key ingredient to address some of the industrial use cases discussed in Section 6. **metadata search** will allow to perform searches based on project-level metadata, from simple information (e.g., project name or hosting place), to more substantial ones like the entity behind the project, its license, etc.

**content search** conversely, content search will allow to search based on the content of archived files. Full-text search is the classic example of this, but in the context of Software Heritage content search can be implemented at various level of “understanding” of the content of individual files, from raw character sequences to full-fledged abstract syntax trees for a given programming language.

## 10 CONCLUSIONS

Software Heritage is taking over the long overdue task of collecting all publicly available source code, with all its development history, organizing it into a canonical structure, and providing unique, intrinsic identifiers for all its parts, enabling its better fruition while ensuring its long term preservation.

This is a significant undertaking, that faces a broad range of challenges, from plain technical complexity to economic sustainability.

To maximize the chances of success, the core design principles of Software Heritage include technical choices like archival minimalism and deduplication, up to organizational decisions like running the project as a multi-stakeholder, transparent, non-profit initiative, open to collaboration with other digital archival initiatives for all non source code aspects of software archiving.

Software Heritage is run as an open collaborative project, and we call for digital archivists and computer scientists to critically review our work and join the mission of archiving our entire software commons. Challenges abound, but we believe they are worth it.

## ACKNOWLEDGMENTS

Software Heritage is a far reaching project that could not exist without the dedication and vision of many people and institutions that we would like to sincerely thank here.

First and foremost Inria as an institution, and its CEO Antoine Petit in particular, for providing the support needed to move from paper to reality. Jean-François Abramic, who spent countless hours to make the project take off, as chair of the advisory board and “compagnon de route” of the first author. The Software Heritage development team, for their tireless commitment, with shout out for Guillaume Rousseau, Antoine R. Dumont, and Nicolas Dandrimont—team members since day one.

Last but not least the special people who worked internally at the many sponsor organizations convincing them to commit resources to this mission and the many others, too numerous to list here, who contributed along the way.

## REFERENCES


ABSTRACT
This paper uses a collective case study to reveal similarities and differences in the operations and service models of nine distributed digital preservation services. The study uncovers a wide range of organizations and technical variations among the nine services, but finds that they can be grouped into three basic service models.

KEYWORDS
Digital preservation service models, digital libraries, digital curation, case study

1 INTRODUCTION
During the 2016 Annual Steering Committee Meeting of the MetaArchive Cooperative, the Steering Committee initiated an environmental scan of services provided by comparable digital preservation consortia and vendors. This study is a high-level overview of digital preservation service models, and includes an exploration of organizational aspects such as governance, support and training, documentation, community facilitation, outreach and communication, marketing, and membership; as well as technical aspects such as functionality, setup and configuration, content, ingest, storage, security, access, and integration; and finally, compliance with selected National Digital Stewardship Alliance Levels of Preservation[1].

Digital formats are highly sensitive to obsolescence, corruption, and degradation, elevating the importance of strategies and infrastructure for preserving digital records. Successful digital preservation “combines policies, strategies and actions” that ensure digital content survives in a usable form with an overarching goal to safeguard public records, scientific advances, and cultural heritage[2]. Many libraries engaged in digital preservation are “increasingly committing to the use of large-scale, comprehensive, distributed digital preservation systems” [3][4]. Such systems help libraries preserve greater amounts of data efficiently and cost-effectively according to standards and best practices.

A number of programs and providers offer similar digital preservation services, many of which are backed by the same third party vendors. However, they differ in their organizational models, strategies, and sometimes, architectures. The increasing number of options allows libraries to choose among several categories of service provider. Libraries might make financial and technical commitments based on the stated values of their service providers and peers within their community, how rigorously a service provider follows established standards, or based on the cost of the service and how it fits in the library’s budget and strategic plan. While we did not undertake a cost comparison since that information would be proprietary and/or negotiable, this study compares organizational and technical aspects of large-scale digital preservation services so that libraries can be better informed when deciding which providers make the most sense for them.

2 METHODS
Through documentary analysis, surveys, and interviews[5], we reviewed operations and service models of nine (MetaArchive, APTrust, DPN, TDL, DuraCloud, Preservica, Chronopolis, Rosetta, and Arkivum) distributed digital preservation service providers. We analyze governance, organizational structure, support and training, documentation, community facilitation, outreach and communication, marketing, membership, compliance with NDSA Levels of Preservation, functionality, setup and configuration, content, ingest, storage, security, access, and integration. We conducted a collective study[6] because these programs provide comparable functions yet have different organizational models and our purpose is to “reveal the properties of the class” of programs[7]. This analysis is a collective case study because it is an exploration of multiple programs in order to investigate distributed digital preservation services[8]. Schultz and Skinner[9] conducted a similar study that compared underlying technologies for three distributed digital preservation systems: Chronopolis, University of North Texas, and MetaArchive.

We created a rubric to compare each program across the categories listed above. Some of the information to complete this rubric comes from the organizations’ websites, and we received additional information through personal communication with executive directors or other program officials. Because we spoke with the representatives of these organizations about facts about their respective organizations, rather than opinions, Virginia
Tech’s Institutional Review Board waived the review requirement. We compared the profiles of the nine service providers and analyzed similarities and differences within the sample along the lines delineated by the question categories. The themes and the comparative analysis together describe how the nine programs differ in their approach to managing digital preservation services.

We identify several limitations to this study, some of which are general to case study research[8][10][11]. Additionally, due to the informal nature of the semi-structured interviews, some participants may have interpreted questions differently from each other, or the follow-up questions may have gone in different directions in each interview. As a result there is limited basis for comparison in some categories. The third limitation is that this is a fast-moving area and some of these programs are planning and initiating new services at the time of writing. Some findings therefore may be out of date by the time of publication.

3 RESULTS
This section organizes the findings into categories for comparing similarities and differences among the program in the sample.

3.1 Organizational Aspects

3.1.1 Governance
All organizations in the sample provide their users with digital preservation services, but each one articulates its mission in distinct ways. The mission statements published on their websites each note the importance of collaboration or connections with other preservation and access systems. The Arkivum and Preservica websites state the sectors they serve without focusing on the content. Other providers stated that their purpose is to ensure the longevity of cultural heritage digital content, but do not provide business model details in their mission statements. DPN and Preservica list types of digital resources they preserve while Chronopolis and TDL discuss the mission of preserving digital content more generally. APTrust, Chronopolis, and DuraCloud each identify a values statement, and DPN indicate that they have one in development. The other institutions in the sample do not have a formal values statement.

There are three distinct organizational types with equal distribution in the sample. 1) APTrust, TD, and Chronopolis are each legally constituted as a part of a university. 2) Preservica, Arkivum, and Rosetta are commercial services. 3) DPN, DuraCloud, and MetaArchive are non-profit limited liability companies. Leadership varies across the sample, although MetaArchive, DuraSpace, DPN, APTrust, Preservica, and TDL are governed by boards or steering committees. MetaArchive, APTrust, DPN, and TDL all accomplish technical and non-technical development work through committees. Rosetta and Preservica have active user groups that meet regularly online, and at regular meetings and conferences. Arkivum has an executive board which includes investors and senior management, but does not have a formal user group.

3.1.2 Support and Training
Most of the services provide training (see Table 2-Member Support and Training[13]) in the form of new customer orientations, instructional webinars, workshops, and/or video tutorials. DPN identifies itself as more of a catalyst organization instead of an educational organization, so they partner with other programs such as APTrust and DuraSpace for user support, and with AVPreserve and Educropia for curriculum development. TDL, APTrust, Arkivum, and MetaArchive provide in-person orientation and training either remotely or on site. Arkivum, Preservica, TDL, DuraCloud, and MetaArchive all offer training via webinars, and DPN plans to offer them in the future. APTrust and TDL both provide informal in-person workshops on-site to new members, whereas MetaArchive and DuraCloud offer in-person workshops on an irregular, ad hoc basis. Preservica provides briefings and workshops targeted at specific users and at conferences. DuraCloud, Preservica, Arkivum, TDL, and Rosetta all offer video tutorials through YouTube or Adobe Connect. APTrust also has training videos on YouTube but notes that they are slightly dated. All services offer support and troubleshooting via email, phone, or both. DPN offers informal support with legal agreements in digital preservation, but relies on APTrust and DuraSpace for formal technical support for members.

3.1.3 Documentation
The availability of comparable documentation in the form of Frequently Asked Questions and technical specifications varies among the nine providers. Rosetta, TDL, DPN, DuraCloud and APTrust, and to some extent MetaArchive have open technical documentation. Chronopolis, Arkivum, and Preservica on the other hand do not make their technical documentation openly accessible but it is available to customers. Among the three vendors, only Rosetta makes its documentation available to non-subscribers, including their API data model, system integration models, and user guides. DuraCloud and APTrust provide technical information on an openly available wiki (DuraCloud) or web site (APTrust) that includes knowledge bases of common issues, release notes, and detailed information on features and services. DPN provides an FAQ with documentation and code on their openly available GitHub. In lieu of a knowledge base, Arkivum has an online collection of case studies and white papers. Preservica, Chronopolis, and TDL do not have an open knowledge base. TDL had one in the past and may develop another one. MetaArchive provides technical specifications for hardware on its website, but in-depth technical resources and a knowledge base are currently restricted to users.

Table 1-Member Community[12] demonstrates opportunities for community engagement associated with each preservation service. Community discussion are monthly calls or online forums. In-person gatherings are annual or bi-annual meetings. Mailing lists distribute service and product updates.
3.1.4 Community Facilitation

Nearly all of the providers we spoke with have a community, user, or customer-focused staff position. This position is responsible for communicating with users about services, and for facilitating community relations, meetings, and events. All except Chronopolis and Arkivum provide mechanisms for community discussion and product updates, either via mailing lists, Google Groups, or subscriber user forums. Each organization in the sample has a program in place for facilitating conversations with and between users. MetaArchive hosts regular monthly community calls that provide an opportunity for regular communication between community members. DuraSpace publishes a quarterly newsletter that includes updates about DuraCloud and ArchivesDirect. MetaArchive, Rosetta, DPN, DuraSpace, and TDL hold annual user meetings. APTrust members meet twice per year. Preservica facilitates user group meetings at conferences, and Rosetta, in addition to hosting an annual user group meeting, hosts quarterly working group web meetings and meets with advisory groups as needed.

3.1.5 Outreach and Communication

Table 3-Outreach and Marketing[14] illustrates that all of the programs engage with their community via social media. Most use Twitter, but some also use LinkedIn, YouTube, or Facebook. The commercial providers and DPN also sponsor conferences, and exhibits as well as join their non-profit and academic colleagues as conference presenters and panelists. Arkivum attends sector-based conferences in medicine and other fields.

3.1.6 Marketing

Preservica, Arkivum, and Rosetta engage in marketing, though MetaArchive, TDL, and APTrust do not. DPN has published flyers and co-sponsored events such as PASIG and Digital Preservation 2016. DuraCloud has exhibited at conferences, but does not engage in marketing. Chronopolis is developing promotional materials, but has never exhibited at a conference.

3.1.7 Membership

There are slight variations in membership models and composition. MetaArchive has twenty-two members and over sixty participating institutions including consortia, academic libraries, public libraries, archives, and museums. Nearly two hundred institutions use DuraCloud, including cultural heritage and commercial enterprise users. DPN has over fifty members, including universities, consortia, and one commercial entity (Figshare). DPN anticipates future membership to include public libraries and cultural heritage institutions. APTrust has sixteen members, all of which are ARL member academic libraries, but is expanding membership to include public libraries and liberal arts colleges. Chronopolis is not a membership organization, but the replication nodes are at University of California at San Diego, National Center for Atmospheric Research, and the University of Maryland Institute for Advanced Computer Studies, and most users are at those institutions. TDL has twenty-two members drawn from higher education institutions, though they considering expanding their membership model. Arkivum serves approximately one hundred institutions, though some share a consortium account.

3.2 Technical Aspects

3.2.1 Setup and Configuration

MetaArchive members that host a storage server node follow specifications, instructions, and support to set up a server to connect to the MetaArchive network. Preservica provides a hosted Cloud Edition that requires no installation, as well as an Enterprise Edition that requires a local server and storage (e.g. Windows or Linux servers with Network Attached Storage) and a MySQL database for managing metadata. DuraCloud users set up an account and then utilize a web dashboard to manually ingest content, or they can install a DuraCloud Sync tool to automate content ingest. Chronopolis and TDL both use SSH or DuraCloud. APTrust uses an API. Arkivum sends hardware for the customer to install, and then Arkivum runs the software installation and configuration.

3.2.2 Content

Table 4-Content and Ingest[15] demonstrates that all programs in the sample are format agnostic in that they support all file formats, content types, metadata schemas, and structures. Furthermore, they all support large file sizes as well as BagIt bags. All of the services except for MetaArchive and ExLibris have a drag-and-drop ingest interface, though ExLibris and MetaArchive do have a simple graphic user interface for uploading files through a web browser.

3.2.3 Ingest

There is wide variability in ingest methods within the sample. MetaArchive has two ingest processes—one for public, live, web-based content, which is crawled and ingested via LOCKSS[16] plugins setup with parameters to automatically ingest repository content, and another for non-public content that members place on a web server with simple directory listings. APTrust partners bag their own content using the LOC BagIt specification, and then submit their bags through the APTrust API. Partners can track the bag’s progress through final ingest into AWS. DuraCloud users can add content manually via drag-and-drop to a web dashboard, or they can automate ingest through the DuraCloud Sync Tool. DPN, Chronopolis, and TDL also use DuraCloud software to upload content. TDL users can also use a desktop client that syncs to the network from a folder, or they can use a command line method that pushes from a server to DuraCloud. Preservica users can upload content via a web dashboard, a local SIP creator, a networked transfer agent, or Preservica’s Cloud Edition Bulk Upload Service. Arkivum customers move files onto a networked file share. Customers who use Arkivum’s Perpetua product can also use AWS for ingest.

3.2.3 Storage
which is managed by a local systems administrator. Each node
3.2.5 Access
firewall and port settings to restrict access only to other network
encrypted. The setup of MetaArchive server caches includes
MetaArchive member server cache communications are SSL
TDL uses SSH and requires login and access control settings.
RackSpace). All DuraCloud content uses https for encryption and
cloud storage service provider protocols (e.g. Amazon,
encryption keys. TDL and DuraCloud’s content security employs
configurations can be as granular as file level permissions.
Arkivum data uses file encryption and customer-supplied
encryption keys. TDL and DuraCloud’s content security employs
cloud storage service provider protocols (e.g. Amazon,
Archivematica, they can download ingest packages from that
interface. Preservica’s Universal Access module allows users to
search, find, and download publicly available content, though
details are not available for bulk download and restoration
processes. Preservica has many different levels of accessibility
that can be implemented, in addition to user-defined restrictions.
APTrust members restore content through the API, and TDL
members use the DuraCloud interface, but TDL staff retrieve
content from S3 and Glacier.

Nearly all of the programs operate as hosted services so that
members do not need to install, run, or maintain servers or other
infrastructure. Three of the services (DuraCloud, APTrust, TDL)
offer replication through Amazon Web Services Glacier and
S3. Preservica and Arkivum both offer managed hosted services,
as well as enterprise versions that run on local infrastructure.
MetaArchive, as a cooperative, has unique requirements for its
members, who must purchase and host servers to run LOCKSS
software and replicate other members’ content. This configuration
reduces MetaArchive’s membership fee by transferring IT costs to
member staff time and hardware. Alternatively, members can pay
a hosted storage service fee.

All services in the sample provide regular fixity checking.
Arkivum has the most frequent intervals with monthly checks.
TDL reported biannual checks. Most participants did not report
fixity check frequency.

3.2.4 Security

DPN security is managed by contributing partners. APTrust uses
data integrity and siloing through AWS to secure its data. APTrust
has scheduled a security audit for the coming year. Preservica did
not provide details about security beyond access roles and rights.
All users must authenticate with a username and password, and
each user account is set up with a series of roles which allowing
users to see content based on their defined roles. These
configurations can be as granular as file level permissions.
Arkivum data uses file encryption and customer-supplied
encryption keys. TDL and DuraCloud’s content security employs
copy files back out through a fileshare, or if the customers use
Archivematica, they can download ingest packages from that
interface. Preservica’s Universal Access module allows users to
search, find, and download publicly available content, though
details are not available for bulk download and restoration
processes. Preservica has many different levels of accessibility
that can be implemented, in addition to user-defined restrictions.
APTrust members restore content through the API, and TDL
members use the DuraCloud interface, but TDL staff retrieve
content from S3 and Glacier.

3.2.5 Access
Preservica, Arkivum, APTrust, Chronopolis, DuraCloud, and
TDL all restrict access through authenticated user credentials. MetaArchive’s access system differs in that each
member institution is a storage node and provides its own server
which is managed by a local systems administrator. Each node
hosts a copy of data from other institutions on the network, but
only designated system administrators have user accounts for
member server caches. No member is permitted to access another
member’s stored collections. Login pages are only accessible to
the host member institution and the MetaArchive central staff, and
SSH is required for remote access.

Each preservation system has a separate method for restoring
content. MetaArchive members submit a request to MetaArchive
central staff who establish a secure connection to the nodes with
copies of the requested content, and then the software constructs
an uncompressed ZIP package and makes it available for the
member to download. DuraCloud users utilize a retrieval tool to
download content to a local environment. Arkivum customers


4 CONCLUSIONS
4.1 Technical

Three main service models emerge in this study. Most of the
services follow a consistent hosted service model. APTrust,
Chronopolis, and TDL are all legally constituted within public
universities and use third-party applications such as DuraCloud
software or Amazon Web Services Glacier and S3. MetaArchive
is unique in its distributed network of nodes, where each
member/partner provides its own storage hardware and
contributes more staff time. In this regard, MetaArchive operates
more like a cooperative. DPN is different because it operates as an
independent umbrella organization that links other programs
together, adding further replication and geographic distribution.
4.2 Organizational

Prior to data collection we anticipated seeing significant differences between commercial and non-commercial providers, and these differences did manifest in organization and mission—some providers are privately held commercial firms and other providers have a non-profit model. The non-profit providers are either legally constituted within public universities or as independent LLCs. While mission statements and business models articulated by the commercial entities exhibit competitive business services around data preservation, the non-profit organizations used language expressing a direct interest in sustainability of the content itself for posterity’s sake.

Several of the services in the sample expressed plans or an interest in expanding their services to public libraries. This expansion is good news for the preservation of cultural heritage materials and data in public libraries. Among the sample the opportunity to expand services to this population will create new opportunities and challenges in coming years.

4.3 Geographical

All of the surveyed programs except Arkivum are based in the United States. Arkivum and all of its data are stored in the UK-based data centers. Understandably, the scale of size and distance are different in the U.S. and U.K. so Arkivum’s nodes are a few hundred miles apart as opposed to a thousand miles or more, as is the case with some storage nodes and third-party data centers in the U.S. In spite of this difference, Arkivum’s data centers are still far enough apart to have different disaster threats. Furthermore, Arkivum is the only service in the sample that offers offline secure storage in its preservation model.

While we are not aware of recent studies that parallel the scope of this one, it is a fertile area for further research. Some participants in this study are preparing to announce new services at the time of writing so a follow up study would yield different findings. Additionally, a larger and more diverse sample would be useful. Eight of the nine participants are U.S.-based, and one is based in the U.K. We became aware of other service providers that we would include in a follow-up, yet even these are exclusively within English-speaking countries. Aside from the scope, a few other subjects warrant further investigation. An analysis of cost versus effort could serve institutions investing in distributed digital preservation. Spalenka[18] writes that administrators or digital preservation program managers “should not assume that an application bundling many digital curation and preservation functions together with a single user interface will necessarily provide an entirely comprehensive and worry-free experience” because some service providers assume that users have already addressed the basic preservation needs. A tool that identifies existing resources and gaps in digital preservation architecture and matches that assessment with digital preservation services that best address the gaps would be a useful application of these findings.

Hopefully this study assists decision makers identify the best digital preservation service providers for their organization, and expands our participants’ understanding of the digital preservation service environment while also helping them identify strengths, weaknesses, opportunities, and threats of their respective programs and service models.

ACKNOWLEDGMENTS

We would like to thank Carly Dearborn (Purdue University), Sam Meister (MetaArchive), Mary Molinaro (DPN), Bradley Daigle (APTrust), Kristi Park and Ryan Steans (TDL), David Portman and Tom Hodge (Preservica), Sibyl Schaefer (Chronopolis), Matthew Addis (Arkivum), and Heather Greer Klein and Bill Bannon (DuraCloud) for their time, contributions, and cooperation in this study.

REFERENCES

[5] http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/61411/DDP_EnvironmentalScan.pdf?sequence=1
ABSTRACT
The field of digital archiving is situated in a tenuous position between the archives and information technology, where the humanity of archival labor is often erased on one side and that of users may be ill-considered on the other. To that end, the quality of technical documentation is paramount to the success of digital archiving efforts, but it often falls victim to a lack of empathy for practitioners, users of our tools and collections, and those represented in the archives. This paper aims to provide initial community suggestions for creating better, more useful documentation built around the concept of empathy.

KEYWORDS
Documentation, Digital Archives, Digital Curation, Ethics

ACM Reference format:

1 INTRODUCTION
In Rethinking Repair, Jackson presents two potential worlds: one is “always-almost-falling-apart”, while the other is one of constant progress, reinvention and improvement. The crux, he argues, is repair; it is the ongoing maintenance of our technologies [15, p. 222]. Indeed, we must “love our monsters” [19], but so often the focus on innovation and newness overshadows the maintenance needs of existing technologies. In the field of digital archiving—which, for practical purposes, we consider here as a related component of the greater digital preservation field—we feel the impact of technological degradation acutely and yet our own maintenance work is lacking. Ongoing development of technical solutions (including software, applications, and hardware), to the challenges faced in managing born-digital content within libraries and archives often neglects repair work. This is especially clear in the realm of technical documentation, a piece of digital archiving work that is not often incentivized. Further, often when technical documentation work is undertaken, it is done so to reflect only the knowledge, experience, and needs of the staff and institutions creating such documentation, which thereby leaves out large populations that should be represented in digital archiving work.

Drabinski [11], based on the seminal work of Bowker, argues that the fundamental act of documentation is one of building a professional narrative. In this case, we argue that the development of better, more empathic technical documentation practices is one of the most fundamental elements needed for building a more diverse field, which will in turn make us better at representing both our users and those who are included in our archives.

The field of archival science has recently explored the issue of empathy as part of the archivist’s role (e.g., Caswell & Cifor). We aim to build on this exploration by arguing that the same critical lens should be focused on our digital tools and practices. Additionally, there are extensive theoretical studies on representation in technology and in the online realm (e.g., Nakamura; Wolmark) and we seek to build a bridge between that theoretical work and the functional practice of those undertaking technical work in digital archives. There is a need for sustained inquiry regarding how we address issues of representations of race, gender, and inequality in technical arenas [17], and we argue that such an inquiry should be integrated into every aspect, rather than separated as its own subfield. The purpose of this paper is to provide initial suggestions for the digital archiving community in order to develop a more human-focused, empathic approach to technical documentation. We argue that by creating documentation that is underpinned by empathy for ourselves, our users, and those in our archives, we will improve our community, our tools and software, and our practices.

2 THE STATE OF DIGITAL ARCHIVING TECHNICAL DOCUMENTATION

2.1 Current State
A discussion of the current state of technical documentation in digital archiving must begin with the fact that, while documentation is ubiquitously known to be of importance from both a technical standpoint as well as a preservation standpoint, it is generally not prioritized. On an individual tool level, documentation is often prone to institutional specificity and the single-point-of-failure issue [12]. Specifically regarding the implementation of open source software (OSS), libraries cited poor documentation as second only to the need for skilled staff that could provide support for an OSS system in terms of barriers [9, 26]. In reviewing the principles of various technical development communities, this dearth of focus on documentation
is apparent. The Hydra community, for example, only mentions documentation in the context of code in its overarching principles, and does not provide much guidance to developers in creating usable documentation [28]. The Spruce Mash-Up Manifesto asserts that development projects should “make it easy to use, build on, re-purpose and ultimately, maintain” but does not specifically highlight the need for good documentation [27]. These well-intentioned principles serve a point by Bayer and Muthig, that often documentation processes are defined without justification, so those producing it are not clear about “what is being done with the documents they produce and for whom they are producing them” [3, p. 2].

In terms of documentation to support digital archival efforts at the institutional level, digital preservation policies tend to be the most ubiquitous, as growth in both administrative policy focus and digital content in cultural heritage institutions have helped solidify the need for these policies to exist [25]. Standards and certification also necessitate related types of high-level documentation that aid practitioners in understanding their institution’s specific landscape for digital archiving, such as collection development policies and mission statements. For example, many larger institutions strive to meet the requirements for Trustworthy Digital Repositories (TDR), for which there are extensive explanatory documentation requirements. Levels of TDR-readiness vary widely, however, and so does the quality of organizational and technical documentation that ultimately supports digital archiving.

At a digital archives practitioner level, the poverty of technical documentation is felt acutely in systems and tools that require the content creators and users to change their needs in order to be able to use the technical solution. However, the field is becoming more adept at requesting user-focused development with better documentation. In creating a list of Minimum Necessary Requirements for Developing OSS Tools for Digital Preservation, attendees of the OSS4Pres 2.0 workshop cited “Provide publicly accessible documentation” as the first item of necessity. “Keep documentation up-to-date and versioned” followed that requirement, as well as the need to provide “a documented process for how people can contribute to development, report bugs, and suggest new documentation.” Participants also suggested well-documented integrations with other tools and systems, and the need for validated use cases [10]. All of these items refocus development projects on the needs of users and future stewards of the tools and software.

2.2 Relevant Developments in the Field

In the larger field of digital preservation, momentum has been building in terms of reconsidering traditional beliefs and practices, embracing concepts like ethical collection practices and community building. These initiatives reflect the growing interest and need for preservation across communities, such as activists, laborers, researchers, artists, and others. New efforts by some of the digital archiving projects discussed below also reveal the diversity of both the user base intended to use the tools, software, and workflows being developed by the digital preservation community, as well as that of the new voices that digital archivists are striving to include in archival collections. Where archives of the past have often represented those with a certain level of power, recent projects have been working to establish the visibility of those accidentally and purposely erased from previous archival memory.

3.2.1 Documenting the Now. Documenting the Now (DocNow)¹, a Mellon-funded project that began in 2016, aims to collect and preserve social media associated with historic events. The project is unique in its focus on representing the needs and desires of content creators whose content it aims to preserve, delving deep into the new ethics of digital archiving. DocNow has specifically addressed issues such as the ethics of collecting data in the first place if it puts content creators at risk of harm or imprisonment. The project’s work is also unique in that much of the discussions are openly documented and available on platforms such as DocNow’s public Slack channel, which will allow the decisions made by the project’s team and collaborators to gather wider input from the community, and to have a strong and lasting impact of future ethics embedded in tools and practices for digital archiving.

3.2.2 Mukurtu. Mukurtu, an online digital content management system, has a stated goal “to empower communities to manage, share, preserve, and exchange their digital heritage in culturally relevant and ethically-minded ways” [20]. The project has contributed to the accessibility of technical solutions for digital archiving that ensure alignment with cultural practices. Began as a collaboration between researchers and members of the Warumungu community, Mukurtu provides access controls that align with the needs of cultural protocols, as well as traditional knowledge labels to prevent misinformation or appropriation of cultural objects and memory [21].

3 THE NEED FOR EMPATHY IN TECHNICAL DOCUMENTATION

3.1 Defining Technical Documentation

Our use of the term technical documentation covers that of tools, software, and workflows. Based on the Open Preservation Foundation’s (OPF) software maturity model, we define this type of documentation as including “source code, comments, technical documentation, installation manuals, user documentation” [23].

The concept of what is good in terms of documentation varies depending on the use cases and context in which that documentation is created [3]. In laying out a method for assuring the quality of documentation, Knodel and Naab [22], define content and representation as the two main elements for review. Good documentation knows its audience and purpose in terms of content, and has high levels of consistency, understandability,

¹ www.docnow.io
completeness, traceability across documents, and extensibility. The failure of documentation, on the other hand, is complex, but there are two main issues at play in the context of digital archiving. First, in archives, we often erase the impact of our work by not providing public documentation of our regular practices and ongoing work [1]. Second, in the context of digital work, Harihareswara [14] contends that the larger technology field “systematically undervalues the jobs and roles that require empathy and has deeply gendered associations with hospitality and empathy.” In digital archiving, then, it can be argued that we are prone to downplaying both the human element of our own work and the humanity of our users.

3.2 Empathy

In order to provide tools, software, workflows, and organizational solutions that actually work for people undertaking digital archiving tasks, reframing technical documentation in terms of empathy is paramount. While the term is somewhat diffuse in its meanings, here we utilize Coplan’s definition, which is “a complex imaginative process in which an observer simulates another person’s situated psychological states while maintaining clear self-other differentiation” [7, p. 5]. Additionally, we attempt to apply concepts from Kouprie and Visser [18], who demonstrate how a framework of empathy can be constructed as a pragmatic process by using the phases of “discovery”, “immersion”, “connection”, and “detachment” to enhance the empathy of designers for users, with the ultimate goal of improving the use and adoption of tools by others who are unlike themselves.

Focusing on empathy in this way requires imagining the needs of another person whose goals and lived experience are very different from one’s own, while still maintaining critical distance. This is key, as the differences between stakeholders using technical documentation within digital archiving are what will begin to push the field into a more comprehensively human direction. As Jules [16], aptly pointed out in the National Digital Stewardship Alliance (NDSA) Digital Preservation 2016 keynote, we must embrace the “unbearable whiteness” of our field, going beyond just race to address underlying issues of overemphasis on “professionalism” and “standards and technical know how.”

4 BUILDING BETTER DOCUMENTATION

There is an opportunity for those working in digital archives to take an approach to technical documentation that allows the field to move closer toward the inclusion of a wide community of practitioners of varying skill sets and experiences. Additionally, it is an opportunity to contribute to a more thorough understanding of our own developing practices, standards, community needs and sustainability. It is also important to acknowledge the ongoing discussion of adoption and technical documentation related to the OSS and library community [6, 9, 12, 26], which will undoubtedly continue to help shape the discussion of digital preservation community documentation at large.

4.1 Empathic Community Guidelines

In codifying community technical documentation guidelines based on empathy, we seek to provide initial steps that reflect both the need for improved documentation related to the adoption, maintenance and use of software and technical resources, as well as the lowering of barriers to find ways to better connect with the myriad users and preservation communities that create, maintain, teach and use the various systems and tools that further our work. These suggestions serve as a basis to think more comprehensively about everything from how-to documents for users to educational or technical workflow documents.

4.1.1 Establish a Method to Receive and Integrate Feedback, Iterations, and Updates. As Dowding et al. [10] argued, establishing users to verify tool, software, and workflow use cases is key, and we argue here that providing a mechanism for receiving ongoing feedback will add essential value. This concept involves some of the aspects of Kouprie and Visser’s [18] first two phases of “discovery” and “immersion”; user feedback and study is imperative for empathy immersion. To address these issues, we must be lucky enough to elicit user feedback and comments, this input should be used wisely. Organizations could take time each month to cull questions and answers from participants within a Google Group, listserv, Slack channel or internal meetings. This information could then be used to build out documentation and identify gaps in the information that is provided to the greater community. This interactive, iterative community approach has been shown to positively impact technical documentation [8].

The BitCurator community provides a robust example of how this type of approach could benefit both users and maintainers of the software. This community sees detailed public questions in their dedicated Google Group based on practical, user case studies on a regular basis that are answered by staff or other users with a good deal of care and technical expertise. BitCurator also provides a detailed “Quick Start” guide that accompanies their software that attempts to address the myriad issues than can arise with installation of the software [4]. This guide demonstrates an iterative approach to documenting, as a new version is released along with new versions of the software, and a wiki with documentation is also maintained.

However, though the community is very active, questions and answers with useful information are eventually buried within years of Google Group pages, and are not always included in the newest “Quick Start” guide or other technical documentation for the software. These questions range from troubleshooting hardware to detailed error reports and use cases. This not only raises questions about the sustainability of this type of community documentation approach, but can leave users - especially those new to digital forensics or digital archiving - to sift through years of questions to unearth a relevant solution. This approach can also create redundancies in answers from BitCurator project staff, and sometimes leaves questions unanswered. To address these issues, and to make use of the invested user and staff community, a regular, automated or scripted solution for a project as large as BitCurator could be used to extract pertinent or new questions and answers for use in iterative technical documentation such as the “Quick Start” guide. This way, documentation could better refine common issues gathered and observed from the larger community, define use cases, and identify patterns or gaps that technical documentation needs to address. For smaller communities or projects, this could be achieved manually with similar results.
4.1.2 Better Explain Errors and Provide Examples. Many of the following suggestions relate to the “connection” phase of the empathic process as defined by Koubrie and Visser [18], in that those writing technical documentation may connect with users by recalling their own experiences first learning code, or a new tool, and then using those experiences to aid documentation design. Empathy for users, therefore, may be translated pragmatically in technical documentation to include:

Providing clear explanations of common errors that those who are not yet well versed in a tool or project may encounter, and what these errors mean. This specific approach means that common user errors can be learned from and decoded via technical documentation, allowing for greater understanding of underlying concepts, or the structure of a tool or program. An example of this type of empathic approach can be found in the documentation and exercises in Learning Python the Hard Way, an online coursebook for learning the Python programming language. The course details common questions and errors from new users with each exercise, along with why and how these errors might occur [24]. Additional related examples include fairly minimal efforts that nevertheless can translate into highly useful concepts for diverse user bases, such as clearer and more verbose error codes that can allow users to seek distinct further information [5].

Providing pragmatic, specific examples that reflect a variety of skill sets and functions related to a concept, tool or software. Providing examples can benefit both those developing tools or software by serving as use cases or user stories, as well as those wishing to learn the tool itself. An example of this approach can be found in technical documentation for some application programming interfaces (APIs), such as the Digital Public Library of America², where samples and guides provide step-by-step instructions for the most basic functionality of the API, along with notes about why and how specific elements were used.

4.1.3 Lower All Possible Barriers. “It is open but not available.” Davidson and Casden [9] express this sentiment in recognition of the wide variety of resources and technical skill levels that are the reality for most cultural heritage organizations outside of large research institutions, and how this reality affects the adoption and use of open source software and tools. The authors argue that software cannot truly provide utility to a large and diverse community until gaps are identified and programmatically addressed to further lower the barriers of adoption. Perhaps most importantly, they pinpoint that “c[ommunity members who would help address many other types of users (serving many other types of institutional users) never have the opportunity to participate” when barriers to do so continue to be too high, and therefore the more inclusive community desired remains unrealized. Though technical documentation cannot solve all related barrier issues, parallel supporting arguments can be made in the form of improving the type and scope of documentation for these and other projects that reflect the needs and resources of a technologically-diverse digital archives community. In addition to intuitive programmatic changes and thorough testing with a variety of users, suggestions for documentation that also aids the continued lowering of implementation barriers include:

Moving from a lack of specificity to extreme detail. Technical documentation should show users more granular detail for everything from creating exact command line strings to visual examples that demonstrate settings for common operating systems. While this approach may undeniably take more time initially, this specificity will help developers, maintainers, and users, as it introduces a way of creating standard, documented steps than can be more easily reviewed if something goes wrong. This approach can then enable users to feel more comfortable investigating issues independently in the future or assisting others in the community with use. The web archiving subscription service Archive-It and BitCurator community documentation are good examples of this type of approach, and understandably so; their contributing feedback communities of users range widely, and both have worked through several iterations of development and documentation.

Defining terms clearly and avoid elusive language or jargon. The word “instance” provides an excellent example of a term used with regularity between developer, library, and various user worlds that can have a variety of meanings in different contexts. For a developer, this term may be used more frequently to refer to an object belonging to a certain class, whereas users may be more likely to use “instance” as a singular example. A reference guide or glossary for terms is a familiar concept, and would be a helpful addition to consider. Avoiding jargon within technical documentation is also imperative, as specialized language emphasizes differences and prioritizes the professionalization that Jules [16] argues against.

4.1.4 Establishing Better Timing for Documentation. In a study examining open source developer documentation, Dagena and Robillard found that when developers were encouraged to update their documentation with each code change, the practice “...led to a form of embarrassment-driven development, which in turn led to an improvement in the code quality” [8, p.127]. Koubrie and Visser also suggest, “a process of empathy in design practice requires a structured investment of time” [18, p. 447]. This investment of time and reflection informs the final emphatic phase of “detachment”; when the designer steps back out of the user world with an increased understanding that is then used to drive new design insight. Both concepts compel the idea that early, iterative documentation can help improve the work being done whether it is related to community code or use cases. Often, technical documentation is something that happens after all other work is done, and it is difficult for most to provide imperfect or “unfinished” work, even for internal use, much less for public consumption [2, 26]. However, it has been shown that following this type of strategy can lead to developers who, “...document their changes as quickly as possible after realizing that they often improved their code while documenting” [8, p. 130].

There is opportunity in the digital preservation community, that writing empathetic technical documentation early and often will strengthen the overall effectiveness of approach. This means that if draft documentation can also be considered a means to not just detail strategy or craft code, but also to structure empathic phases by integrating user feedback, recording errors and providing detailed examples, and critically reviewing approaches to an issue, it has an improved chance to more fully serve the greater community. This draft technical documentation need not be

---

² https://dp.la/info/developers/codes/
perfect; the point is to get the ideas, but tested, concepts down and allow them to be reviewed [2, 13].

4.1.5 Improve Community Availability. At the iPRES 2015 OSS4Pres workshop, one of the most important issues highlighted by participants related to the successful adoption and use of open source tool was the need for a publically-available, centralized technical documentation source that related to the more granular, problem-solving activities and issues that can arise when integrating OSS into digital preservation workflows. Requests for “end-to-end workflows” and “case studies” prevailed, alongside requests that documentation need also be kept up to date and versioned [10, 12]. Building on these recommendations, particularly that a public, centralized hub need not be built from scratch, but could instead be modeled after existing wikis such as the Community Owned digital Preservation Tool Registry (COPTR), a community hub might also include efforts to integrate with both international and national digital archiving and preservation initiatives such as the NDSA, the OPF, or the Digital Preservation Coalition. This approach would increase communication channels and encourage a wider variety of practitioners and communities to write, search and provide feedback for technical documentation. However, it is equally important that access and submissions to such a hub also stem from smaller or burgeoning sources of documentation such as the Society for American Archivists Electronic Records blog [3] or community digital archiving efforts, as these sources will further enrich it. Providing examples of useful technical or instructional documentation, as well as calls for specific documentation types might also stem from the hub to fill gaps as defined by the user community.

5 Conclusion

These community technical documentation guidelines represent the beginning of a broad structure based on empathy that can be built around the various digital archiving communities. Further work remains to be done, particularly around the need for building an empathy framework. However, approaching documentation in this way is necessary to improve sustainability, innovate practices, and remove technological barriers to better serve our growing community.

REFERENCES


2 http://coptr.digipres.org/
4 https://saers.wordpress.com/
Designing and implementing a digital preservation training needs assessment

Findings from the Bodleian Libraries’ institutional repository

Sarah Mason
Bodleian Libraries
University of Oxford
Oxford OX2 0EW
sarah.mason@bodleian.ox.ac.uk

Edith Halvarsson
Bodleian Libraries
University of Oxford
Oxford OX2 0EW
edith.halvarsson@bodleian.ox.ac.uk

ABSTRACT
This paper describes the design and pilot of a digital preservation training needs assessment at Bodleian Libraries (University of Oxford). The assessment was designed to establish training needs among staff, with the purpose of creating targeted digital preservation training based on gaps identified. Measurements for the assessment were developed around the DigCurV [7] Framework skill descriptors combined with a literature review of other skills frameworks. The measurements were developed into a set of interview questions, which was piloted on nine members of staff from the Oxford University Research Archive (ORA) in the winter of 2016. Findings from the assessment have influenced the design of training modules which frame digital preservation concepts around workflows in the ORA digital repository. The modules take as their starting point the awareness level, knowledge, practical skills, problem solving approaches and preferred learning style of staff within the ORA team.

KEYWORDS
Auditing, skills, workforce development, training, DigCurV, institutional repository, digital preservation

1 INTRODUCTION
This paper describes the design and trial of a digital preservation training needs assessment undertaken at Bodleian Libraries (University of Oxford). In the winter of 2016, the Digital Preservation at Oxford and Cambridge project 1 trialled a set of interview questions on staff members from the Oxford University Research Archive (ORA) 2. The purpose of the assessment was to record what staff knew about digital preservation, and which transferable skills could be drawn upon to deliver tailored staff training in this area.

Digital Preservation at Oxford and Cambridge (DPOC) is a two-year project funded by the Polonsky Foundation to research digital preservation policy, staffing and technical challenges present at Bodleian Libraries and Cambridge University Library. The project’s outcomes and recommendations will form the basis for long-term, sustainable digital preservation programmes at each institution. Each institution has appointed 3 Polonsky Fellows to undertake this work in the following roles: Policy & Planning, Outreach & Training and Technical Specialist.

ORAl the Oxford University Research Archive, is the University of Oxford’s institutional repository and maintains the scholarly output of its members. It preserves research publications, journal articles, conference papers, working papers, theses, reports, book sections, unpublished academic work, and more. ORA-data was launched in 2015 as an archival store to help researchers archive, share and cite research data.

Findings from the assessment revealed that awareness of digital preservation risks was good among ORA staff. This knowledge was often acquired through the practical experience of working with digital material in ORA. Interestingly, despite being able to describe core concepts using practical examples, “jargon” digital preservation terms were only recognised by members of staff who had undertaken formal academic training on digital preservation. Having digital preservation terms as part of a staff member’s vernacular helps with accessing digital preservation literature and concepts. However, the assessment illustrates that “jargon” is perceived as intimidating and can also create a barrier for entering into discussions in the digital preservation field.

Furthermore, the assessment found that staff felt uncomfortable speaking about digital preservation with depositors due to a lack of up-to-date knowledge of the technical end-to-end workflows used within ORA’s underlying submission and repository service. This lack of confidence is a key skill to address; one of the strongest motivators staff expressed for learning about digital preservation was the ability to provide good service and advice to depositors and users. As preservation plans for research outputs are required by many UK funding bodies, staff in ORA are increasingly receiving queries from academics about digital preservation. Based on these findings, DPOC concluded that digital preservation training was better delivered in-house, tailored to Bodleian Libraries’ local context, rather than outsourced to external providers. DPOC are currently developing customised training modules around ORA workflows and repository software to contextualise digital preservation within the service. The modules take as their starting point the awareness level, knowledge, practical skills, problem solving approaches and preferred learning style of staff within the ORA team.

2 ASSESSMENT DESIGN
2.1 Reviewing Skills Frameworks
Before designing a set of interview questions for the assessment, the DPOC project needed to identify key skills for staff working
in digital preservation. A number of skills frameworks were consulted including the ARA Competency Framework [2], the CILIP Professional Knowledge and Skills Base [6] and the DigCCurr Matrix from the University of North Carolina [11]. However, as there were no available interview question templates to work with, the DPOC team needed to design their own. The DigCurV Framework [7] formed the DPOC project’s list of skills for creating interview questions. DigCurV was chosen as it is the most comprehensive framework among those assessed, and provides good granularity as it is tailored around different skill sets for different types of roles (Executive, Manager and Practitioner lenses). Also each skill in the DigCurV Framework has been assigned a unique number, making coding of the literature review and interview transcripts clear and concise.

2.2 Refining DigCurV Skill Descriptors

There are 110 skill descriptors in DigCurV [7], far too many to condense into a concise set of interview questions or online questionnaire. The volume of skill descriptors in DigCurV posed an issue during the assessment design, as addressing each one would take several hours and become too disruptive for staff. DPOC therefore reviewed DigCurV [7] and the earlier frameworks, which produced a reduced list of 71 skills for managers and practitioners. Skills were removed if they applied solely to the Executive lens, which was deemed out of scope for the training needs assessment. Skill descriptors were combined where the DPOC project assessed that there was overlap between them or if a skill was too unclear and lacked clarification. Since the DigCurV [7] Framework lacks clear definitions for all 110 descriptors, the DPOC team created a glossary for the remaining skill descriptors.

A literature review [1, 3–5, 10, 11, 13, 16, 19] produced a final list of 63 of the most commonly associated digital preservation skills. A further review of Bodleian Libraries’ training programmes, policies and job descriptions, narrowed the list to a further 43 skills for use in the training needs assessment. This list, combined with the literature review, yielded a shortlist of the top 20 skills for digital preservation that focused mainly on:

- technical skills (KIA1.15, KIA5.1, KIA5.2, KIA5.4, MQA3.12),
- metadata (KIA4.5, KIA4.6),
- communications skills (PQ2.1),
- domain and digital preservation knowledge (KIA1.1, KIA2.5),
- preservation planning (KIA1.16, KIA3.4),
- the designated community, access and searching needs (KIA3.6, KIA4.1, KIA4.2) and
- legal requirements (PC1.1, PC1.2, PC2.1, PC2.4) [7].

Using this list of descriptors, the DPOC project created a set of interview questions which can be delivered in a manageable 1–1.5 hour session.

2.3 Methodology

The DPOC team considered a number of qualitative and quantitative methods for the assessment, and settled on running semi-structured interviews. The use of semi-structured interviews allowed participants to explore alternative areas of interest, aspects which cannot be adequately captured in a structured interview or online questionnaire [18, p. 57] [8, p. 29]. Interviews were run with one member of the ORA team at a time, as this provided participants with confidentiality not available in a focus group setting. Another strength of one-on-one interviews was that the interviewers received input equally from each participant [18, p.171]. The flexible nature of semi-structured interviews also allowed the interviewers to clarify problematic phrasing and vague terminology identified by participants during the interview. The feedback resulted in a second version of the assessment with revised questions; samples of the question revisions are in Table 1. It was trialled in a second round of interviews in April and May 2017; results from these interviews are currently being analysed.

Two different sets of interview questions were designed and used in the pilot interviews, depending on whether or not a participant was considered to be a manager or practitioner. This differentiation was based on role responsibilities, not the participant’s job title [12]. Later, a manager/practitioner combined set of interview questions was added; together, these formed the training needs assessment that would be trialled on the ORA team.

<table>
<thead>
<tr>
<th>Table 1: Interview question revisions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original question</strong></td>
</tr>
<tr>
<td>What is your understanding of a digital file format?</td>
</tr>
<tr>
<td>Using your understanding of a digital file format, what formats do you work with in your role for your collection materials and administrative tasks?</td>
</tr>
<tr>
<td>From your professional awareness or experience, can you describe to me an example of a complicated information rights issues for a digital deposit?</td>
</tr>
<tr>
<td>Thinking about your digital collections, are you aware of them having any formal accreditation, certification or being subject to compliance audits?</td>
</tr>
</tbody>
</table>

3 PILOT SAMPLE

The DPOC project interviewed 9 members of the ORA team using the first version of the interview questions. The interview sample included:

- 2 managers,
- 5 metadata assistants,
- 1 senior metadata assistant, and
- 1 curator of research data.
The metadata assistants in ORA have a broad range of professional and academic backgrounds. Their time working with the service ranged from 5 years to a few months—although some have previously worked for Bodleian Libraries in other roles. Notably, one of the metadata assistants had completed a Masters level module on digital curation, and was able to respond in more detail around digital preservation concepts and terminology.

The interviews were completed over a two month period with each interview averaging between 1–1.5 hours. The same two members of the DPOC team took part in each interview to ensure that questions were delivered consistently across the sample [17, p.1193]. This method allowed one person to focus on accurate note taking, negating the need for tape recording, which the Oxford DPOC team felt would be too cumbersome to transcribe and had the potential to alter interviewees’ responses [15, p.169].

4 FINDINGS FROM THE PILOT INTERVIEWS

4.1 Core DigCurV Skills

The assessments identified a number of core skill strengths as identified by the selected DigCurV [7] skill descriptors. Not surprisingly, due to the type of material ORA staff work with, all staff had a good understanding of rights management, as well as strong communication and metadata editing skills. Staff also displayed good problem solving skills when required to learn new technology. Two members of staff had experience of using UNIX command lines and scripts.

4.2 Digital Preservation Risks and Concepts

Staff were able to give a number of examples of digital preservation risks, drawing on experience of their role reviewing digital file submissions for ORA. The most common risk areas described were hardware obsolescence (55% of respondents), file format ‘obsolescence’ and software obsolescence (55% of respondents). One staff member also mentioned storage failure and low quality metadata as risks to digital material. Although staff were able to describe these risks, only two recognised the digital preservation terms and concepts listed in one of the questions (the list contained references to the OAIS model and concepts, normalization and fixity). In several cases, the interviewers were able to frame what these concepts referred to by pointing to examples already supplied by staff earlier on in the interview. This led the DPOC team to start considering the possibility of using these work experiences and the ORA workflow as an entry to digital preservation concepts.

4.3 Digital Preservation and the ORA Service

While many preservation risks were understood by staff, there was much uncertainty about how risks can be mitigated. A re-occurring theme throughout the assessment was that staff did not know, or had out-dated knowledge about technical workflows and preservation practices associated with ORA, including the role and implementation of Bodleian Libraries’ Fedora instances and Hydra. Opinions were split among the 9 staff interviewed regarding whether or not ORA carries out digital preservation activities: 4 members of staff were unsure about the answer, 2 thought that ORA is doing basic bit-level preservation and 3 thought the service currently does not do any preservation activities (see Figure 1).

4.4 Manager Buy-in

Both managers showed a clear understanding of digital preservation, associated risks and techniques to mitigate those risks. This finding was encouraging, as it demonstrated managerial commitment to digital preservation. One manager expressed awareness that confidence with digital preservation was lacking in the ORA team, and believed that training would ensure staff could “answer queries and understand the wider environments which they work in.” This support for digital preservation training included the manager’s willingness to release staff to attend any available training sessions and provide additional feedback on session content.

4.5 Practitioner Problem Solving

The practitioner set of questions delved into problem solving approaches. When metadata assistants were asked how they would handle a scenario of reviewing a digital file with an unknown or missing file extension, their responses showed that most were comfortable experimenting with digital files and doing online self-directed research. The questions around problem solving approaches also revealed that 5 out of 6 metadata assistants relied heavily on
the knowledge base built up within their local team to address scenarios that required them to learn a new task or program. Several staff explained that if they received problematic or unidentified digital files, they would initially ask their immediate team for advice to see if anyone had previous experience of that file format.

4.6 Practitioners’ Preferred Learning Styles

The importance of the local team was also evident when staff were asked about their learning style preferences. The metadata assistants expressed that they preferred small groups for interactive training modules, but that larger groups for lecture-style training was good, provided they were not required to interact with other participants. This often reflected confidence in interacting with a new subject. One participant commented that “I would feel more comfortable having training with my closest team [the metadata assistants]” and another that “I don’t like working in a group with people I do not know.” However, in terms of general introductory modules, the same participant commented that “I think it is best to open it to all [staff], since then you get a much more varied perspective on things.” This suggested that for general awareness topics a large group approach is acceptable, but that ORA-specific or other in-depth workshops should be kept small and be limited to the core practitioners.

5 MODULE DEVELOPMENT

Based on the findings and skills profile of ORA staff, the DPOC project are currently developing training material for a tailored digital preservation module around ORA. Developing the training module for ORA has proven to be an interesting exercise in itself, as the DPOC team members are new to their organisations. Training is therefore being developed alongside a technical repository review of ORA as part of the DPOC project [14]. This has required extensive collaboration with technical support staff around Bodleian Libraries, in order to inform the DPOC team of ORA’s workflows and underlying infrastructure. For organisations looking to frame their digital preservation training in a similar fashion, it is evident that getting buy-in and advice from a variety of managerial and service support staff will be necessary to develop meaningful, targeted training modules.

5.1 Tailored Training Module: ORA

The module currently in development covers digital preservation concerns in end-to-end ORA workflows, from submission of research outputs by academics, to long-term storage and management by the library. It also includes explanations of how bit-level preservation is addressed in ORA’s current storage infrastructure. DPOC hope that framing digital preservation in a familiar context will help contextualize and introduce digital preservation terminology, and directly impact staff’s understanding of their own role within the service.

5.2 Hands-on Workshops

As well as having modules around the ORA workflows, the DPOC team are preparing hands-on workshops for analysing digital files using tools like DROID, JHOVE, VeraPDF and the BitCurator suite. As some staff have command line experience, a mixture of GUI and command line tasks will be run. Although these tools will not be used by reviewers on a day-to-day basis, knowledge of how they work and why they are useful tools will help staff when speaking about digital preservation to depositors and users.

5.3 All-staff Digital Preservation Awareness Module

During the assessment pilot phase, there was an interest in developing digital preservation awareness training that could be delivered to all library staff, not just those currently working with digital collections. Due to this demand, an adapted and brief online questionnaire was later sent out to all members of staff at Bodleian Libraries. The questionnaire enabled the DPOC team to triangulate some of the themes which were raised from qualitative interviews with ORA staff [9]. They used the findings to develop an all-staff digital preservation awareness training module.

This module includes a unit on personal digital archiving, which has been run multiple times. Several other units are currently being developed on digital preservation terms and concepts—these will also cover some basic language from the OAIS reference model. Although the assessment has shown that OAIS terminology is not necessary for identifying digital preservation risks, the DPOC project would like to explore if familiarity with OAIS terminology makes participation in discussions more accessible. These awareness modules will mix lecture-style, interactive and online delivery methods.

5.4 Training Evaluation

The success of the ORA and general digital preservation awareness training will be measured by pre- and post-module evaluation. Where two versions of a similar module or hands-on workshop will be trialled, one metadata assistant will attend and evaluate the two versions, facilitating revisions and improvements to the training materials. This method will allow the DPOC team to ensure an effective training programme is in place at the conclusion of the project.

6 NEXT STEPS

Alongside designing the ORA digital preservation training itself, the DPOC project will also be developing and trialling the next phase of the training needs assessment. The pilot trial of the sets of interview questions on ORA highlighted that there is also a need to design one for software developers working with digital collections. The current practitioner interview questions were found to be unsuitable because its focus does not align with skills required for a software developer. Further research is required to define the ideal skills set for software developers and other technical staff working with collection material. Once done, a specific set of developer interview questions will be developed. When completed, it will be trialled on developers working with ORA services.

The interview questions, online questionnaire and glossary of skill descriptors have been developed into a draft training needs assessment toolkit. This will enable other institutions to audit the skills and training needs of their staff. The authors invite comment and feedback on the draft toolkit, available on the DPOC project website: www.dpoc.ac.uk.
7 CONCLUSION
Findings from the assessment pilot have been invaluable for developing training modules around digital preservation at Bodleian Libraries. While there are a number of outsourced training solutions available, the assessment revealed that training modules benefit from being tailored to Bodleian Libraries’ local context. Staff in ORA want and need to be informed about their digital services; they also want to learn about digital preservation in relation to their digital services. While developing training is more labour-intensive, tailored and targeted in-house training will provide staff with the knowledge to speak confidently to users and depositors about their digital service. As the amended set of interview questions is rolled out to other members of staff in Bodleian Libraries, the DPOC project will be able to see whether or not this trend continues across other teams within the institution.

CONTRIBUTORS
Sarah Barkla, Bodleian Libraries
Alan Mckechnie, Bodleian Libraries
Jason Partridge, Bodleian Libraries
Thomas Stanbury, Bodleian Libraries
David Tomkins, Bodleian Libraries
Hilary Tricker, Bodleian Libraries
David Watson, Bodleian Libraries

ACKNOWLEDGEMENTS
The authors would like to thank the contributors for their time, ideas and feedback for this paper. They would like to thank the entire ORA team for being generous with their time during the training needs assessment. Also, thank you to the DPOC Polonsky Fellows at Cambridge (Lee Pretlove, David Gerrard, Somaya Langley) and the rest of the DPOC team at Oxford (Michael Popham, James Mooney). And, lastly, they would like to thank the Polonsky Foundation for their generous funding of the DPOC project.

REFERENCES
Adding Emulation Functionality to Existing Digital Preservation Infrastructure

Euan Cochrane
Yale University

euan.cochrane@yale.edu

Jonathan Tilbury
Preservica
Abingdon, United Kingdom

jonathan.tilbury@preservica.com

Oleg Stobbe
University of Freiburg
Freiburg i. B., Germany

oleg.stobbe@rz.uni-freiburg.de

ABSTRACT

The emulation of obsolete hardware and software environments to enable information to be read, and to facilitate interaction in a way that simulates the original user experience, is a well-established part of digital preservation solutions. Excellent tools have been developed to work with emulators, but these have remained in the research domain rather than being able to be exploited at scale. This paper explores a real example of how an emulation framework has been added to existing digital preservation infrastructure. This integration has enabled information to be extracted from obsolete hardware, held in a digital preservation system at Yale University Library (YUL), and linked to an appropriate emulator (provided by the University of Freiburg’s Emulation as a Service framework). All of this enables YUL to recreate the user experience of interacting with content using the original software quickly and easily whenever a user requests it. This integration offers the prospect of large scale emulation as a service linked to real preserved data that is in need of this approach and this paper will examine the next steps needed to make this a reality.

KEYWORDS

Software-based Art; Emulation; Preservation Strategy

1 INTRODUCTION

An emulator is “hardware or software that enables one computer system (called the host) to behave like another computer system (called the guest)” 1. Emulators enable users with newer computers to run software designed for older computers, software that may be incompatible with their current computer. In the context of digital archiving, preservation, and access, emulators have a number of useful applications. However, emulators have, for a long time been difficult to configure and use, especially at scale, and require expertise that is rare amongst organizations that have taken responsibility for preserving digital content at scale. This (amongst other reasons) [1] has limited the acceptance of emulation as a viable long term digital preservation method. Fortunately in recent years tools that simplify the use of emulators in these contexts such as JMESS 2 and bwFLA Emulation as a Service (EaaS) [2], have become more readily available and are increasingly being used in production environments. With these emulation tools and services coming of age it is becoming increasingly realistic for digital preservation infrastructure providers and developers to consider how emulation can fit into their products and services. In this paper we describe how Emulation as a Service has already been connected with digital content stored in one digital preservation system Preservica, and discuss the work needed to further integrate and scale this approach for wider use.

2 CONNECTING EMULATION AS A SERVICE (EAAS) AND PRESERVICA: A CASE STUDY

Yale University recently implemented functionality to connect its new bwFLA Emulation as a Service (EaaS) implementation with its digital preservation system Preservica to initially provide access to content on CD-ROMs from Yale University Library’s (YUL’s) general collections. EaaS is a suite of software that simplifies the use of a variety of emulators enabling minimally trained archivists and librarians to make use of emulation technology in typical archival and library workflows. And it does so without requiring significant technical expertise or adding management burden to already busy work schedules by enabling access to preconfigured emulated computers via a web-browser interface. The EaaS software emulates a variety of different computer architectures behind the scenes so the power and value of the EaaS approach to the archivists or librarians it that it abstracts away the details of which emulator is being used and how it is configured, and simply provides the preconfigured emulated computers for use in archival/library workflows.

Preservica 2 is a digital preservation system that combines all the elements of the OAIS model 4 into a single system. It includes tools to ingest both simple and complex data objects, for example ISO disk images, and to store these in multiple data stores with full fixity checking. It also includes a flexible data management capability with an access module to allow the information to be searched, browsed and downloaded. At its core is a full file format preservation suite that identifies and characterises content and if applicable migrates it to new formats. It also includes a number of viewers to allow content to be rendered server side and delivered via a browser. This is built on a registry of file format information.

The initial connection implemented between Preservica and

EaaS is a ‘lightweight’ integration using the standard Application Programming Interfaces (APIs) available in each product allowing access via an emulated computer by combining the following:

1. An emulator
2. A hard drive image with an operating system such as Microsoft Windows 95 installed on it
3. (Possibly) some software for interacting with the content such as Adobe Acrobat Reader

At Yale University Library (YUL) these assets are preserved within Yale University’s local “Enterprise Edition” installation of Preservica. Within Preservica there are a number of collections that preserve the required assets:

1. “Base-Images” collection
2. “Software” collection
3. “Derivatives” collection
4. One or more content collections

The Base Images collection in Preservica contains hard drive images (single files that capture all the content on a hard drive) that have a minimal set of software installed on them such as an operating system. These, along with a configuration file, are all that is necessary to provide a basic emulated computer via one of the emulators available within EaaS.

Base Images are not provided as part of the bwFLA EaaS software suite and have to be curated by users of EaaS. At YUL the images were created by cloning or ‘imaging’ hard drives from original computers using digital forensics software (See Figure 1).

In most cases these images function without any changes within the appropriate emulator. In some cases, the images need to loaded in the emulator and have new hardware drivers installed in order to be compatible with the new emulated hardware.

Base Images could benefit from centralized solutions. Imaging original hardware to obtain base images creates an emulated computer environment is as close as possible to a functioning original machine ensuring a maximally authentic experience for the end-user in such a way that the user can seek out the original hardware to validate the experience of the emulated version against the original. However a wider implementation of this approach would not be sustainable even short-term due to a lack of available original hardware for each organization and long-term due to hardware degradation through natural causes. In the short/medium term this could be mitigated by establishing a network of hardware museums that could provide a valuable source for base-images and emulation validation services and in the longer term through thorough documentation of the original machines that will help to enable testing of future emulators. The preservation of the base-images will further ensure we have maximally authentic emulation experiences available.

To access Base-Images the EaaS service can be triggered to pull new content (using the Preservica REST API) from any pre-configured collections in Preservica and organize it into its administrative interface using buttons in the EaaS GUI.

Figure 1: Digital forensics software and hardware being used to capture the content of hard drives from original computers.

Users can then use the EaaS GUI to configure base environments. Once base images have been created or captured, and added to EaaS, staff can then add software applications to them. YUL’s digital preservation policy framework states that: “YUL will ensure access to hardware and software dependencies of digital objects and emulation or virtualization tools by [...] Preserving, or providing access to preserved software (applications and operating systems), and pre-configured software environments, for use in interacting with digital content that depends on them.”

YUL has begun preserving software in compliance with this policy statement within a “Software” collection in Preservica. The “software” collection contains a set of installable software binaries used to install software. These in turn can be used to provide a maximally authentic rendering and/or interaction experience for researchers to access preserved digital content. An example might be the installation file for Adobe Acrobat Reader 4.05 from the year 2000. Currently the software installation binaries are all preserved in Preservica wrapped within disk image files (e.g. ISO files) to simplify their use with EaaS. However it is possible to submit files or folders of files to EaaS and have it automatically wrap them within a virtual CD or Hard Disk Drive (HDD) image and attach them to an emulated computer. The software binaries in the YUL collection are organized in a standardized structure with standardized metadata. To capture and arrange that metadata YUL is using a custom Metadata Encoding and Transmission Standard (METS) implementation that references a persistent external master database, Wikidata.org, used to capture and store generic software documentation (such as title, publisher, input and output formats, etc). The METS file is also used to directly store localized information such as license keys, and local holdings information such as local media identifiers. In future iterations of this integration the additional software metadata necessary may be cached locally in Preservica’s internal preservation metadata registry which already documents some relevant software applications.

---

3 BIOS files and other cross-environment configuration files are considered part of the EaaS framework.
5 Policy%20Framework%20V1%200.pdf
After ingesting a new binary or set of installation media into Preservica archivists or librarians can click in the EaaS GUI to "sync object environments" and import the software into the EaaS interface. They can then launch an emulated base environment and have the installation media for a selected piece of software automatically inserted into the emulated computer’s virtual drive (either within a virtual CD/DVD drive or an emulated hard drive).

The staff member can now use the standard software installation process to install the software, shut down the emulated computer, and click a button in the GUI to save a new "software environment". The software environment is now ready to be exported for ingest back into Preservica for long term Preservation and future retrieval. This export and ingest process is currently manual however it is expected to be automated in coming months.

The software preservation workflow at YUL requires there is a basic record for the software in Wikidata, that the wikidata reference is in a METS file for the software, that it and the media images are ingested into Preservica, and that the software collection is synchronized with the EaaS GUI by clicking a button in the interface. Following this the software is installed and configured on a base-environment to create a "software environment". EaaS is then used to execute the software and document further metadata about it, such as the software’s import and export formats (or `open` and `save-as` as the case may be) and the default format associations of the software as assigned in the operating system (in order to know which software will execute by default when a file of a particular format is opened). As of this article’s publication date this metadata is being stored locally but in coming months it will be added to Wikidata to be shared with the wider digital preservation community.

Software environments can also be layered on top of one another enabling the creation of software environments that meet many different needs. As a worked example a user may configure a software environment that includes Microsoft Outlook in order to access the content stored in "PST" email archive files in their original context. On accessing the PST file they may discover that it contains a lot of .psd (Photoshop) files as attachments to the email. They can then go back to the Outlook software environment and layer Adobe Photoshop on top of it to create a new software environment that includes both Outlook and Photoshop. This will ensure that the Photoshop files attached to emails in the PST file can be accessed by users browsing the email records. Each of these shareable and reusable software environments can be exported and preserved in Preservica for preservation and future utilization.

YUL are initially using EaaS for providing access to content on CD-ROMs (CDs) that make up part of their general collections and date back as far as the late 1980s. The CDs contain a huge variety of content from government and business datasets, to interactive video and image content, to conference proceedings, and even computer games. The software requirements for accessing the content on these CDs are fairly minimal as the CDs were generally designed to come bundled with all software that was necessary to use them and simply require standard operating system base images along with some freeware such as Adobe Acrobat, VideoLAN Player (VLC), or Netscape Navigator.

3 ADDING CONTENT TO SOFTWARE ENVIRONMENTS

After configuring software environments in EaaS users can synchronize content from any collection in Preservica that EaaS has been configured to pull from, and which is structured according to a predefined (but configurable) substructure pattern within a Preservica collection allowing the EaaS automatically identify where the installation media are. METS metadata is also included to document information about which disc of a multi-volume set of media needs to be inserted into the drive first. This can be important when only one disc from a set includes the setup files or software, or when one disc is required to initiate an interactive experience such as a game or encyclopaedia, before switching to an additional disc to provide additional content. The documentation shows how the end-user GUI for EaaS can provide a button enabling users to change discs when interacting with the environment.

Once software environments have been configured and the content collections ingested into Preservica and synchronized to EaaS, content can then be associated with software environments to enable it to be automatically accessed by end-users in an appropriate software environment. Using the Emulation as a Service software organizations can choose between implementing three different workflows for matching emulatable software environments with digital objects in their collections:

1. Automated
2. Semi-automated

Using the automated workflow the user selects an object to interact with via emulation. The content of the object is characterized, extracting file formats and creation dates of the files it contains, and matched to a database of available software using metadata held on each environment and an algorithm that attempts to find the environment containing the most compatible software. The results of this matching are either saved for a future request or are discarded and regenerated next time the object is requested. The latter approach may be advantageous for archives that are growing their software collection as at any point in time
the "best" environment for interacting with any particular object may have changed as new software was acquired.

Sometimes multiple matches may be relevant for a set of content, for example some CDs from the 1990s included both Apple Macintosh compatible content and Microsoft Windows compatible content. In such cases the user can be given the choice of which environment to attempt to execute the content within.

With a semi-automated solution, the characterization and matching happen by default but a staff member confirms that the recommended environment is actually a good match for the files stored on the object by loading the environment and trying it. If there is an appropriate match the staff member can save the configuration as detailed below to be automatically provided by default to future users.

In the third, manual scenario, a staff member manually selects an environment for each object and configures it themselves. This is useful for complex environments or environments in which specialized software is provided with, or required by the object, or where custom environments need to be created for the content.

A core concept utilized in this implementation is that of 'derivative environments'. Derivative environments are complete software and/or content environments that are derived from and depend on a base-disk image that contains the full operating system installation and configuration files required to run the content contained in the derivative environment. The 'derivatives' collection in Preservica is where derivative environments, content files, and metadata files are preserved after being configured in EaaS. Derivative environments are created by staff in the processes outlined above whenever a software application or set of content is added to a base environment or software environment and a button is clicked in the EaaS GUI to save the environment as a derivative "object environment" (Fig. 6).

When saved as a derivative environment the EaaS software encapsulates the small set of data that was added to the base environment, or existing derivative environment, along with metadata instructing EaaS how to reinitialize the base with the derivative in real-time when requested in by users in the future. Periodically, derivatives can also be exported and re-ingested to Preservica for long term preservation.

4 ACHIEVING SCALABILITY

The use of EaaS with Preservica at Yale University is now relatively seamless and requires few advanced technical skills from the staff who will be continually configuring new derivative object environments as YUL adds more content to the system in the future. Such configuration is currently being undertaken by digital preservation staff but given the non-technical nature of the work involved, this task may be assigned to other library and archives staff in the future. For end users, the ability to click a link in a web browser and have an historic computing environment seamlessly load within the web browser is both minimally burdensome, and transformative from a research perspective. However, this ease of use for these users (be they researchers or non-preservation librarians and archivists) belies significant preparatory work on behalf of both the digital preservation team, the EaaS software developers and the community of metadata creators whose contributions behind the scenes enable this approach. Fortunately, there are many alternatives for scaling these activities to share the load and achieve economies of scale for all digital preservation practitioners to benefit from.

This use of derivative environments, instead of capturing full copies of the entire environment each time a change is made, has two major benefits that may enable scaling of the use of EaaS in the future. Firstly, it minimizes the additional data storage required to maintain large collections of custom content environments (environments pre-configured to load content such as a single file or set of files on execution). Secondly, it further simplifies and extends the digital preservation community’s options for scaling and providing software and content services using emulation and lowers the storage requirements while doing so. It also offers potentially advantageous opportunities for the seamless provision of services software and emulation tools across multiple organizations. For example, this approach could enable one provider to host base environments, another to host the software-layer derivatives that point to the base environments, and a third to host the content environment derivatives that point to the software environments. All of those layers could then be brought together using an EaaS implementation hosted by an additional provider and all of this could happen seamlessly and in real time from the end-user’s perspective.

The UNESCO PERSIST project 6, The Software Heritage Foundation 7, and The Software Preservation Network 8 are working to enable access to software archives. Using the EaaS functionality that enables the layering of derivatives, and their local preservation environments, they could potentially each play a role in providing software for use in emulation solutions without requiring end-users to each preserve the software themselves.

The involvement of these larger bodies promises the possibility of a solution to one of the largest outstanding issues, the permission required by the Intellectual Property owner of the original operating system and software to operate the system. [3] By encouraging these IP owners to allow access to older versions of their software under given licence terms, EaaS becomes a commercial reality. The licence terms are likely to dictate who can use the software and under which circumstances, and a centralised service could enforce these licence terms in a consistent manner to the satisfaction of the IP owner.

Applying the semi-curated approach YUL has implemented for it’s 6,000-10,000 CD-ROMs would not be sustainable for a larger scale use of EaaS for providing access to single files. In contrast migration enables content to be reused by ensuring it is available in file formats that modern software can interact with, and Preservica already provides comprehensive tools to automate migration at scale. However, emulation is sometimes necessary e.g. where custom software is involved, such as is the case with many of YULs CDs. In other cases, emulation is desirable due to its ability to provide access to the content using original software,

---

ensuring rendering or interaction issues are minimized. Unfortunately, manually creating custom content-environments for each file in a collection of millions of files would not be viable due to the cost. However, a future potential large-scale application of EaaS would be to implement it as an automated "universal viewer" within preservation and access systems. This approach might entail file format information being provided to EaaS from a digital preservation system and EaaS automatically associating the file with an appropriate preconfigured software environment. To enable this the digital preservation system would have to be able to provide the file format information via its API. Alternatively, the digital preservation system could do the matching with an external software-environment library during the process of ingesting the file into the digital preservation system. In that scenario, the digital preservation system could then just provide the content file and the identifier for the environment needed to interact with it via its API to EaaS. The environment could then be edited in real time to wrap the file in a disk image, attach the image to the environment and insert a link into e.g. the Windows start-up folder (in the case of a Microsoft Windows-based software environment) on the main emulated hard drive. This would force the system to automatically open the file when the emulated computer was loaded. Scattered across many software applications this approach thus has the potential to provide a 'universal viewer' service.

Within Preservica, the 'Universal Viewer' application of emulation technology could also be integrated within the system by including the (open source) EaaS framework as a Preservica toolset to sit alongside other simple file viewers. In Preservica processes can be scaled horizontally (enabling multiple processes to run concurrently) by adding additional processing servers. This is a relatively simple but semi-manual step. If EaaS were integrated more tightly into Preservica a dynamic scaling mechanism may be required such as the ability to spin up additional processing servers in the cloud on demand for use in emulating additional computers. However, investment to achieve that result may be advantageous for a number of reasons. In particular the association between software environments and file formats, and the preservation of derivative environments and other assets in the emulation stack could be more tightly integrated and seamlessly automated within the Preservica system. The addition of technology seamlessly linking EaaS technology into Digital Preservation products such as Preservica is dependent on a resolution of the licencing issue of the operating systems and software packages being run.

5 CONCLUSION

The marrying of EaaS and Preservica at Yale University has enabled Yale University Library to ensure its content can be made accessible via emulation while simultaneously being preserved in a robust digital preservation system. In addition, the base images, software and derivatives can all also be preserved in a trustworthy digital preservation system with little additional staff effort. The work to reach this point has been significant but the existing tools, 

Preservica and bwFLA Emulation as a Service, with their generic APIs and sophisticated out-of-the-box functionality, have made much of this integration relatively straightforward to implement.

There are still gaps in the emulation and software preservation services landscape that will prevent less well-resourced organizations from implementing a similar approach. Fortunately, the features available in EaaS imply clear pathways to implementing software archives as a service, pathways that may be palatable to the software IP owners. In addition, the benefits of pursuing this and applying EaaS to create a "Universal Viewer" are hopefully obvious to all.

Further integration of emulation services into digital preservation systems such as Preservica seems inevitable. Fortunately, their existing functionality, such as Preservica’s preservation metadata registry, standard APIs, and ability to scale horizontally, ought to make this a reasonably straightforward process.

REFERENCES


10 Note that emulator hardware demands are fairly light when emulating older computers. Multiple emulated computers can share one CPU thread and memory (RAM) requirements are insignificant (e.g. 16 Megabytes for 1990s era IBM-compatible PCs)
Phantoms of the Digital Opera:
The need for long term preservation of born-digital actors and multimedia objects using methods that permit ongoing new creations

Dena L. Strong
Technology Services
University of Illinois at Urbana-Champaign
1304 W. Springfield, Urbana, IL 61820
dlstrong@illinois.edu

ABSTRACT
While the rapid improvement in computing power and digital tools have offered a vast new realm of creativity to more independent creators than ever before, that same rapidly changing pace also offers a unique threat to those digital creators: What can you do when the length of time you need to finish your project is longer than the length of time your computer’s OS or software is supported? How can a small creative or research group best ensure that their work can still be actively worked with in five years – or twenty-five years?

In this paper, I’ve assessed digital preservation recommendations over time, discussing the gap between the recommendations for preserving a completed work and digital creators’ ongoing needs to be able to create further work. I have three case studies created from longitudinal interviews with digital creators to determine how their own creative ecosystems have held up to the pressure of time, where their systems have begun to degrade, and how they continue their digital work across years or decades.

Currently, either maintaining an aging system or recreating entire worlds with different, newer systems are the best available options for the digital creators I interviewed, though neither is optimal. Hardware and software vendors’ intentional dropping of backwards compatibility prohibits accurate forward migration for many digital media creators. The Pericles model seems promising, but Pericles itself is not open to individuals. Future improvements in emulation, including cloud-based virtualization, may offer a path forward if these services can improve their user friendliness and resolve the version compatibility issues that create obstacles today.

Keywords
Long term digital preservation, digital video creation, digital animation, translation, version compatibility, emulation, digital lifecycle

1. INTRODUCTION
In 2016, Makoto Shinkai made waves throughout the digital animation world by releasing a record-breaking feature film called Kimi no Na wa (Your Name); he was the director, the writer, the cinematographer, and the editor, and it became Japan’s highest-grossing animation title and fourth-highest grossing film of all time [28].

In many ways, those waves originated 14 years earlier, with Shinkai’s Hoshi no Koe (Voices of a Distant Star), a short film for which he had personally created everything but the soundtrack. And he did it with what now sounds like an antique: A 400 Mhz Mac G4 with 1 Gb memory and a 300 Gb hard drive, along with Lightwave 3d 6.5, Photoshop 5, and AfterEffects 4.1[31].

In the years since, digital technology has opened doors for creators worldwide by allowing them to create a complete work single-handed. But digital creators are also at the mercy of the software and hardware upgrade cycle in ways physical creators aren’t. A sequel to Hoshi no Koe would likely need to rebuild the entire digital world, unable to reuse any of the character or scenery models or animations involved in its original creation.

I’ve interviewed three digital creators with similar long-term digital media projects ranging from six to twenty years in duration. All the artists I interviewed have the same problem: The working lifespan of their projects is significantly longer than that of their digital creative infrastructure.

Physically creative professionals working on long term projects need to maintain their own personally-managed creation-capable digital preservation systems without a clear set of best practices, because (for the most part) the long-term preservation best practices are determined by library and archive professionals, not digital media creators.

Digitally creative professionals working on long term projects need to maintain their own personally-managed creation-capable digital preservation systems without a clear set of best practices, because (for the most part) the long-term preservation best practices are determined by library and archive professionals, not digital media creators.

Digital creators need preservationists’ help in order to keep their own works alive – but preservationists’ usual goals often don’t encompass creators’ needs.

Libraries and archives emphasize read-only formats to preserve authenticity and provenance, as well as to ensure
that content is available to end users in the future. In general, creators and researchers who are still amid the creative process are not considered the target audience for digital preservation. Creators and researchers are usually directed toward storage instead, but the problem they face is both larger and more enduring than file storage.

Preservationists are accustomed to considering how to store the equivalent of a completed roll of film, not the equivalent of an entire opera house with a full prop and costume department, lights, sound, rigging, and actors. But digital creators and researchers need to maintain the entire opera house – and the smaller their team is, the longer the span of time they’ll need it available, because that small team plays every role in the production one at a time.

In this paper, I’ll assess what the professionals in the field consider(ed) to be the best solutions at various points in time, and contrast that with three longitudinal and qualitative case studies of individual media creators whose works span years or decades of development.

If the digital preservation ‘state of the art’ could be made more inclusive of individual creators’ and small working groups’ needs, we could support the ongoing use of the creation suite (the digital opera house), not just a single static production (the roll of film).

When a digital actor can’t continue to make new interactions with its digital environment, then that actor is effectively dead, and the only remnants are its completed works stored in a read only format – a phantom cast on the screen in static form, not a dynamic and living being in an active creative space with continued potential.

If the preservation field can’t provide this creative sustenance across digital generations, who else could?

2. LITERATURE REVIEW

In the scholarly literature surrounding digital preservation, preserving digital creation and editing capacity ranges from a secondary consideration that may be an acceptable loss [11, 12] to actively prevented by converting from editable formats to read-only formats [19, 32].

One school of preservation thought emphasizes mechanically comparing a migrated file is to its original [8, 9], and another emphasizes migration to new formats on demand [18], but most of the approaches to either emulation or migration assume that the original is a fixed item that is not to be modified. The Pericles project [1, 5–7, 14] is an exception, but it’s restricted to specific content from European agencies and not structured to support independent creators’ ongoing active-use needs.

While there are quite a few papers discussing the preservation of digital playback formats, not many discuss the preservation of creation formats. The history of digital preservation recommendations for other multimedia and non-multimedia formats shows that the best guesses made are not always correct.

Looking back at 2000, while assessing the long-term preservation risk of several digital formats, Lawrence et al considered Lotus 1-2-3 an important enough format to include it in their project, but no current spreadsheet system offers the ability to read its file format. However, they considered TIFF at similar risk, but versions of TIFF still appear in recommendation standards fifteen years later [19, 32]. We can’t rely on our own current format recommendations remaining viable decades into the future; we can’t know if our guesses today will be any better than Lawrence’s were in his time.

While Lawrence and his collaborators considered leaving a file unchanged less risky than migration, Mellor et al. considered migration the best possible solution to digital preservation [18]. Their recommendation course was to preserve originals in various formats and devote the bulk of their work to a type of ‘skeleton key’ master-conversion system that could migrate content to new formats on demand.

The concept seems simple, but it also makes assumptions that others haven’t supported. One is that it’s easy to keep the source format’s bitstream intact to continue to make copies; another is that migration is an unequivocally good solution to the problem. All of the digital creators I interviewed would love to migrate their works to a newer format – but the software and hardware vendors have made accurate forward migration of editable working copies impossible for them.

Becker and Rauber wrote multiple papers on languages used in verification of files that were migrated to other formats, and are much more reticent about the value of migration on demand than Mellor et al. In the XCL language paper, Becker and Rauber present the primary options as emulation or migration, and consider emulation at least as viable as migration. Migration presents serious concerns for them, even when converting static formats, as the requirements for conversion are dependent on context. In addition, they didn’t consider continued editing capability an unacceptable loss. But for working multimedia creators, I posit that the ability to edit files is an unacceptable loss. Here again, we see the difference between the perspectives of creators of repositories and creators of dynamic content.

One of the most closely related fields of study is digital game preservation, which also addresses the preservation of born-digital dynamic content with specific hardware and software requirements. There’s still a gap between preservationists’ and practitioners’ perspectives in this realm, though. In Bettivia’s paper, she notes that “significant properties” of a game are often interpreted by preservationists to mean “metadata,” but computer scientists and gamers find other properties like the game experience itself more significant. [10]

Still, the Preserving Virtual Worlds focus doesn’t quite match this paper’s focus either. Their focus is on the players of the games and the re-creation of their experiences, not on the game creators, except when re-
creation on a different platform is considered as a preservation strategy[10, 17]. This may be a viable option, and in some cases may be the only one available [10, 17, 33], but isn’t optimal. Time a digital media creator spends in re-creating an existing work on a new platform can’t be spent creating new works.

In addition to the philosophical approaches to preservation, the difference in scale between small creative groups and the larger industry is a key consideration. Most profitable film or game companies employ a large enough staff to pay for a dedicated digital preservationist, but independent media creators rarely have these resources.

Schumacher et al. surveyed a group of library professionals from smaller Illinois universities, and notes, “Practitioners at smaller institutions often do not have time to stay abreast of the frequent developments in the field of digital preservation, may not have the expertise or technical infrastructure necessary to install and maintain complex software solutions, and frequently lack the funds to pay for complete, ready-to-use solutions that may exist.” [21] (p. 4-5)

Schumacher’s statement could apply to researchers, digital media and content creators, independent video professionals, and other groups not traditionally considered full “digital repositories”, but editable-repository needs are critical to many of them. The ability to reuse files and environments saves time and effort that can be applied to creating new works.

Schumacher’s team takes nearly the opposite approach from Mellor’s. Rather than using limited resources building and maintaining a single tool, Schumacher’s team leveraged tools built by others to focus on preserving the materials unique to their institutions. In the process, they concluded that digital preservation is a continually shifting gradient scale, and a solution ‘good enough for now’ fits their needs better than a perfect solution in the future.

They investigated seventy tools, and their top recommendations for end-to-end preservation tools included Archivematica, DuraCloud, and Preservica. They concluded there was no clear one-stop solution, but provided a range of recommendations for different-sized institutions.

Their assessment of the needs of very small institutions came close to addressing the situation of independent digital media creators; still, their focus was on preserving read-only rather than editable files, they didn’t address unique hardware preservation, and the prices of all the systems but Archivematica would put them out of reach of the digital creators I interviewed.

In some cases, expansive research gathered over a longer time than a hardware or software generation faces similar issues, particularly for preserving data. Smit et al. undertook a project called PARSE.insight to assess the state of preservation among major publishers; they found that the publishers are fairly well positioned to preserve their own published papers, but that the research behind those papers is much more precarious, with 69% of publishers not offering any research preservation options [24]. Authors would like to be able to digitally preserve their research along with the publications, but there is debate over whose responsibility it is, due to the larger, more varied, and less-standardized nature of the data [24].

This gap is very similar to the digital media creators’. If the goal of digital preservation is purely defined as readibility, the content distribution formats would be sufficient – but researchers need content creation formats to be able to add to research data that has been generated over years or decades.

A digital preservation strategy for content consumers is very different than one for active content creators and maintainers. I was glad to see references to the need to talk to a system’s user community in several papers [5, 8, 10, 11]. But the ability to preserve content changes — let alone maintaining a specialized environment to continue making new creations on an old system – is often given short shrift.

The Pericles project is a notable exception, looking at digital preservation methods for dynamic digital creations for both artists and researchers [5–7, 33], but has drawbacks here and now. First, it only preserves content from European space agencies [1]. Second, it’s still under development. Third, even if it were opened to independent media creators, there’s no way to estimate participation costs. Fourth, the Pericles model is complex, multi-layered, and not reproducible by an independent creator. So, while Pericles’ theoretical approach could help independent creators over the long term, they can’t benefit from it today.

3. METHODS

To explore the digital preservation needs of media creators over a span of time, I performed a literature review including scholastic publications about digital preservation. I also conducted personal, qualitative interviews with three local digital media creators across a span of decades to see whether and, if so, how their creative processes and environments had changed.

I interviewed digital 2D animator and filmmaker Nina Paley, digital 3D animator and filmmaker Shaun Mills, and multilingual translator David Fleming. In Paley’s case, our formal interviews began in 2014. In the case of Mills and Fleming, I’ve been learning about their digital creations and working infrastructure since we first met in 1997, and I performed formal interviews with them in 2014. I followed up in 2017 to see whether and how the passage of years have changed their working environments.

I chose a small group of known interviewees to follow up with the same people across years and see whether their own assessments of their best available options held true over time.

4. FINDINGS
Interviews with the digital creators and findings have been trimmed for length. More information is available at [http://go.illinois.edu/digitaloperasupplements](http://go.illinois.edu/digitaloperasupplements).

**4.1 Nina Paley, digital and stop-motion filmmaker, *Sita Sings the Blues, This Land is Mine, and more***

Paley has the most urgent digital preservation needs of the three creative professionals I interviewed, since she has no other collaborators for her most enduring projects.

From her perspective, “computer hardware really reached its best level of price-to-functionality in about 2010, and software in about 2005” (personal conversation, 2014). Changes Adobe made when acquiring Flash in 2005 and that Apple made to Final Cut Pro in 2011 removed both functionality and backwards compatibility, leaving her and many other creators with the choice of updating at the cost of recreating their previous work, or not updating and facing increased obsolescence and difficulty in consultation.

Because Paley was a studio of one, she had originally decided to stay with her old tools. In 2014, Paley didn’t see an alternative to preserving the old hardware and software. But by 2017 she had split her digital “opera house” in half. For her own decade-long projects, she depends on aging collaborators on commercial projects, some of her newer work is done in an entirely different and newer environment, capable of 4K output but incompatible with her ten-year project and the rest of her creative works.

**4.2 Shaun Mills, integrated live action and CGI filmmaker***

Shaun Mills is an independent filmmaker with a degree in film production. He’s been working on completing his own science fiction movie, *The Gatherers*, since 2009. He’s already captured the actors’ footage and has spent the past several years creating digital effects.

Mills has been unable to find a working video content preservation system that fits his needs; he hasn’t been able to find a version control system that would capture the interactions between Lightwave and AfterEffects files, rough cut footage, and digitally manipulated PNG files needed to assemble the final frames.

After losing hundreds of hours of work because the 3D rendering and video effect files he created in AfterEffects in 2009 were no longer usable after a software update dropped backwards compatibility five years later, he’s created his own “brute force” digital preservation method: He exports every single frame of his two-hour movie to PNG, requiring 24 separate files for each second of film, then re-importing the PNGs for the next rendering pass. It takes a long time to render a scene, to the point where he’s considering reducing his paid work hours to have more time to work on his film (personal conversation, 2014). Follow-up in 2017 reveals that he’s still using his frame by frame approach; he’s kept his own backwards compatibility, but at the price of excruciatingly slow progress.

**4.3 David Fleming, professional film, game, and anime translator***

David Fleming, a professional translator working in the media translation industry since 1997, has faced the digital preservation dilemma many times. When a long-running series he’s translated 10 to 20 years earlier is be released in a new edition or format, if he can’t access his earlier digital materials because of version incompatibility, then he has to reproduce hundreds of hours of work for each season of a series.

Many of the files he works with are provided to him from international sources in compressed formats that don’t fit US media authoring standards, meaning that much of his billable work time is spent in reformatting video, in addition to translating and subtitle timing. He stores terabytes worth of working video footage on a home-built RAID server.

Changes to hardware and software have required Fleming to retain outdated versions of operating systems to continue to use his open source and custom commissioned tools; the modern corporate alternatives would block his backwards compatibility and cost tens of thousands of dollars, prohibitive for an individual freelance worker. In the mid-2000s, Fleming hired programmer Keith Hays to write a customized Excel plugin that would make it possible for Fleming to use open software packages VirtualDub and SubStation Alpha to capture subtitle timing information in an Excel spreadsheet. Hays’ plugin worked with Excel 2003, but wasn’t compatible with later versions, and Hays has been unable to create an updated version that works precisely enough for Fleming’s split-second subtitle timing needs.

**5. DISCUSSION***

I agree most with the approach taken by Schumacher’s [21] group, looking for inexpensive “good enough” solutions for the moment, while acknowledging that they may not be sufficient in the future. Lawrence *et al.* [16] follow the same principle. Mellor’s [18] monolithic solution seems unlikely to be viable for small and independent users.

The Pericles project, seems to be a larger-scale version of Schumacher’s approach – their scenarios involve assessing individual projects’ needs and customizing the solutions with the aid of digital toolkits [14, 33].

Cost and suitability continue to be a prohibitive factor for most of the available solutions. Pericles has potential, but isn’t available to independent creators yet, and under its current focus may never be.
Of the specific systems explored by Schumacher et al., three have price points of thousands of dollars per year, and the fourth requires technical knowledge and produces read-oriented packages. By comparison, online retailers offer considerable storage space for under $1000. Neither solution solves the problem of software compatibility.

On the video creation side, no good answer has been offered for Paley and Fleming’s frequent issues with vendor-planned obsolescence. Mills’ manual approach is less version-dependent, but at the cost of a considerable time investment that prevents him from making progress at sufficient speed to be financially supported by his work. A free, open-source equivalent of their tools would make long-term access safer from planned obsolescence, but they would require open source versions of many software packages. Corporations have little motivation to create open alternatives to their products, and independent media packages. Corporations have little motivation to create open alternatives to their products, and independent media creators tend to focus more on making their art than on learning to program new tools for themselves. (Fleming, by far the most technical of the creators I interviewed, still chose to hire a programmer for the Excel plugin – and an Excel plugin is vastly simpler than a video manipulation program like AfterEffects.)

Currently, either maintaining the older system or recreating their works in a newer system are the best available options for the digital creators I interviewed. Neither emulation nor forward-migration have offered them viable options.

In a future where emulation will become more accurate and less expensive, that may become a more viable option to extend the lifespan of digital environments. However, thus far hardware-specific and graphic card-specific version incompatibilities have prevented emulation from maintaining an aging digital environment for any of the creators I interviewed.

Cloud infrastructure offerings like Amazon Web Services provide tantalizing glimpses of effectively hardware-independent emulation and scalability, but as of this writing, effective use of AWS needs more technical skill and cloud-development-specific dedication than the digital creators I interviewed have available.

In theory, it sounds ideal – a major company like Amazon lends its vast pool of expertise to the business of hosting systems for thousands of large and small companies and organizations already, and the “container-based” model of thinking is designed to be more portable than many other development types.

If an individual creator could simply replicate his or her own creative environment in a platform-independent container, charged by usage, not charged when idle, and scaled to meet his or her needs, then they could preserve their digital worlds for as long as they were needed. Unfortunately, we’re not at the stage where AWS containers offer the ease of installation, use, and maintenance or the solution for hardware-specific requirements that digital creators need. In my department, even system administrators with decades of experience need help adapting to AWS in time-and-cost-effective ways, and they want to keep their systems as well patched and secured as possible. Amazon doesn’t yet offer anything like a walled garden for secure use of out of date environments, out of date software, and specific graphic card emulation – but, conceivably, they could.

6. CONCLUSIONS

Over the fifteen years since Makoto Shinkai’s Hoshi no Koe, many more independent digital creators have begun to follow in his footsteps – and a very specific irony stands out in conversations with them.

Shinkai’s first film considered the personal effects of time dilation on two human lives that are out of sync with each other, an artifact of light-years of intervening space, with increasing gaps of time between any communications from the past to the present and the future.

Fifteen years later, the greatest threat to each of these new creators’ works is the personal effects of time compression across digital environments that are out of sync with their creators’ human lives, an artifact of a digital generation gap across the technical equivalent of light-years’ hardware, software, and encoding differences that offers fewer and fewer points of creative compatibility between past, present, and future.

Right now, we simply don’t have a good option for preserving the digital opera houses – but we haven’t had nearly as much time to solve the problem as the people who work in physical preservation. Truly effective and simple-to-implement emulation could offer aging systems a new lease on life, but we’re not there yet.

Still, user-friendly cloud-based emulation is an area where digital preservation, computer science, and digital creativity could merge to offer a new world. I can imagine a future where preservationists and computer scientists have worked together to offer cloud-based emulation of a range of historic and modern systems and software, easy for digital creators to access and use, without depending on the specific Final Cut Pro version and Matrox graphics card a video creator’s used since their project’s inception. We haven’t reached that easy emulation world yet. But maybe, by Hoshi no Koe’s 20th anniversary, we could.

7. ACKNOWLEDGEMENTS

I wish to thank David Fleming, Shaun Mills, and Nina Paley for sharing their time and their experiences with me over the course of several years’ investigative conversations; Rhiannon Bettivia for her support and insight on previous drafts; and Heidi Imker, Dan Tracy, Elise Dunham, and Doug Atkinson for their encouragement and help in structuring and formatting this paper.

8. REFERENCES


[2] Anthony Willis’s answer to Video Editing: What’s a


A PDF Test-Set for Well-Formedness Validation in JHOVE - The Good, the Bad and the Ugly

Michelle Lindlar
TIB - Leibniz Information Centre of Science and Technology
Welfengarten 1B
Hannover, Germany 30167
michelle.lindlar@tib.eu

Yvonne Tunnat
ZBW - Leibniz Information Centre of Economics
Düsternbrooker Weg 120
Kiel, Germany 24105
y.tunnat@zbw.eu

Carl Wilson
OPF - Open Preservation Foundation
c/o The British Library
Boston Spa, United Kingdom LS23 7BQ
carl@openpreservation.org

ABSTRACT
Digital preservation and active software stewardship are both cyclical processes. While digital preservation strategies have to be reevaluated regularly to ensure that they still meet technological and organizational requirements, software needs to be tested with every new release to ensure that it functions correctly. JHOVE is an open source format validation tool which plays a central role in many digital preservation workflows and the PDF module is one of its most important features. Unlike tools such as Adobe PreFlight or veraPDF which check against requirements at profile level, JHOVE’s PDF-module is the only tool that can validate the syntax and structure of PDF files. Despite JHOVE’s widespread and long-standing adoption, the underlying validation rules are not formally or thoroughly tested, leading to bugs going undetected for a long time. Furthermore, there is no ground-truth data set which can be used to understand and test PDF validation at the structural level. The authors present a corpus of light-weight files designed to test the validation criteria of JHOVE’s PDF module against “well-formedness”. We conclude by measuring the code coverage of the test corpus within JHOVE PDF validation and by feeding detected inconsistencies of the PDF-module back into the open source development process.

KEYWORDS
file format validation, PDF, test data, quality assurance, JHOVE

ACM Reference format:
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION
File format validation is a central task in digital preservation processes, giving insight into the degree with which the digital object complies with the specification of the file format it purports to be. For complex formats such as PDF, which allow for a multitude of content types and variations, such as embedded AV material or embedded and non-embedded fonts, validation poses a challenging problem. While software to validate digital object’s against PDF profile requirements such as PDF/A or PDF/X exist, they typically focus on the requirements of the profile and do not take the syntactical and structural requirements of the underlying PDF format into account [8]. As of today, the go-to validator for the PDF format is the open source tool JHOVE [23]. The initial development of JHOVE dates back to 2003-2008 and the tool has been widely used by digital archives since.

Digital preservation and active software stewardship are both cyclical processes. While digital preservation strategies have to be regularly reevaluated to ensure they continue to meet technological and organizational requirements, software needs to be tested with every new release to ensure that it functions correctly. Despite JHOVE’s widespread and long-standing adoption, the underlying validation rules are not formally or thoroughly tested, leading to bugs which can go undetected for a long time. Formal testing for complex software such as file format validators has to be automated. However, a requirement for such automated testing processes is a ground-truth as a point of reference, ideally manifested in a light-weight test set. This test set can be used to check the validator’s capability to enforce specific clauses in the format specification. In the case of PDF validation in general and JHOVE specifically, no such test-set has been available until now.

This paper describes the authors’ efforts to narrow this gap by building a light-weight test-set for PDF validation. The test set focuses on the validation against structural and syntactical requirements of the PDF file format as described in the ISO 32000-1:2008 standard for PDF 1.7. It will not look at particular profile requirements such as those described in the ISO 19005 series for PDF/A. As the standard does not make a clear distinction between well-formed and valid requirements, these are derived by looking at required structural parts of any PDF object, namely: a header, a body consisting of a minimal set of objects, a cross-reference table and a trailer (see Figure 1). While JHOVE only supports PDF features up to version 1.7, the cases implemented in the test set are common to all PDF versions. The aim of test set is threefold:

(1) to establish a ground truth for what is not well-formed
(2) to test the JHOVE software against that ground truth
(3) to improve automated regression testing
Section 2 of this paper will introduce the concept of file format validation and give insight into the development of JHOVE in general and the PDF module in particular to provide a contextual framework for the test set work. Section 3 will introduce the methodology used for the construction of the test set as well as for measuring and describing the automated regression testing gap. Section 4 describes the test set itself and the results of running the JHOVE PDF-module across the test set. To introduce a second point of reference, each test file is also rendered using a suitable application. While the ability to render a file correctly does not guarantee that it is well-formed, incorrectly displayed content or the failure to render often indicates that the file is not well-formed. Section 5 discusses the impact of the test set as a ground-truth and as an improving factor in current JHOVE code as well as in existing automated regression processes. We conclude with section 6, highlighting possibilities for further work building on the test set described in this paper.

![Basic PDF structure](image)

### Figure 1: Basic PDF structure

**2 BACKGROUND AND RELATED WORK**

File format validation is a challenging task. Section 2.1 describes the motivation behind and general approach to file format validation, sections 2.2 and 2.4 illustrate how this challenge was met in the development of the JHOVE framework and the PDF-module, respectively. As aforementioned, very few efforts have been undertaken to validate against standard requirements of PDF at a structural and syntactical level, a recent exception being Caradoc [8]. Written in OCaml and first introduced in 2016, Caradoc is still in its beta stage. As opposed to JHOVE, which skips over unsupported structures thus considering them valid by default, Caradoc sets out to take a whitelisted approach, considering unknown features as suspicious by default and flagging them as invalid. While this is a thorough approach, the current implementation of Caradoc only serves as a proof-of-concept, containing rules for a very limited number of PDF features. Due to this, the vast majority of "real-life" PDFs currently fail validation with Caradoc.

Hence, the digital preservation community largely relies on JHOVE for validation - despite known bugs. The adoption of and ongoing work on JHOVE will be introduced in section 2.3, further motivating the relevance of and urgent need for thorough regression testing and ground-truth data.

#### 2.1 File Format Validation

File format validation is the process of checking an object’s conformance to syntactic and semantic rules of the format it purports to be. As such, it is closely related to file format identification. While most pattern based identification tools such as DROID or file rely on short signatures such as magic numbers, full identification requires an analysis of the entire bit-stream and a comparison to the structure and semantics prescribed by the file format’s specification [1]. To illustrate, consider the following minimal PDF code of the file minimal_test.pdf:

```plaintext
%PDF-1.4
%EOF

Minimal_test.pdf is identified as PDF 1.4 by standard file format identification tools. JHOVE, however, recognizes that the object is Not well-formed, indicating problems at the basic structural level of the file format level which the object purports to be. Ideally, the normative syntactic and semantic rules used to check the validity of an object are taken from the file format’s authoritative specification. However, in many cases a specification may not exist or may not be publicly available. Format specifications not written within an official standardization context present another problem. These are often ambiguous and therefore open to interpretation. Ambiguities in the PDF specifications published by Adobe have lead to a rather broad interpretation of the file formats syntactical and semantic makeup. This, in return, has lead to PDF rendering software being forgiving towards many violations, resulting in files which are strictly speaking invalid but are still renderable and usable.

Format validation is usually broken down into two conformance levels - determining whether an object is well-formed and valid. The W3C Extensible Markup Language Standard [31], for example, clearly defines the constraints of a well-formed XML object. In short, a well-formed XML document must contain exactly one root element, consist of one or more correctly nested and delimited elements and follow the regulations specified for entities. While well-formed XML objects comply with the XML specification, valid XML objects comply with an XML schema. In short, well-formedness addresses the syntactic correctness while validity describes the semantic correctness of an object’s conformity to the file format it purports to be. JHOVE file format modules adhere to this two-tiered conformance checking. The validation rules implemented in the TIFF module, for example, determine a file to be well-formed if the beginning of the file is an 8 byte header followed by a sequence of Image File Directories, which in return are each composed of a 2 byte entry count and a series of 8 byte tagged entries. The module defines an object as being valid if it meets certain additional

---

1 At the point of writing the latest available version is JHOVE 1.16 framework, which wraps PDF module v1.8
2 The rendering application used is Adobe Acrobat Professional 11.0.15.2
3 See https://github.com/ANSSI-FR/caradoc for beta 0.3 of the Caradoc tool
4 At the point of writing the latest available version is JHOVE 1.16 framework, which wraps PDF-module v1.8 across the test set. To introduce a second point of reference, each test file is also rendered using a suitable application. While the ability to render a file correctly does not guarantee that it is well-formed, incorrectly displayed content or the failure to render often indicates that the file is not well-formed. Section 5 discusses the impact of the test set as a ground-truth and as an improving factor in current JHOVE code as well as in existing automated regression processes. We conclude with section 6, highlighting possibilities for further work building on the test set described in this paper.

---

5 The rendering application used is Adobe Acrobat Professional 11.0.15.2
6 See https://github.com/ANSSI-FR/caradoc for beta 0.3 of the Caradoc tool
7 Tested with: DROID 6.2.1, Signature File V88, Siegfried 1.5.0 (both identified via signature pattern); TrID/32 v2.24 (findings: 100% PDF without PDF version output) and File for windows v5.03
While JHOVE works under the assumption that 1 object = 1 file = 1 while odd numbers indicate development releases (e.g. 1.15). Early where it remains open source under a GNU Lesser General Public License (LGPL). Since then, the OPF has offered software supporters and members the chance to steer maintenance and development activities through the JHOVE product board. With the move to the OPF GitHub repository, the versioning method has changed. Minor version numbers are used for production releases (e.g. 1.16), while odd numbers indicate development releases (e.g. 1.15). Early major conceptual changes in the approach to file format characterization were the introduction of a more sophisticated data model. While JHOVE works under the assumption that 1 object = 1 file = 1 format, JHOVE2 allowed for complex objects, shifting the module to 1 object = m files = n objects. A second change was made by decoupling file format identification from validation. While JHOVE conducts file format identification by iteratively calling each existing module until one reports the file to be valid, JHOVE2 relied on DROID for initial file format identification. With the move towards JHOVE2, Harvard University Library’s JHOVE developer Gary McGath left, asking for continued custody of the JHOVE code [19] which he facilitated through a SourceForge project. In the following seven years, McGath oversaw the release of 11 versions, including several updates to the PDF-module. In 2013 McGath moved JHOVE to GitHub, which had by then overtaken SourceForge as the code platform of choice [24]. In March of 2014, Gary McGath announced that he could no longer maintain the software by himself [18], leading to the Open Preservation Foundation (OPF) taking over JHOVE stewardship and moving the code to the OPF GitHub repository, where it remains open source under a GNU Lesser General Public License (LGPL). Since then, the OPF has offered software supporters and members the chance to steer maintenance and development activities through the JHOVE product board. With the move to the OPF GitHub repository, the versioning method has changed. Minor version numbers are used for production releases (e.g. 1.16), while odd numbers indicate development releases (e.g. 1.15). Early

Validation rules for the JHOVE PDF-module will be discussed further in section 2.4.

2.2 Brief History of JHOVE

JHOVE is by no means a new tool to the digital preservation community. The idea of the JSTOR/Harvard Object Validation Environment dates back to 2003 [7]. Partially funded by the Andrew W. Mellon Foundation [24], initial development of the driver and API layers as well as the ASCII, UTF-8, PDF, TIFF, GIF, JPEG and XML modules took 10 months and involved 1.35 full-time equivalents (0.10 project management, 0.25 senior analyst, 1 developer)[7]. With the initial release of version 1.0 in May of 2005, work on JHOVE continued under the auspices of the JSTOR Electronic-Archiving Initiative (now Portico) and the Harvard University Library until 2008. In late 2008, the California Digital Library, Portico and Stanford University secured Library of Congress funding under the National Digital Information Infrastructure Preservation Program (NDIIPP) for a follow-up JHOVE2 project. The project, which ran for two years, was based on the observation that the original JHOVE, even though extensively used, had “revealed a number of limitations imposed by idiosyncrasies of design and implementation” [2]. JHOVE2 was conceptualized to be a complete re-factoring of the software, allowing for simplified integration, containing streamlined APIs and including modules for file formats previously not covered in JHOVE. Two important tool benchmarks have been conducted, focusing on the identification aspect of JHOVE in juxtaposition to other tools [27] or comparing the validation output of JHOVE’s wave [30], TIFF [29, 16], and JPEG [28] modules to the output of other tools that can characterize and validate the respective file format families.

2.4 PDF Module

The PDF-hul module has been frequently updated since JHOVE’s 1.0 inception in 2008. With the exception of the 1.6, 1.9 and 1.11 framework releases, every JHOVE release has seen updates to the PDF module, resulting in new versions of the module. Improvements range from handling of parameters in accordance with the specification (version 1.3) to optimizations of the parser and the module’s memory use (version 1.10) [25]. Changes made in the PDF-module may change the outcome of validation results for a file. As such, JHOVE’s most recent 1.16 version included PDF-hul version 1.8, which fixed two major bugs in the code. These had been present from the start of development, leading to false validation errors relating to invalid page dictionary objects and improperly constructed page trees [17]. While a number of fixes have improved PDF/A validation [25], JHOVE has been proven unsuitable for PDF/A validation [10] [12]. The coverage of PDF versions hasn’t changed since PDF-hul 1.0; for “plain” PDF, JHOVE support PDF 1.0-1.6. While PDF is backwards compatible, features introduced in 10http://wiki.opf-labs.org/display/Documents/Home. 11http://wiki.opf-labs.org/display/Documents/JHOVE+issues+and+error+messages
Table 1: Lines and Percentage of code executed by unit tests per module

<table>
<thead>
<tr>
<th>Module</th>
<th>No. of Files</th>
<th>Lines of Code</th>
<th>Class Coverage</th>
<th>Module Coverage</th>
<th>No. of code lines in module (only *.java files)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIFF</td>
<td>18</td>
<td>1,253</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1,253</td>
</tr>
<tr>
<td>ASCII</td>
<td>1</td>
<td>398</td>
<td>0.00%</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>GIF</td>
<td>2</td>
<td>60</td>
<td>0.00%</td>
<td>N/A</td>
<td>60</td>
</tr>
<tr>
<td>GZIP</td>
<td>5</td>
<td>611</td>
<td>89.33%</td>
<td>90.44%</td>
<td>611</td>
</tr>
<tr>
<td>HTML</td>
<td>41</td>
<td>9,571</td>
<td>0.00%</td>
<td>0.00%</td>
<td>9,371</td>
</tr>
<tr>
<td>IFF</td>
<td>4</td>
<td>219</td>
<td>0.00%</td>
<td>38.29%</td>
<td>219</td>
</tr>
<tr>
<td>JPEG</td>
<td>9</td>
<td>895</td>
<td>0.00%</td>
<td>0.00%</td>
<td>895</td>
</tr>
<tr>
<td>JPEG2000</td>
<td>69</td>
<td>7,633</td>
<td>0.00%</td>
<td>0.00%</td>
<td>7,633</td>
</tr>
<tr>
<td>PDF</td>
<td>61</td>
<td>10,581</td>
<td>0.00%</td>
<td>0.00%</td>
<td>10,581</td>
</tr>
<tr>
<td>TIFF</td>
<td>61</td>
<td>14,457</td>
<td>0.87%</td>
<td>0.00%</td>
<td>14,457</td>
</tr>
<tr>
<td>WAVE</td>
<td>27</td>
<td>3,183</td>
<td>40.62%</td>
<td>6.87%</td>
<td>3,183</td>
</tr>
<tr>
<td>XML</td>
<td>9</td>
<td>1,498</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1,498</td>
</tr>
<tr>
<td>Totals</td>
<td>309</td>
<td>50,705</td>
<td>N/A</td>
<td>2.97%</td>
<td>N/A</td>
</tr>
</tbody>
</table>

newer versions are currently not supported by JHOVE. The complexity of the module reflects the complexity of the PDF format itself. All Adobe PDF specifications are freely available via the company’s website\(^{12}\), however, as described in section 2.1, ambiguities in the specifications have lead to differing interpretations of the file format’s syntactical and semantic restrictions. This complexity resulted in the PDF module consuming significant resources to complete\(^{13}\) [7], and is also reflected in the continuing work on the module as described above.

But what does the module base the validation outcome on? Section 2.1 described the general two-tiered conformance approach taken by JHOVE. While the differentiations between well-formed and valid is rather straightforward for well specified file formats such as TIF\(F\), the situation is unfortunately more complex for the PDF file format. The module description itself states that a PDF is considered well-formed if it meets the criteria defined in Chapter 3 of the PDF Reference, breaking this down further into the following requirements \(^{15}\): “In general, a file is well-formed if it has a header:%PDF-m.n; a body consisting of well-formed objects; a cross-reference table; and a trailer defining the cross-reference table size, and an indirect reference to the document catalog dictionary, and ending with: %EOF”. Unfortunately this statement remains vague, concrete rules breaking these high-level requirements down to e.g. dictionary or object level are missing. The PDF-module documentation carries on stating that in addition to further requirements, well-formedness is a prerequisite for validity


\(^{13}\) It may surprise that the original JHOVE developers stated that the complexity of the PDF-module was superseded by the HTML-module, elaborating that while other modules were designed to terminate the validation process at the first error, the HTML module typically encountered so many errors that it had to be designed to recover from errors and continue [7].

\(^{15}\) Adobe: “The rules described here are sufficient to produce a well-formed PDF file” became ISO: “The rules described here are sufficient to produce a basic conforming PDF file”

3 METHOD

In this section we briefly introduce software testing methods. In particular, we are highlighting the use of test corpora to measure code coverage as a form of software testing. We conclude this section with a brief description of the process used to build the light-weight test set put forth in this paper.

3.1 Software testing

JHOVE is a large established code base, currently comprising over 500 Java files and 100,000 lines of code. There are modules that validate files against twelve format specifications, each requiring its own specialist knowledge. Good automated testing is the only way to ensure that a large software project functions as expected, but how do you ascertain how well a piece of software has been tested? Before addressing JHOVE testing we define a few key software testing terms: The term code coverage is used to express how much of a code base has been tested. It can be measured empirically for a given test scenario by measuring how many lines of code are executed when you run it. You then divide this by the number of lines of code in the project to give a percentage figure. This task is carried out by automated coverage tools used by software testers and developers.

Testing can be carried out in different ways, one of the most effective forms is known as unit testing. These are small, discrete tests written by programmers when developing and fixing code. They are usually executed automatically every time that the code is [13]. Regarding limitations of the validation by the module, the PDF-module documentation only states that data within content streams as well as encrypted data is not validated. While these criteria may seem straight-forward, in reality they include thousands of possibilities. Furthermore, while Adobe’s specification included the “well-formedness” terminology [4], the ISO standard replaced this with “conformance”\(^{14}\) [11]. The ISO standard clearly states, that it does not specify "methods for validating the conformance of PDF files", carrying on, however, by describing that “conforming PDF files shall adhere to all requirements of the ISO 32000-1 specification and a conforming file is not obligated to use any feature other than those explicitly required by ISO 32000-1.” [11]. With 5,471 occurrences of the word “shall” as the ISO verbal form for a requirement and no clear differentiation between well-formed and valid, achieving rule-based conformance checking remains a lofty goal. Nevertheless, is the most suitable point of reference for what is syntactically and structurally valid and what is not. Due to the complexity of both the file format and module, it should come as no surprise that validation errors like those recently fixed in JHOVE 1.16 went unnoticed for many years. However, when relying on a validation tool for digital preservation decision-making, the tool’s output must be correct and complete. Ideally, a test routine exists, which checks the tool’s output against a ground-truth for every new release. While such a ground-truth needs to exist on both, well-formedness and validation levels, the authors have limited the scope to criteria determining well-formedness of the object.
compiled. These test the smallest components, usually a few lines of code, that make up a software package.

**Integration testing** is an alternative, complementary approach to unit testing. Rather than test low-level units, integration testing focuses on larger software components, including the delivered software. These tests can often take a long time to run so aren’t run for every code change. Integration tests are usually run as a final test before delivering release candidate software.

**Black box testing** describes an approach to testing that concerns itself solely with software functionality and ignores the internal details. Nearly all software testing above the unit testing level is carried out as black box testing.

### 3.2 Testing JHOVE

Historically, JHOVE hasn’t had a rigorous, public testing policy. There are very few unit tests, the unit testing code coverage figures for the JHOVE modules as of v1.16 are shown in Table 15.

For each module two coverage figures are given, each measures the percentage of code that’s executed by unit tests, the first for the module control class the other for the supporting classes. The test coverage was measured using the Jacoco software and its accompanying Maven reporting plugins. Overall, there’s less than 4% coverage of the codebase. Writing unit tests for 100,000 lines of code is at least 2 person years of effort and isn’t practical given the development resources available to the project. Clearly another approach is needed to establish confidence in JHOVE’s results.

### 3.3 Using Test Corpora and Measuring Coverage

One method suited to testing JHOVE is to create a corpus of ground truth test data designed to test the modules functionality. A good test should be atomic if possible, examining the codes handling of a specific validation issue. One large, valid PDF document might execute a high percentage of the code base without providing any insight into the manner in which JHOVE deals with validation issues at all. Currently two data sets for testing exist: The first dataset is comprised of example files that accompany the JHOVE code base. These have presumably been used by developers to test that the modules were working, without providing a formal, rigorous test corpus. The regression testing suite that the OPF is currently developing uses this data and compares the XML output of different versions of JHOVE to ensure that the results haven’t changed. This is typical black box testing, treating the software as a unit that takes input files and produces XML output without worrying how the software does this. This test set comprises approximately 80 files that cover the PDF, TIFF, HTML, GIF, JPEG, JPEG 2000, XML and ASCII modules. The other test set is currently being assembled as a community effort. The aim is to gather a set of test files that between them generate every one of JHOVE’s 153 PDF validation errors. First efforts, undertaken as part of the OPF DIG group and the JHOVE hack day have put forth 44 files. However, as these files come from institutions’ productive archival workflows, there are some associated problems: About a third of them are subject to access regulations and may not be shared publicly. Secondly, as these are “real-world” examples, the PDF files are typically large and complex, making it hard to understand which specific part of the digital object triggered a validation rule. A ground truth test set is currently lacking.

### 3.4 Building a test corpus

In order to conduct straightforward black-box functional tests we introduce a set of PDFs built specifically to test the requirements of well-formedness. We limit ourselves to well-formedness, as it forms the prerequisite for a valid file. The JHOVE well-formed statement is split into the requirements for the structural sections, shown in Figure 1: header (section 4.1), body (section 4.2), cross-reference table (section 4.3) and trailer (section 4.4). For each section, the JHOVE requirement - if available - forms the starting point of the process shown in Figure 2. In a second step, the ISO 32000-1:2008 standard is checked to transform the high-level criterion into individual requirements / test cases. These requirements are then implemented as a test file, which is validated using JHOVE and rendered with Adobe Acrobat Professional. To produce atomic tests against syntactical file format requirements, a minimal well-formed and valid PDF (“Hello_World.pdf”) was created and used as the basis for all other test files. This file consists of a single page which includes one font definition as a resource and a single text stream as a content. The graph including respective object Ids is shown in Figure 3. Information on test cases and accompanying test results are captured in a google spreadsheet.

### 4 TEST CORPUS

This synthetic test corpus consists of 90 files, which are available on the JHOVE github repository [25]. The test set contains the minimal PDF file (minimal_test.pdf) given as an example in section 2.1, the starter file “Hello_World.pdf” as well as 88 test cases which are derivations of “Hello_World.pdf”. Test cases were created using the
process described in section 3.2. To allow files within the test corpus
data set to be referenced, we introduce a naming scheme for the test
files and cases. The scheme is based on a scalable ontology, which
follows the basic PDF structure as shown in Figure 1. The main
four sections are numbered with the body section (T02) branching
off in separate subsections for possible object types. For this paper,
the object types document catalog (T02-01), page tree node (T02-
02), page node (T02-03), page resource (T02-04) and stream object
(T02-05) are analyzed further. While different types of objects are
possible for page resources and stream objects, we only focus on
font (T02-04-01) and text respectively. Each test case is numbered
according to the section in which the deviation is introduced in
(Txx-xx-xx), followed by a 3 digit number ((xxx). In addition to the
test case ID, the file names contain a brief description of the feature
tested, e.g. “T001_header_invalid-major-version.pdf”. All of the test
files are listed in a google spreadsheet19. The created test corpus
has been made available on GitHub19.

4.1 PDF Header

The PDF-hul documentation prescribes that the PDF Header con-
ists of the first line of the file which must contain the five characters
"%PDF-", followed by a version number. According to ISO32000-
1:2008, the version number is of the form 1.N, where N is a digit
between 0 and 7 [11]. Adobe’s documentation states simply that
the syntax is %PDF-M.m., where M is the major and m the minor
version [3]. In addition to the general syntax, each Adobe specifica-
tion explicitly names the header relevant for the respective format
version, e.g. %PDF-1.6 for PDF 1.6 [4]. Beginning with PDF 1.4 the
version may also be included in the document’s catalog dictionary.
This information is relevant for adequate identification of the file
format version, it does not replace the structural requirement of
the header.

Test cases are based on the syntax requirement as well as by using
plausible version numbers. Deviations from the syntax are added to
the first 5 chars %PDF- as well as to the version notation M.m.
An additional test case is formed by removing the header. In total 7

4.2 PDF Body Well-formed objects

The JHOVE PDF-hul module requires that "the file has a body,
consisting of well-formed objects". PDF supports five basic and
three compound objects: Integers / real numbers; strings which
must be enclosed in parentheses; names, which are introduced by a
forward slash; boolean values; the null object, denoted by keyword
null; arrays consisting of an ordered listing of objects, e.g., [1 0 0 0];
dictionaries and streams [33]. Objects are linked to each other via
indirect references. Additionally, PDFs are divided into nodes as
objects (obj). The content of these nodes, such as a page tree node,
are regulated in the standard, prescribing required and optional
objects within. This paper focuses on the five PDF node structures
required to construct a minimal PDF as shown in Figure 3: the
document catalog (4.2.1), the page tree node (4.2.2), the page node
(4.2.3), the page resource node containing the resource font (4.2.4)
and the stream node containing text (4.2.5).

4.2.1 Document catalog. The JHOVE well-formedness descrip-
tion only mentions the document catalog in conjunction with the
trailer (see 4.4), which must contain an indirect reference to the
document catalog. However, the document catalog is the root of
the PDF’s object hierarchy graph (see Figure 3), containing refer-
cences to all other objects which define the document’s content,
outline, threads and attributes. The document catalog can also con-
tain information about how the document shall be displayed, e.g.,
defining a default page other than page 1 to first show when the file
is rendered. While a number of key pairs are possible within
the catalog dictionary, only two are required: Type with value Catalog,
which defines the object type of the dictionary, and Pages, which
contains an indirect reference to the page tree node [11].

Test cases present a missing or inaccessible document catalog
(T02-01-001 - 002), a missing or incorrectly defined Pages indirect
object (T02-01-003 - 004) and a missing Type key value pair (T02-
01-005 - 007)./newline Analyzing the results of running JHOVE
across the test set revealed some troubling behavior. While JHOVE
correctly recognizes when the document catalog is removed com-
pletely, it does not appear to cross-check the indirect reference in
the trailer and the actual object number of the document catalog.
While the file (T02-01-002) is not well-formed according to the stan-
dard and cannot be rendered by Adobe Acrobat, JHOVE reports the
file as being well-formed and valid. The suspicion that referenced
object numbers are not cross-checked against the target objects is

ipres-paper-pdfs/modules/PDF-hul

Figure 3: Graph and according PDF object Ids for Hello_World.pdf file
further confirmed by the test case containing an incorrect indirect reference in the Pages value (T02-01_004). Here, the file is also flagged as well-formed and valid by JHOVE but cannot be rendered by Adobe Acrobat, as it fails to locate the page tree node needed to display the pages. Similarly, the requirement that the value for Type must be Catalog is not checked. Invalid entries are considered Well-formed and valid by JHOVE. However, as opposed to the other false negatives, which could not be rendered by the reader software, the test file with an invalid Type key value (T02-01_006) renders correctly and without a warning.

While the validation against the values result in incorrect outcomes, the routines implemented in JHOVE do correctly check the existence of the required key values Pages and Type - the absence of which results in objects being reported as Not well-formed with "Invalid object definition" error messages (T02-01_003, 005, 007).

4.2.2 Page tree node. While the JHOVE well-formedness requirements do not mention the page tree object, the object is required in order to interpret and access the file’s content correctly. Every PDF object has at least one page tree node, which forms the basis of the hierarchical structure of page leaf nodes and/or further page tree nodes. Required information within the page tree node dictionary is the dictionary Name with required value Pages, the number of children the node has (further page tree nodes or pages) given in the Count key value pair, and lastly an array of indirect objects of the children in the Kids key value pair. If the page tree node is not the root page tree node, i.e., not the first page tree in the document, a Parent key value entry is also required. As our lightweight test set only contains a single page tree node, this value is not required and is not checked within the test set.

The test cases present a missing page tree node (T02-02_001), missing or malformed Type (T02-02_006, T02-02_009) and Count (T02-02_007 - 008) key value pairs and a missing or malformed Kids array (T02-02_003 - 005).

The result of running JHOVE across this test set is similar to that of the previous section - a missing page tree node is detected by JHOVE, resulting in a Not well-formed object and an error message stating "Invalid object definition". The same result - Not well-formed and "Invalid object definition" - is obtained when validating the test cases with missing required keys Type (T02-02_006) and Count (T02-02_008). In both cases the file renders correctly in Adobe reader, however, Adobe attempts to fix the error and prompts the user to save the changes upon closing the file. As in the case of the document catalog key value pairs, the values for Type and Count do not appear to be checked by JHOVE - invalid values such as an incorrect integer value for Count result in well-formed and valid files according to JHOVE despite the fact that they clearly violate the requirements defined in the standard.

In the case of the array Kids, values appear to be at least partially checked, invalid indirect objects such as a reference to self (T02-02_002) and no kids (T02-02_005) or a combination of existing and non-existent children (T02-02_003) are caught correctly, resulting in a not-well formed outcome and error messages such as "Excessive depth or infinite recursion in page tree structure" for self-reference or "Invalid object definition" for missing kids. The test case containing only one, non-existent object as the only array entry for the Kids key (T02-02_004) was interesting, as it resulted in a well-formed, but not valid outcome. It is unclear, why this is considered well-formed as opposed to, e.g., the combination of valid and invalid array entries found in test case T02-02_003 which are flagged as not well-formed by JHOVE.

4.2.3 Page node. No specific requirements concerning page objects are given in the JHOVE PDF-module description, however, per standard a PDF file must have at least one page object. The page object dictionary can contain 30 different key pair values, but only 4 of them are required: Type, Parent, MediaBox and Resources. MediaBox and Resources can be inherited from ancestor nodes in the page tree. Other keys are only required under specific conditions, e.g., StructParents is required, if the PDF contains structural content items. It’s interesting to note that Contents is not a required key value pair. If no Contents object is referenced, the page is simply blank. As our sample file contains one Contents object, we are testing the Contents reference despite the fact that it is only an optional key.

The test cases present a missing page object (T02-03_006), and missing or malformed Type (T02-03_001, T02-03_002), Parent (T02-03_003 - 005), Resources (T02-03_008, T02-03_012), MediaBox (T02-03_008 - 009) and Contents (T02-03_010 - 011) key value pairs. As in the analysis for the other body test cases, a missing node (T02-03_006) as well as missing key value pairs (T02-03_001, T02-03_003, T02-03_008, T02-03_012) result in Not well-formed files with "Invalid Object definition" errors. This includes the test cases where the Contents key value pair is missing (T02-03_010) - an interesting result, considering that the Contents key value pair is optional according to the standard. An absence of Contents should therefore result in an invalid, but not directly in a not-well formed status. Moreover, an invalid value entry for Contents results in a well-formed, but not valid output (T02-03_011), which is surprising as in previous dictionary cases plausibility and correctness of values was rarely checked and never lead to well-formed, but not valid outputs. The respective test file cannot be rendered by Adobe Acrobat and leads to the rendering application crashing - clearly not the result we expected for a well-formed but not valid object. Indirect values for Parent (T02-03_004) and Resources (T02-03_007) go unnoticed, resulting in Well-formed and valid outputs. Checking against inconsistencies for the Type values (T02-03_002) lead to an interesting discovery. After having changed the Type value from the required Pages to Catalog, JHOVE returned the file as Not well-formed. This was surprising, as the previous test cases had returned invalid entries for the Type key value pair as well-formed and valid. Further analysis revealed that when the Type was then changed to Font and re-validated the outcome was Well-formed and valid, even though the value was still wrong. This leads to the assumption that some but not all values for Type are checked and the value Catalog always leads to further analysis by the software. Further test cases which were handled correctly by JHOVE's validation routine and were identified as Not well-formed are a wrong object type for the Parent value, which expects a single indirect reference (T02-03_005) and the wrong number of parameters for the MediaBox.

4.2.4 Page Resource - Font. Well-formedness criteria for page resources in general and fonts in particular are not addressed by the description of JHOVE well-formedness criteria. As per standard,
resources for a page, such as images or font, may be described in resource dictionaries which can be included in dedicated page resource objects. Alternatively, resources may also be directly described within content stream objects. This paper only briefly examines page resources for fonts. The use of fonts in PDF is a particularly complex subject [5], hence this is only a high level analysis focusing on font dictionaries and associated data structures. This includes a look at the information a conforming reader requires to interpret the text and position the glyphs correctly. Required key value pairs in the font dictionary are Type, BaseFont and Subtype. The test cases present a missing or invalid resource object (T02-04-01_005, T02-04-01_004), a missing or malformed Subtype (T02-04-01_005, T02-04-01_006) and missing BaseFont (T02-04-01_002, T02-04-01_006) key value pairs. As is the case with other tests, a missing dictionary Type key value pair (T02-04-01_003) resulted in a not well-formed JHOVE result, while an invalid value - in this case Page instead of the expected Font (T02-04-01_004) - produces a well-formed and valid result. The Subtype key identifies the font type, ISO32000 lists the following valid values: Type0, Type1, MMType, Type3, TrueType, CIDFontType0 and CIDFontType2. A missing SubType key value pair (T02-04-01_005) from the font dictionary results in an unreadable file, which JHOVE correctly catches as Not well-formed with an "invalid object definition error". A more distinct error message indicating the magnitude of the error would be helpful here. A wrong value for SubType (T02-04-01_006) unfortunately goes unnoticed - resulting in a well-formed and valid JHOVE result. The BaseFont key contains the actual font - for Type1 fonts this is typically the name it is known by in the respective font program. A missing BaseFont (T02-04-01_001) results in Adobe being unable to render any text which uses the font. JHOVE catches this error as a well-formedness violation. Again, the error message is "invalid object definition" as well as an additional "no document catalog dictionary" error. This appears to be a resulting downstream error in parsing dictionaries and is misleading here. A BaseFont with the wrong value (T02-04-01_002) also results in a "no document catalog dictionary" error, which is again misleading. Adobe Reader renders the text using an alternative, default font.

4.2.5 Stream Objects – Text. JHOVE states no specific requirements for stream objects. Streams are used to store binary data or text to be displayed on a page. For this paper, we only examine a simple, uncompressed text stream. As per ISO standard, the requirement for printing text on page is a Stream dictionary including the Length key value with the number of bytes between the stream and the endstream keywords. The stream dictionary must be followed by a descriptor stating the position of the text on the page, a beginning text object marker (BT), operators to choose the text font (Tj), size, the text itself, the font show operator (Tj), the end text marker (ET) and the endstream keyword. The standard mandates that no extra bytes other than white space are allowed between the endstream and the endobj markers. The test cases check the presence of the required text operators (T02-05-01_001 - 008, T02-05-01_012), correct syntax of the text object (T02-05-01_009 - 011), missing keywords stream (T02-05-01_13) and endstream (T02-05-01_14), missing or invalid Length key value pair (T02-05-01_015, T02-05-01_016) and extra bytes between endstream and endobj keywords (T02-05-01_017). Running JHOVE across the test files showed that all text operators are checked by the PDF-module. The absence of the operators is detected, resulting in a Not well-formed output (T02-05-01_001 - 008, T02-05-01_012). As these errors typically result in the rendering application being unable to open the file, it is particularly important that JHOVE detects them. However, the accompanying error message for the test cases: "No document catalog dictionary", appears to be a down-stream error and does not indicate that the problem is in the stream object or more specifically in a missing text operator, which would be important information for the user. The PDF-module correctly detects missing stream (T02-05-01_13) and endstream (T02-05-01_14) keywords and missing or invalid Length key value pairs (T02-05-01_015, T02-05-01_016). Validation issues were detected when processing the actual text streams. String objects can be either encoded as literal strings enclosed in parentheses, or as hexadecimal streams enclosed in angle brackets [11]. Our lightweight test set includes a literal string. However, missing opening or closing parentheses (T02-05-01_009, T02-05-01_010) as well as a substitution with brackets (T02-05-01_011) goes unnoticed by JHOVE, returning a Well-formed and valid result. This is especially grave as the reader fails to render the files, showing the message "An error exists on this page. Acrobat may not display the page correctly. Please contact the person who created the PDF document to correct the problem".

4.3 Cross reference table

The cross reference table enables random access to the various objects contained in a PDF and is an essential element of any PDF file. JHOVE acknowledges this, the mandatory presence of the cross reference table is mentioned in the well-formedness conformity statement. However, as with other objects, JHOVE gives no further requirements for the table. While conforming PDF implementations may divide information between multiple cross-references streams, cross-reference sections and cross-reference tables, this paper only examines a lightweight PDF with a single cross-reference table. Per standard, the cross-reference section must start with the xref keyword. If the file only contains one table and has never been updated, as is the case with our test file, the second line should start with 0 and include a second number stating the number of entries in the table - in our case 6. Finally, the table contains one entry for each object. Each entry is exactly 20 bytes in length and consists of the off-set (10-digit), the generation number (5-digit) and a keyword indicating whether the object is in use (n) or free (f). [11] The test cases present a missing cross-reference table (T03_001), missing xref keyword (T03_002), missing or invalid number of entries (T03_003, T03_004), missing or invalid offsets (T03_006 - 008), invalid entry keywords (T03_009) and invalid generation numbers (T03_010). The test files for missing cross reference table and xref keyword, (T03_001, 002) as well as for an invalid number of entries (T03_003 - 005) were detected correctly as Not well-formed. An interesting outcome in this context was a software bug when performing the test against missing number of entries (T03_003), producing the following error: "edu.harvard.hul.ois.jhove.module.pdf.Keyword cannot be cast to edu.harvard.hul.ois.jhove.module.pdf.Numeric". This is a java exception thrown when the application has tried to
perform an undefined data conversion. Only one test case for the cross-reference table led to an incorrect validation result. (T03_010) - here, an invalid generation number for an entry was proclaimed as \textit{Well-formed and valid} by JHOVE.

4.4 PDF Trailer

The PDF-module description states that to be well-formed, a file must have "a trailer defining the cross-reference table size with an indirect reference to the document catalog dictionary, and ending with: \texttt{%%EOF}" \cite{11}. The PDF standard is more specific in its requirements, stating that the trailer must start with the trailer dictionary, consisting of the \textit{trailer} keyword and key-value pairs enclosed in double angle brackets. Two key value pairs are required for all PDFs: \texttt{Size} of type integer, which holds the total number of entries in the cross-reference table, and \texttt{Root} of type dictionary, which contains the indirect reference to the catalog dictionary or root object of the PDF. Other key value pairs are reserved for particular PDF properties, such as \texttt{Encrypt} and \texttt{ID} for encrypted documents, as well as \texttt{Prev} for objects with more than one cross-ref section.

The trailer dictionary is followed by the \texttt{startxref} keyword and the byte offset counting from the beginning of the file to the last cross-reference section. The trailer and the object closes with the end-of-file marker \texttt{%%EOF} \cite{11}. For the test cases, The JHOVE well-formed criteria stated above are broken down into the existence of a properly formed trailer (T04_008 - 010), the existence and validity of the mandatory key value pairs \texttt{Size} (T04_015, T04_016) and \texttt{Root} (T04_011 - 014). The last line of a PDF file contains only the end-of-file-marker \texttt{%%EOF}. As PDF files are typically read starting with the trailer, the EOF-marker is an essential keyword, indeed most applications will not parse or render the file if it is missing \cite{3}. The test set contains a number of invalid variations of the \texttt{%%EOF} tag (T04_001 - 007).

While not explicitly mentioned in JHOVE's conformance requirements, the offset of the cross-reference section prefaced by the \texttt{startxref} keyword is an essential part of the trailer. Because of this, offset and keyword are also included in the test set (T04_018 - 019). With one exception, every test case pertaining to the trailer dictionary or the cross-reference table byte offset produced the expected status message \textit{Not well-formed}. The exception was an unexpected program termination in the test case which omitted the closing brackets. It seems as if the module gets stuck in an endless loop. Another unclear case were two error messages which appear to be incomplete ("4" and "Null"). One arose in the case of an incorrect value of type integer given in \texttt{Size}, i.e., not the correct size of the cross-reference table entries. Here, one of the errors simply returns the value "4". Further experiments have shown if the number of entries does not equal the value of \texttt{Size}, the error message returns the value stated in the \texttt{Size key} value pair. This should be replaced with a more detailed message. The "Null" error appears to have similar origins, as it was produced by the test case with a missing \texttt{Size} entry. Two of the test cases produced validation errors rather than well-formedness violations: "Size entry missing in trailer dictionary" and "Root entry missing in trailer dictionary". While both test cases result in a \textit{Not well-formed} status due to a subsequent error ("No document catalog dictionary"), the missing entries should be reason enough to fail the syntactical check, as the standard clearly states that \texttt{Size} and \texttt{Root} keys are mandatory elements of the trailer dictionary. Further dictionary errors thrown, such as "Malformed indirect object reference" or "Improperly nested dictionary delimiters" are generic to all dictionaries and also appear in other objects.

The main problem found when validating the \texttt{%%EOF} tag test cases was that JHOVE is tolerant towards data being present after the \texttt{%%EOF} tag. Following incremental updates a PDF file might contain several \texttt{%%EOF} tags, still the last line must be \texttt{%%EOF}. The test case containing junk data after the tag (T04_005) passed JHOVE validation as \textit{well-formed and valid}. Furthermore, the ISO standard states that the \texttt{%%EOF} tag should be present in the last line of the file. JHOVE validation simply follows the requirement that \texttt{%%EOF} is the last string in the file, regardless of it being on a line of its own (T04_001), resulting in a \textit{well-formed and valid} file.

5 DISCUSSION

As part of the work presented in this paper we have developed a light-weight test set for JHOVE’s PDF validation routine against well-formedness requirements derived from the ISO 32000-1:2008 standard for PDF. As presented in section 1, our aim has been to:

1. establish a ground truth for well-formedness criteria
2. test the JHOVE software against that ground truth; and
3. improve automated regression testing

Within this section we will briefly discuss if and how the test set meets these goals.

5.1 Establishing a ground truth

The JHOVE PDF-module’s description of the requirements which need to be fulfilled to be considered well-formed are fairly vague. Particularly the definition of what comprises a well-formed body is high level. Neither a ground truth test corpus nor a clear overview of the validation criteria enforced by the JHOVE PDF-module are available.

In this paper, we have checked the criteria presented in the PDF-module documentation against the concrete requirements stated in the PDF ISO 32000 standard. Using a light-weight test object, consisting of a minimal set of structural objects, we have produced ground truth data for a small sub-set of criteria for the validation against structural and syntactical requirements. As the ISO standard does not include a differentiation between well-formed and valid, we have defined well-formed as the required syntactical and structural aspects of the PDF graph in general and the object’s used within in particular. This approach shows that the line between well-formed and valid for PDF is unfortunately by no means as straightforward as the XML example included in section 2.1. This will be especially challenging when the work described here is taken forward, looking at the optional structural elements of PDF, such as object requirements for linearized PDFs, which have fixed requirements in themselves. Is a linearized PDF only well-formed when all requirements for linearization are met? Or is any requirement violation occurring in an optional structural part only a violation of validity?

5.2 Testing JHOVE against ground truth

The test cases included in this paper have shown clearly how error-tolerant rendering software can be. Files syntactically wrong at
very elementary levels may happily be rendered by tools such as Adobe Acrobat without a warning. However, we’ve also produced examples where the reader failed to render objects which JHOVE deemed to be well-formed and valid, further underlining the invaluable asset of a ground truth test set. By running the test cases against the JHOVE PDF-module, we have discovered a number of inconsistencies between expected outcome for a test-case and the de-facto validation result returned by JHOVE 1.16 / PDF-hul 1.8. These inconsistencies are being raised as issues on the GitHub project site where they will be picked up by the JHOVE maintainers. As a first step, we have opened issues on the JHOVE GitHub repository for 9 of the discovered inconsistencies:

- issue #207: PDF version checking incorrect
- issue #208: Inconsistent catalog indirect reference and object number
- issue #209: Inconsistent Pages indirect reference and object number
- issue #210: Value for Type in Document Catalog not validated
- issue #211: Indirect reference to not existing object in page tree node Kids array results in well-formed and valid
- issue #212: Value for Type in Page Tree Node not validated
- issue #213: Consistency between /Parent and Kids for Page Tree Node and Page Object not checked
- issue #214: Parenthesis around literal strings are not checked
- issue #215: Error message - JHOVE expects integer, gets string in cross-reference stream with missing value

### 5.3 Improving test coverage and regression testing

A corpus based testing methodology has been described in section 3.2. In this section we examine the utility of the test set produced with this paper as a test corpus for JHOVE PDF validation. This was tested by measuring and analyzing the coverage for the individual files and the test set as a whole. The "hello_world.pdf" seed file covered 36% of the code in the PDF module’s control class and 26% of the module code. Examining the coverage figures reveals that when the problems presented in the test files were undetected by JHOVE the coverage figures for the test are identical to the seed file. That makes sense as the application found no problems in the file so continued processing the file resulting in more of the code being executed. Because of this we’ve omitted these results from our summary as they obscure the coverage figures for the discrete tests that fail as expected. Table 2 shows the coverage figures produced by running JHOVE over the test classes described in this document.

Regarding these test figures, it’s worth noting the manner that the combined coverage figures are never higher than around 36% for the controlling class and 26% for the module. Furthermore these figures are identical to the coverage generated by the seed file. This means that many of the tests are exercising the same areas of the code base reflecting similarities in the test cases themselves. It’s important not to get too carried away with test coverage in this respect. The real use of these test file is not the amount of code they execute but in which parts of the code they execute. Synthetic test corpora can be crafted to exercise specific sections of the code base. Initially this can be done by studying JHOVE’s validation criteria and the PDF standards as is the case for the work presented here.

<table>
<thead>
<tr>
<th>Test File</th>
<th>Class Coverage</th>
<th>Module Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>T01 Header tests</td>
<td>9%</td>
<td>17%</td>
</tr>
<tr>
<td>T02-01 Document Catalog tests</td>
<td>21%</td>
<td>18%</td>
</tr>
<tr>
<td>T02-02 Page Tree tests</td>
<td>21%</td>
<td>20%</td>
</tr>
<tr>
<td>T02-03 Page Object tests</td>
<td>26%</td>
<td>22%</td>
</tr>
<tr>
<td>T02-04 Resources tests</td>
<td>19%</td>
<td>18%</td>
</tr>
<tr>
<td>T02-05 Stream tests</td>
<td>19%</td>
<td>18%</td>
</tr>
<tr>
<td>T03 Cross Reference tests</td>
<td>28%</td>
<td>24%</td>
</tr>
<tr>
<td>T04 Trailer tests</td>
<td>19%</td>
<td>20%</td>
</tr>
<tr>
<td>All Test Files</td>
<td>36%</td>
<td>26%</td>
</tr>
</tbody>
</table>

As test coverage increases it’s possible to see which areas of the code remain untested and use this as another guide when creating test files.

Even after JHOVE validation has been tested and verified to function as expected the test corpus has a vital role in ensuring that this remains the case. The data set can be used to regression test new releases of JHOVE to make sure that fixed bugs are not inadvertently reintroduced as new development takes place. Finally we’ll examine coverage figures for the two other test data sets introduced in section 3.3 - the initial JHOVE example files and the OFF JHOVE error corpus - and compare the figures to this corpus for context.

Due to the complexity of the files within the other corpora, they produce greater coverage of the PDF module’s code. However, the high coverage comes as a price: due to their complexity, they are also of less use in identifying and fixing problems. For complex, "real-world" files it is much harder to predict which areas of the code they will test, in turn making it harder to diagnose problems as there’s more code to analyze. Furthermore, many bugs can be caused by interactions between features that weren’t properly considered when developing the software. In summary synthetic test files are ideal for testing how the software deals with the individual elements of the specification. Real world files are excellent candidates for testing software’s function when presented with more complex files as well as measuring performance but only after confidence in the software has been established through formal testing using synthetic test data.
6 SUMMARY

In the work presented in this paper, the authors have created a light-weight test set for the validation of PDFs at the structural well-formedness level. The test set consists of 90 files and 88 test cases and is publicly available via the JHOVE GitHub repository26. Additionally, the outcome of the validation and rendering tests described in this paper as well as the detailed figures for the code coverage of the test set in regards to the JHOVE PDF-module are described in a spreadsheet available online27.

The authors have shown how the test corpus can be used to serve three purposes: to establish a ground truth for what is not well-formed, to test the JHOVE software against that ground truth and lastly to improve automated regression testing. Inconsistencies discovered in running the ground truth data against JHOVE are being fed back into the development process via GitHub issues. Furthermore, the test set and process will be shared with the JHOVE community, hoping to stimulate discussion around the methodology used and triggering further efforts in extending the test data to cover more features of PDF files and other format modules.

If we want the software we use to be fit for the lofty, long term goals that the digital preservation community aspires to it needs to be tested thoroughly. This testing needs to be public and demonstratively complete, and who better to make sure that this is the case than the community who use the software. Only when this testing is in place will the JHOVE status of Well-formed and valid carry the lifetime guarantee we want it to.

REFERENCES


PDF/A considered harmful for digital preservation

Marco Klindt
Zuse Institute Berlin (ZIB)
Takustr. 7
Berlin, Germany 14195
klindt@zib.de

ABSTRACT
Today, the Portable Document Format (PDF) is the prevalent file format for the exchange of fixed content electronic documents for publication, research, and dissemination work in the academic and cultural heritage domains. Therefore it is not surprising that PDF/A is perceived to be an archival format suitable for digital archiving workflows.

This paper gives a rather short overview about the history and technical complexity of the format, its benefits, shortcomings and potential pitfalls in the area of digital preservation with respect to aspects of accessibility and reusability of the information content of PDF/A.

Several potential problems within the creation, preservation, and dissemination contexts are identified that may create problems for present and future content users. It also discusses some of the risks inherent to PDF/A for parts of the preservation community and suggests possible strategies to mitigate problems that might prevent future human or machine-based usability of the data and information stored within digital archives.

CCS CONCEPTS
• Information systems → Digital libraries and archives; Document structure; Content analysis and feature selection; Information extraction; Data encoding and canonicalization; Structured text search; • Human-centered computing → Accessibility systems and tools; Accessibility technologies; • Applied computing → Digital libraries and archives; • General and reference → Validation;

KEYWORDS
Portable Document Format, PDF/A, file format analysis, risk assessment, accessibility, content extraction, preservation policy, text tagging, archival format requirements

1.1 Motivation

Enabling potential consumers to use and assess digital information in the future is the fundamental goal of digital preservation systems. They implement the required workflows and procedures by providing and establishing the processes, human experts and technical infrastructure. Part of the archive’s mission involves assessing the uncertainty about future developments in both technical and academic practices. Anticipation of the future is not an easy task and involves constant review of existing technological risks and procedures and a potentially changing designated community of information users. The accepted reference model for digital preservation systems is the Open Archival Information System (ISO 14721:2012, OAIS)[11] (also available as the magenta book from [4]).

Partners, who deposit digital objects in our digital preservation system[17], are oftentimes unsure whether to use PDF or PDF/A as the file format for textual data and ask for our guidance on that subject. Based on our risk assessment, general observations and discussions within the preservation community, we concluded that it would be useful for everyone in the community to have a discussion about risks and strategies involving PDF/A in a digital archive.

An anecdotal side note: one of our colleagues in our digital preservation working group is blind, so we have to ensure the
accessibility of information or documents that we produce. He acts as a litmus test. If he cannot read the information contained in a digital object, an algorithm will also have difficulties extracting and processing that kind of information.

A short disclaimer: I don’t have a solution but instead present some strategies on how to deal with (predominant textual) documents in the field of digital preservation. The scope of the discussion is electronic documents deemed as content containers for long-term preservation from the cultural heritage and academic domain, not business-related records or documents from the print publishing domains.

An apology might also be in order: I am aware that the title of this paper is a bold and provocative statement. "Considered harmful” articles have a long tradition in the computer science domain to convey meaning. Examples for glyphs are the letters of latin script, Japanese syllabaries or punctuation marks. These glyphs are available as electronic typefaces that consist of lists of drawing instructions for rendering the glyphs. The pages are represented as scalable vector graphics saving storage space in comparison to bitmaps. They do not suffer from pixelation and have to be rasterized (converted to raster images) in the resolution of the output device prior to being displayed or printed. Laser printers had specialized (converted to raster images) in the resolution of the output device and might prevent correct identification of older PDFs. The objects in the body are the components that represent the content of the document is organized in pages which allows faster navigation to a certain page without requiring the execution of all the PostScript commands of the preceding pages.

2 BACKGROUND

2.1 A short history of PDF

PDF is a file format that captures the layout of pages. It was developed by John Warnock and others at Adobe in the Camelot Project[31] in the early 1990s to replicate the convenience of sending documents with a fax machine in computer systems.

The prevalent page description method at the time (in desktop publishing) was using the interpreted PostScript language. PostScript programs contain instructions for laying out geometric forms (lines, curves etc.) and glyphs on rectangular planes (pages). Glyphs are graphical symbols that can be recognized within a writing system to convey meaning. Examples for glyphs are the letters of latin script, Japanese syllabaries or punctuation marks. These glyphs are available as electronic typefaces that consist of lists of drawing instructions for rendering the glyphs. The pages are represented as scalable vector graphics saving storage space in comparison to bitmaps. They do not suffer from pixelation and have to be rasterized (converted to raster images) in the resolution of the output device prior to being displayed or printed. Laser printers had specialized hardware to provide the processing power and lots of memory that were required for rasterization. Hardware to handle on-screen display of those documents was not widely available in desktop computing at that time.

PDF reduces the computational burden of the display device by executing the necessary PostScript programs during the creation of the PDF file. A single file saves thus only the graphic command results (called objects within PDF) required to render the pages, embedding raster image data along with font type information or even the digital fonts themselves. Another advantage over its ancestor is that the document is organized in pages which allows faster navigation to a certain page without requiring the execution of all the PostScript commands of the preceding pages.

The fixed page layouts of documents could (and still can) be faithfully displayed by limited computing devices or printed in high quality while being small enough to be sent through electronic networks.

Adobe extended the PDF specification multiple times over the years to allow for more features like encryption, transparency, device-independent colors, forms, web-links, javascript, audio, video, 3D objects and many more[18].

The usage of and commercial success began with the release of the free Acrobat Reader 2.0 in 1996 for PDF 1.1 and licensing all patents royalty free for everyone using its format. It became the de facto exchange format for electronic documents and version 1.7 was finally standardized by the International Standards Organization as ISO 32000-1[15] in 2008.

2.2 Technical introduction to PDF

Let’s begin with a brief introduction of the technical foundation of all PDFs.

2.2.1 File structure of PDF. A basic PDF file consists of four sections: a header, a body with objects, a cross-reference table, and the trailer. An example is shown in table 1.

<table>
<thead>
<tr>
<th>Header</th>
<th>0 %PDF-1.5</th>
</tr>
</thead>
</table>
| Body   | 16 obj <<  
| Objects| /Pages 2 0 R >> endobj  
|        | ... 4 0 obj <<  
|        | /Length 53 >> stream BT  
|        | /Fi 11 Tf 10 40 Td (Lore Ipsum)Tj ET endstream  
|        | endobj |
| XRef   | 384 xref  
| Cross-referencetable | 0 5  
|       | 0000000000 65535 f  
|       | 0000000016 0000 n  
|       | ... trailer  
|       | << /Root 1 0 R /Size 5 >> startxref  
|       | 384 %EOF |

Table 1: File structure of PDF and “Lore Ipsum” example

The header specifies the version of the PDF file. Until the 10.1.5 and 11.0.01 updates Adobe Acrobat products have historically opened a PDF as long as the %PDF-header started anywhere within the first 1024 bytes of the file. No checks were performed on the extraneous bytes before the %PDF-header[8], which can be a security risk and might prevent correct identification of older PDFs. The objects in the body are the components that represent the content of the

2Using a lite version of "Lorem Ipsum": “Lore Ipsum"
document. These objects for example are fonts, pages, text, sampled images, rendering instructions and so on but also data structures such as strings, arrays or dictionaries. Text in the context of PDF describes operators that paint text using character glyphs defined in fonts and not text in the usual sense. Starting with PDF version 1.5 objects can also be stored as object streams (which can be encoded or compressed using filter algorithms to save space).

As the objects can be stored in any sequence in the body, the cross-reference table (xref-table) stores the location of each object within the file stream for faster random access. Finally, the trailer contains the location of the cross-reference table, its size and a reference to the object containing the document catalog, the starting point of the object hierarchy. The trailer has to end with %EOF marking the end-of-file.

PDF supports incremental updates of its content. New objects, a new cross-reference table and a new trailer can be appended to the end of the file, if the content of the PDF is updated, without the need to rewrite the whole file. As objects can be marked as deleted in the xref-table there is no need to delete the corresponding objects in the body section.

2.2.2 Parsing a PDF file. As PDF is a quite complex file format, this is just a brief description of the necessary steps taken by an application in order to display the document’s content.

The parsing of a PDF file begins with checking the header signature to identify the version and to look for the last (the most recent) end-of-file marker. The xref-table is located via the \texttt{startxref} entry in the trailer and read into memory. The trailer also points to the document catalog via the /Root element.

The objects referenced in the document catalog are then parsed in order. The root object in table 1 for example only refers to the second object (the string \texttt{2 0 R}). The body section continues as

2 0 obj <<
/Type /Pages
/Count 1 >>
endobj

There is one child (/Kids) object of /Type /Pages. The page object

3 0 obj <<
/Parent 2 0 R
/MediaBox [0 0 612 792]
/Resources <<
/Font << /F1 <<
/BaseFont /Arial /Subtype /Type1 /Type /Font
>> >>
/Contents 4 0 R
/Type /Page >>
endobj

of type /Page defines the dimensions of the media box (the rectangular canvas for that page) and the resources used. Here only a single font /F1 is used. PDF can also define different rectangles useful in print like crop boxes, bleed boxes, trim boxes, and art boxes (refer to the PDF reference[7] for additional information).

The content of the page is contained in the fourth object and renders the symbols \texttt{Lore Ipsum} by executing the glyph drawing instructions from the Arial font on a certain location.

The encoding of the glyphs to render happens to be 7-bit ASCII code points with no additional positional parameters. A code point is a concept in the character encoding terminology and is used to distinguish between the binary number in an encoding and the abstract character in a particular graphical representation. As the primary focus of PDF is the reproduction of page layout, most strings are most certainly not as simple and often also contain positional parameters. The encodings from PDFs for example generated by Word 2011 or TextEdit(Mac) as seen in table 2 store an array of strings and geometric offsets. GoogleDocs exports the strings as hexadecimal numbers. Those don’t encode the standard ASCII ordnance but choose to offset it by \texttt{-29} (or \texttt{-0x1d}). A character map links this encoding to to the standard ASCII character set in which this particular font is organized. The character maps mentioned are included in the appendix. LibreOffice Writer’s PDF export on the other hand simply assigns increasing numbers to represent the string and links to those glyph code points via its character map.

3 PDF/A AS A SOLUTION FOR LONG-TERM PRESERVATION?

PDF/A is motivated by leveraging PDF’s characteristics of familiarity, ubiquity, acceptance, portability and reliability across a diverse range of platforms and communities for the purpose of preserving documents in the long-term.

3.1 PDF/A ISO standards

A constrained version of PDF for the purpose of archiving was based on PDF 1.4 and standardized in 2005 as ISO 19005-1[12] (PDF/A-1) with PDF/A-2[13] following in 2011 based on ISO 32000-1 (PDF version 1.7) and PDF/A-3[14] in 2012. The different PDF/A versions are not meant to be backwards compatible as they support different use cases. An overview of PDF/A flavors is given in table 3. The PDF/A standards differentiate between the conformance levels basic (b), accessible (a), and from version 2 onwards the intermediate level unicode (u). Accessible (level a) PDF/A functionalities require tagging of structure and content.

3.2 Tagging and PDF/UA

Assistive Technology (AT) is made up of software tools that can extract meaningful information from electronic documents and provides users with disabilities a means of “reading” and navigating the content. To extract information from content in PDF, tags can be attached to PDF objects from version 1.4 onward. These tags act as markup to denote the logical structure (semantic elements), and logical order (flow) of the content. Tagged PDF should provide markup for any real content in the document in contrast to artifacts like page numbers or other content outside the logical structure. Real content comprises all graphics objects (glyphs) that have been originally introduced by the document’s author. Artifacts are those graphics objects that are not part of the author’s original content. All content shall be marked in the structure tree with semantically appropriate tags (i.e. headings, formulas, paragraphs and such) in
The logical, intended reading order. Content information shall also not be conveyed by contrast, color, format or layout, unless the content is tagged to reflect all intended meaning.

The standard ISO 32000-1 states in section 14.8.2.2.2 note 3: “The purpose of Tagged PDF is […] to provide sufficient declarative and descriptive information to allow it [the conforming reader application] to make appropriate choices about how to process the content.”

Information for appropriate tagging is most of the time readily available to the creation software of the document (e.g. word processor) and has to be used by the tool that creates the tags in the PDF document. Tags can also be attached manually to documents that are already in PDF form, but this process is quite laborious and error prone.

A standard for required tag usage was published by ISO as ISO 14289[10] known as PDF/UA in 2014 (thus after the publication of PDF/A-2/3). Even though being accessible by AT (i.e. software) is a legal requirement in some domains, creating compliant documents is still a complex and cumbersome endeavor. Even assessing compliance to PDF/UA is quite hard: The Matterhorn protocol[24] provides a testing model that defines 31 checkpoints comprised of 136 failure conditions encompassing file format requirements for AT accessible PDF/UA of which some are not applicable to PDF/A (e.g. related to javascript). While 87 failure conditions are determinable by software 47 usually require human judgement or assessment. Failure condition 06-003 for example is machine testable and requires the metadata stream to contain a dublincore:title while 06-004 requires that the title clearly identifies the document in respect to human knowledge, a check that obviously is not decidable by algorithms.

PDF 2.0 ISO/DIS 32000-2 will clarify tag usage identified while working on PDF/UA among other enhancements and is currently under development. At the time of writing a fourth draft is available from ISO[16]. It is reasonable to assume that PDF 2.0 will be the foundation of forthcoming PDF/A flavors.

4 DISCUSSION

The discussion of possible risks and shortcomings of PDF/A for the purpose of digital preservation will be split between observations regarding creation and reuse of PDF/A documents and an attempt to identify or imagine possible (re-)use cases of the future.

4.1 Inadequacies of PDF today

Even without the prospect of problems in the future, PDF(/A) already has some shortcomings today from a usability point of view apart from the accessibility issues mentioned above.

As the primary concern is glyph placement on pages, PDF does not support a standard way of navigation. Although PDFs can
contain a table of contents that link to different sections within a document, page-based navigation is a physical feature of physical paper. The page dimensions are fixed within the PDF, with page sizes based on ANSI US Letter and ISO/DIN A4 being the most common but with different aspect ratios.

PDF also does not provide different perspectives on textual content. Electronic documents may want to provide different views of the text or data, either in multiple languages, diplomatic or critical transcriptions, or from different sources.

Nielsen[23] argued in 2001 that the fixed, page-based layout of PDF is not well suited for on-screen reading in contrast to web pages or other hypertext documents. Lack of a standard way of navigation means that readers are often lost while following elaborate designs. They have to zoom and scroll while reading documents with columnar layouts or articles spread over separate pages. Following links within PDF documents in a reader application without a back button leads to frustration as one has to find again the location where the hyperlink originated. Reading fixed-layout documents is especially tedious on small screen devices like smart phones or high display refresh latency devices like e-readers.

Usability issues aside, Willinsky et al.[32] give an excellent overview about current issues with using PDF in the scholarly environment. They hope, that their observations will influence further development of PDF or even the "Great PDF Replacement Format (GPDRFF)."

In the cultural heritage domain, facsimile pages of digitized books or letters are often compiled into PDF for ensuring page order and to allow for convenient page turning. If optical character recognition results are available they also are embedded into the PDF as a invisible text layer over the corresponding areas in the image of the original. Selecting and copying this text may surprise the reader because OCR engines only recognize characters with uncertainty and the confidence metric values are not included in the PDF for assessment of quality.

Another challenge for data curators or archivists is redaction. Overlaying text with black boxes only obstructs the text but leaves the information in the document. Deleting text blocks in an update process of a PDF file may mark only the reference in the xref-table as deleted while retaining the object itself. It is very hard to be sure that a redaction was successful manually, because even visually identical documents may be presented very differently in structure and encoding.

4.2 PDF/A reuse

The ISO 19005-1:2005 abstract "specifies how to use the Portable Document Format (PDF) 1.4 for long-term preservation of electronic documents. It is applicable to documents containing combinations of character, raster and vector data."

PDF thus primarily encodes page layout information treating text as a graphical representation of glyphs. The purpose for storing structured texts (or data) that contain semantically defined bits of information for conveying knowledge in human and machine accessible form is supported only as extensions to the primary intentions.

An insightful analogue of the difference between human content understanding and machine extraction capabilities would be the visible communication of music. While storing the layout of sheet music is perfectly achievable with PDF the placement of note glyphs on lines with annotating glyphs for bars, clefs and so on, is it easily understood and transformed into audible sound by humans trained in reading musical notation. A machine would have a hard time extracting enough information to reproduce or compare the musical score.

The possibility to faithfully render PDFs on displays or printing devices is therefore not enough for many methods of reuse. Even simple, non-trivial use-cases of information reuse demand a PDF/A a-level conformance.

4.3 Creating PDF/A

The basic conformance level for PDF/A require the glyph information to be present in the PDF file as embedded fonts. For most use cases this is a straightforward requirement, but in some cases it might be prohibited by the license of the fonts used or the font may simply be unavailable for embedding.

PDF/A conformance level a require the representation of the logical structure. Creators "should attempt to capture a document's logical structure hierarchy to the finest granularity possible." (Section 6.8.3 of the standard). Missing appropriate tags can inhibit reuse of PDF content significantly as shown below.

This has to be supported by the creating software. While support for tagging in document creation workflows is widening, this feature is still very poorly supported even in the widespread tools Word for Mac 2011 or LaTeX.

Some of the problems mentioned below can be avoided by software that implements the more advanced (but optional) tagging features available from the standards.

4.3.1 Conversion. Converting "normal" PDFs to PDF/A a-level conformance automatically is not advisable as a lot of information may already be lost during the creation process of the document.

The standard states that "PDF/A-1 writers should not add structural or semantic information that is not explicitly or implicitly present in the source material solely for the purpose of achieving conformance." and that "It is inadvisable for writers to generate structural or semantic information using automated processes without appropriate verification."

The abstract for ISO 19005-2:2011 also clarifies that the standard "is not applicable to specific processes for converting paper or electronic documents to the PDF/A format, [...]" (emphasis added).

The most common conversion tools, Adobe Distiller and the open source software ghostscript, do not offer an option to convert "normal" PDF to PDF/A-a. The latter states in its FAQs that conversion is "basically not possible when starting from a source which is not itself PDF/A-1a compliant"[9]. The FAQs also give a more detailed rationale for not even attempting a conversion.

Successful conversion to b-level conforming PDF/A (i.e. embedding the digital fonts in the document and enforcing other restrictions) is easier to achieve. Licensing problems may arise in converting to PDF/A for example if the copyright holder of digital typefaces does not allow embedding in documents. Fonts with open
licenses like SIL Open Font License\(^5\) circumvent possible restrictions but also complicate conversion due to differences in substitute font dimensions.

### 4.4 Text extraction

Most tools for reading or extracting textual content from PDFs do recover strings suitable for searching within a document or allowing copy-and-paste operations. Full-text indexing only depends on the text strings (words) to find relevant documents. Reusing copied text extracted from PDFs on the other hand oftentimes require removing artifacts like page numbers or footnotes. What constitutes a word and finding word boundaries might be difficult by itself depending on the layout or script of the text. Selecting rows or columns from tables in PDF reader applications often also results in frustration.

In rare cases even full-text indexing can go wrong with PDF/A b-levels, if the encoding of glyphs is somewhat off a standard encoding. A-level documents will have a higher success rate as they do require ToUnicodeMapping and comprehensive tagging.

But even PDF/A a-level conformance may not guarantee full text recovery due to the fact that some tagging features are only recommendations and not mandatory. Hyphenation (the word division at the end of a line) shall be treated as an incidental artifact and be represented as a unicode soft-hyphen (U+00AD) instead of a hard-hyphen (U+002D) as suggested by the standard. "The producer of a Tagged PDF document shall distinguish explicitly between soft and hard hyphens so that the consumer does not have to guess which type a given character represents." It is alternatively possible to provide the /ActualText attribute without the hyphen.

Searching for the string "Rheinland" (German for Rhineland, a part of Germany) in the PDF/A-1a file of the nestor newsletter number 28\[22\] for example would result in no matches in macOS Preview or Adobe Reader as it is stored as a hard-hyphen. The hyphen in "Ostwestfalen-Lippe" is a regular one.

Figure 1: nestor newsletter 28 excerpt

### 4.5 Content extraction

Content extraction is more than mere text extraction as it tries to extract structured and semantically meaningful bits of information or data from a document. A research article for example may consist of a title, author information, abstract, sections, formulas, references, tables, diagrams, and so on, all of which require different methods for identification, extraction and encoding to recover the contained information. The logical structure and physical layout of the document may also be different for the various research communities and journals. Reusable content in contrast to full-text require the extraction of the structure of the text or the narrative flow. Deciding how two blocks of text are chained together if not properly tagged demands the use of layout analysis not unlike that used in optical character recognition software. Naive extraction might interrupt the text flow by mixing the main narrative with footnotes, side-notes, captions, pagination artifacts, or wrong columnar content in a multicolumn page layout.

Two reports from data intensive fields in disciplines that depend on content extraction from text and data published in PDFs as information containers will be examined: archaeology and bioinformatics.

#### 4.5.1 Archaeology

In archaeology the de-facto standard for the sharing and exchange of so-called grey literature is PDF. Grey literature in that field is the main documentation of fieldwork or other archaeological investigation. They often combine descriptive text and reports of findings with rich media such as raster images, vector or CAD drawings, geographic shape files or maps and even screenshots.

As Evans and Moore from the Archaeology Data Service (ADS) in the UK describe in their case study\[6\], these content containers can be easily compiled but have dramatic effect on the reusability, i.e. the extraction of data or datasets with software tools. With a focus on content processing using NLP (natural language processing), they conclude that "the [data] reuse limitations of PDF/A are evident; that is any PDF/A-1 file is designed for 'human consumption' such as reading, printing and copying of text and graphics. [...] However, it is a point that needs to be re-enforced by practical experience, notably the difficulties in using 'text-based' reports for machine-based language processing and indexing."

They also suggest, that "Perhaps the future challenges are not just in ensuring that the PDF/A standard is used consistently and accurately, but that other avenues are explored to enable the information within files is not just limited to the human eye." This is especially true for reusing flattened, embedded objects like maps.

#### 4.5.2 Bioinformatics

Biomedical Natural Language Processing (BioNLP) is trying to help biologists to establish semantic relations between articles published in different journals or fields in biology and between this literature and databases across huge corpora. These relationships are for example protein-protein interactions or gene-disease-phenotype relations.

Ramakrishnan et al.\[25\] for example use a layout-aware text block detection algorithm to extract contiguous blocks of text from PDFs and identify section parts like headings, subheadings, and text body or paragraphs and remove artifacts. They then try to classify these into rhetorical categories like abstract, methods, results, references and so on. The blocks are finally assembled into a structured text that can be further processed with NLP-techniques like Named Entity Recognition. They conclude that although feasible, their method requires prior knowledge and has to be adapted to different journal formats and layouts.

#### 4.5.3 Legal issue: Patents

Apart from the technical difficulties, using methods for text and content extraction from PDF may also be a legal issue. Textual extraction from PDF is considered so involved as to be worthy to be granted patent status from the US Patent Office. US patent No. 9098471\[26\] for example covers a method for document content reconstruction from an "unstructured document format" (sic!) to a markup language in the usual broad description of patent applications.
4.6 Validation

Digital preservation workflows require some sort of checking whether files adhere to the specification of the file format they claim to be. The complex structure of the file format and the sometimes ambiguous specification in the case of PDF and PDF/A made this a problematic endeavor.

For some time, the go-to-tool for PDF/A validation was JHOVE\(^3\) using PDF profiles. As it was discovered that it was not suitable for validating PDF/A files\(^2\), the EU funded PREFORMA project\(^6\) included a provision to create veraPDF\(^7\), a validator which aims at checking conformance of all PDF/A flavors while also allowing for policy checks that are customizable to institutional policy. The goal is to codify the ambiguity of the specification in computer language and provide a comprehensive tool for testing file format validity, taking into account the requirements and constraints imposed by the various PDF/A standards.

This helps a lot but does not address the question whether the content of a PDF file is truly (human and/or machine) accessible and usable with regard to the aspects mentioned above. Being able to validate a file is a necessary condition but it gives no comprehensive answer about potential risks concerning future usability.

4.7 Suitability for long-term preservation

Keeping digital objects discoverable and viable is the core function of digital preservation systems.

Digital archives are tasked with inquiring about and anticipating the needs of a designated community of future users, who might value the preserved content, discern its relevance and should be able to reuse it. But designated communities might change in the future and even the identified designated communities might not know how to (or don’t want to) use the material in its present form and format.

PDF/A is perceived to be an archival solution for digital documents. Discussion within the community revealed the reason for that is three-fold: Firstly, it is marketed as an archival format. The A in PDF/A might stand for “Archive” or “Archival” or simply for the letter “A”; I haven’t found any official explanation for the choice of A in the acronym. The second reason may be that it is used by so many institutions to a point where a critical mass is reached. They cannot altogether err in their risk assessment, so the reasoning is that you simply cannot be wrong when you run with the flock. And thirdly, there does not seem to be a better alternative available (see below).

Comprehensive policies regarding the use of PDF in archives seem to be rare. An analysis of risks and benefits of PDF and content reuse in digital archives has been published by Moore and Evans\(^21\). Another analysis for using PDF/A-3 (which allows for embedded files) has been compiled by the National Digital Stewardship Alliance in its report on “The Benefits and Risks of the PDF/A-3 File Format for Archival Institutions”\(^1\). Using PDF/A as a container for files complicates preservation workflows and might be considered an additional risk.

The benefit and convenience of PDF to easily capture all kinds of textual and graphical information in an electronic equivalent of a stack of paper comes at a cost for digital archives. In the digital preservation workflow technical validation is an essential step to ensure files are valid with respect to the specification of the file format they claim to be. This process will always be costly as it involves manual assessment as the tools are not yet usable for a fully automatic workflow (see this recent report on JHOVE\(^19\)). This burden is lessened if an archival format is less complex and more focused on retaining all or most identified significant properties of the data and information to be preserved.

Despite the reusability issues, exporting to PDF sometimes also results in significant loss of information apart from text structure. Two examples: Spreadsheet formulas and numerical precision are lost, making testing data sets more difficult. Storing OCR results as invisible text over the digital facsimiles loses the confidence values for characters of the recognition software.

In the end, even if PDF/A is validated (by machine) and rendered correctly (by human visual inspection), the availability and validity of structural markup and Matterhorn protocol compliance is extremely difficult and laborious to assess.

4.8 Strategies for long-term preservation

Content in PDF/A form perhaps cannot be avoided altogether and has also already been ingested into archives in huge numbers. Knowing about the risks and benefits is essential for establishing policies regarding submissions. Digital archives have to have strategies and policies in place anyway to avoid being unable to provide useful and relevant content back to archive users. Digital preservation is a process involving not only the archive but also the producers, so there might be the possibility to negotiate better-suited or alternative deposits within the submission agreements.

Some possible strategies for the better handling of PDFs mostly involve the content producers but also create more involved workflows within the archive:

- Negotiate non-PDF documents better suited for their domain and supported by your archive system.
- Consider using PDF/A as a dissemination format only (and therefore use a PDF rendition server only for access not ingest).
- Save the original source documents alongside the PDFs for full text and structure retention. With PDF/A-3 these could be embedded and linked as source of the document.
- Require data producers to implement workflows that adhere to the Matterhorn protocol to assure fully meaningful tagged PDFs (including MathML formulas, semantically tagged data and so on) and to provide /ActualText for every textual information contained in the PDF that is not easily extractable otherwise.

The feasibility to assess and compliance check such PDF/A files automatically remains to be evaluated.

4.9 Possible requirements in the future

As a famous quote says: “It is difficult to make predictions, especially about the future.” But there will always be visionaries that try to push the boundaries of the status quo from the impossible into the
viable. Vannevar Bush envisioned the Memex in his 1945 essay “As we may think”[3], a device to access and organize potentially all human knowledge. This vision to have all relevant information available at your fingertips and to combine bits of information has inspired others like Douglas Engelbart, Steve Jobs or Tim Berners-Lee to create innovative technologies to assist people in accessing, using, combining, and understanding information more easily.

Technology will be the key to accessing knowledge. And therefore technology has to be able to access information. The vast corpus of documents on the web would not be manageable or discoverable (and thus accessible) without search engines that harvest, process and organize information to quickly find relevant pieces of information.

Research papers are generated in such an amount and with such a velocity that even today we depend on machine-based assistance to sift through them. Machine-learning technology to extract and organize information is nascent and might be an essential tool to deal with publications and data in the future. It might even help with extracting content from PDF.

Moreover traditional aspects of academic routine will also change. Organizing information within rectangular boundaries is not inherently given and most of the time adds no additional structure to textual information. The concept of a “Page” is merely a convention due the physical constraints of the medium. Pages are useful for citation in the traditional format of books or journals but with the advancement of digital publishing and linked data technologies it will be more useful to refer to information sets identified (and locatable) by persistent digital identifiers like URIs or IRIs. Relevant excerpts can then become part of the textual narrative and might render traditional references obsolete. Linked data technologies and web annotations[5] require identifiable bits of information (resources) and probably will be part of the scholarly review processes, contextualizations, and sources of new insights.

Even today, with the internet, the expectations of how to access information, how it’s organized, structured, and connected to other pieces of information of relevance are different from the common practice of just some years ago. In the “Teens React to” video[2] about teenager views on a physical World Book Encyclopedia, one can perhaps observe a glimpse of the future: ”It takes forever, this is annoying,” Alix, age 19, said in the video trying to find information. “This is why I don’t use these.” One teenager even wondered why YouTube isn’t mentioned in a book from 2005.

5 FUTURE WORK
Assessing possible structural and semantic reuse of information is not a simple task, even if it is encoded in a structured plain text format (with known character encodings[33]).

Tools and workflows providing Matterhorn protocol, PDF 2.0 and Web Content Accessibility Guidelines (WCAG) 2.0[30] compliant tagged PDF/A files need to be improved to fit into real world content creation processes.

Possible paths that might be worth exploring would be, for example to devise better tools for assessing accessibility, especially accessibility for machine-based methods for content extraction from PDF/A, research machine-learning methods for knowledge extraction to support discovery, linkage, and semantic topic labeling, or to investigate possible alternatives for common document use cases (see below).

Another aspect to further investigate is how to prove authenticity of the content if the archival format is normalized from PDF to some other archival intermediate format as the integrity (and fixity) of the PDF file does not transpose easily to the content itself. How to assess the invariance of the significant properties?

5.1 Alternatives
Today, it seems, there is no viable alternative to PDF as a universal digital container of everything that can be flattened to printed pages. An ideal archival format has to be as simple as possible, able to be validated, retain the identified significant properties of the document depending on the designated community domain, be reasonably adopted within the archival community, and supported by tools to generate dissemination objects.

Although not as widespread as PDF, there are some alternative document formats, containers, and tools that might be worth investigating for certain use-cases.

Some examples for declarative, semantic, or document markup languages are Markdown flavors, HTML/CSS, ODF/OOXML, TEI, JATS, or even TIFF+OCR. Some of them can be converted automatically to PDF easily, others require layout information like Cascading Style Sheets (CSS) or XSL-FO (Formatting Objects). More elaborate semantic markup like TEI/XML may require human intervention.

5.1.1 Markdown. The textual markup of Markdown variants is machine actionable while being human friendly to read at the same time. It is suitable for structured texts (including lists and tables) where the exact layout is not as important. Markdown is not well suited for validation.

5.1.2 HTML/CSS. Hypertext Markup Language (HTML) is the universal language of web documents and supports the separation of semantic markup in HTML with display style commands in Cascading Style Sheets (CSS). Like PDF, HTML/CSS can place graphical elements on rectangular regions. In contrast to XHTML, an XML language, it is very robust to formal errors. WebArchive (WARC) files bundle all necessary components and are already in use in digital archiving. It is also used by the ePub file format (common for eBooks) essentially combines HTML with the corresponding style sheets and (navigation) structure in a ZIP container.

5.1.3 ODF/OOXML. The office document file formats Open Document Format for Office Applications (ODF), native format of LibreOffice, and Office Open XML (OOXML), native format of the Microsoft Office suite, are XML-based and ISO standardized as ISO/IEC 26300 and ISO/IEC 29500 respectively. They retain structural, textual and tabular information alongside diagrams, images, and formulas for content extraction and provide style information for display.

5.1.4 TEI/XML. The P5 guidelines of the Text Encoding Initiative propose a wide-ranging tag set for rich semantic markup of scholarly texts like editions, plays or transcriptions. They are used mostly within the digital humanities.
5.1.5 **JATS.** The Journal Article Tag Suite (JATS) is an XML format used to describe scientific literature and standardized as ANSI/NISO Z39.96-2015. It has been adopted within certain open access journals and repositories such as PubMed Central, some of which require content to be "JATS+PDF". JATS can be validated and converted to PDF, ePub, or HTML.

5.1.6 **TIFF+OCR.** For scanned text-containing artifacts the scanned images could be preserved alongside the OCR results either as ALTO-XML or HOCR.

A universal tool for document conversion is for example Pandoc\(^8\). It is free, open-source software and converts between various document formats. Pandoc includes support or has plug-ins for reading, transforming, and writing Markdown, Office Documents, JATS and other XML-based markup, HTML, and also for LaTeX or ASCII based DocBook. Further investigation may provide insights about its suitability for creating PDFs from these formats within digital preservation workflows if there is the need to provide PDF dissemination copies.

6 CONCLUSION

Digital archives act as facilitators for future research and researchers. They have the responsibility not only to safeguard the information they have been entrusted with, but also to maintain the utility thereof. Because digital archiving and preservation is a process that involves not only the archive but also the data producer, archives have the responsibility to inform about risks, provide training and good practice, and negotiate appropriate measures for content usability if possible.

As appealing as the benefits of PDF/A may appear, even the standard development team was aware of most of the shortcomings of PDF/A. Sullivan reports in her 2003 article (emphasis added): "The intent was not to claim that PDF-based solutions are the best way to preserve electronic documents. PDF/A simply defines an archival profile of PDF that is more amenable to long-term preservation than traditional PDF"[27]

Familiarity of PDF led to fast and widespread adoption of PDF/A as a solution in the field of digital archiving. This fact may have muted prophetic voices demanding the quest for and development of more suitable content containers for research work (text and data) with reuse in mind. After all there seemed to already be available a solution for it. And you cannot be wrong by choosing the accepted standard as preservation policy.

As Nathan C. Thomson quotes Matt Ridley on page 32 in the book "Society's Genome"[28] in respect to sequences in DNA:

"[...] the distinction between two kinds of rubbish: 'garbage which has no use and must be disposed of lest it rot and stink, and 'junk which has no immediate use but does no harm and is kept in the attic in case it might one day be put to use. [...]"

Let's try to retain only the junk but not too much garbage.

Finally, this paper wants to summarize the advantages, risks, and misconceptions about the suitability of PDF/A as an archival file format for long-term preservation. It might start a much needed discussion within the different stakeholder communities to mitigate problems in the future.

ACKNOWLEDGMENTS

This work has been supported by the Senate Department for Culture and Europe of the State Berlin.

The author would like to thank Elias Oltmanns, Kilian Amrhein, Tim Hasler, Wolfgang Peters-Kottig, Heinz Kuper, and innumerable others from various communities for their valuable input, insights and discussions.

REFERENCES


\(^8\)http://www.pandoc.org

A APPENDIX
A.1 Character map: GoogleDocs export
/CMapName /Adobe-Identity-UCS def
/CMapType 2 def
1 begincodespacerange <0000> <FFFF> endcodespacerange
6 beginbfchar <0000> <0020> <0020> <0049> <002F> <004C> <0065> <006D> <0075> endbfchar
2 beginbfrange <0000> <0053> <006F> <0072> <0065> <0070> <0073> <0075> endbfrange
cmap

A.2 Character map: LibreOffice Writer export
/CMapName /Adobe-Identity-UCS def
/CMapType 2 def
1 begincodespacerange <0000> <FFFF> endcodespacerange
10 beginbfchar <01> <004C> <02> <006F> <03> <0072> <04> <0065> <05> <0020> <06> <0049> <07> <0070> <08> <0073> <09> <0075> <0A> <006D> endbfchar
ABSTRACT

There are still a significant number of born-digital artefacts which have yet to be received by memory institutions. These include the works of famous or important authors, artists, politicians, and musicians, some of which are stored on digital media dating back 30 years or more. Many digital objects have escaped the blight of bit rot, and legacy hardware, though rare, is still available for reading the media. Assembling the required equipment and expertise is possible but increasingly difficult.

This is an ongoing challenge which is not confined to floppy disks or obsolete hard drives. The challenge is exacerbated by the reality that many artefacts are not considered for preservation until the creator has passed away. In addition to the past 30 years’ worth of media, today’s storage media will also soon be superseded with newer technology and form a new challenge for future archivists. In order to preserve the content, or the look and feel of the original objects, precautionary actions must be taken which require both knowledge and the right equipment.

Creating disk images from obsolete magnetic media is not a simple task, especially when it concerns more obscure or older formats such as 8-inch disks. There are many challenges to overcome, with each small success leading to a new phase of discovery.

In this case study we recount our progress, through many stages, of dealing with the described challenges with regards to a set of 8-inch magnetic floppy disk media. At the beginning of this project, most requirements for recovering the data from this media were unknown: the original hardware system, the original operating system, and the disk format. In this paper, we describe our approach to uncovering this information, leading to a successful preliminary outcome. This is a cautionary tale which aims to provide some lessons for use in related contexts.

CCS CONCEPTS
• Information systems → Magnetic disks;

KEYWORDS
Digital preservation; disk imaging; disk format; digital heritage

1 INTRODUCTION

There is very little literature in the digital preservation discipline regarding the challenges that must be overcome to retrieve data from obsolete media and render it accessible, complete and accurate. For instance [20] has no use cases on 8-inch disks and the actual information referenced is often rather scarce [4]. Many practitioners in digital libraries and archives presume that someone is aware of the technological changes and takes the necessary action to mitigate it. But digital preservation is still a relatively young discipline and many IT professionals have still not placed sufficient priority on media migration. Additionally, well-established workflows to hand over materials from governance offices to archives still use 20–30 year pre-transfer retention periods, which were reasonable for traditional paper-based records but are much too long for any digital storage medium to be held without intervention by media preservation experts. The situation is even worse in the personal sphere: written documents of famous authors or politicians are still valuable after their death [9] but there is little to no availability of standard processes or policies regarding the transfer of any digital equipment and/or storage media at this difficult moment.

While organizational knowledge exists that There is little existent literature that would recommend removal media of any type to be a worthy archival medium. [10], this is not necessarily the case for the personal domain.

This case study presents the challenges and difficulties faced when attempting to save research data from 8-inch disks and the digital forensic methods needed to succeed. From the very start of this endeavour we had very little information from which to organize our work.

The brief given to us was succinct. There is a set of 8-inch disks from early- to mid-1980s which contain research data. These data are required for current research. Retrieve the files from the disks and make them available on current media.

The hand-over procedure did not generate much information, seventeen 8-inch disks with some additional paper stuck in a few of the disk sleeves were passed to us (Figure 1). The floppy labels of the various brands (e.g. BASF and HP) did not provide any information on the hardware or systems used to create them or on the formatting used.

The few sheets of papers accompanying the disks seemed to indicate that no file system was used at all. It appeared to be head, sector and track information. While the column containing the head information (0 for first side, 1 for the second side of a two sided medium) was rather obvious to deduce, it took a few comparisons to understand the meaning of the sector and track information. There is no record within the university about the hardware systems used at the time and inquiries within the organization only hinted that they had introduced office writers (electronic IBM typewriters) and later CPM machines, both featuring 8-inch drives. Unfortunately, nobody had a clear memory of the exact systems. Due to the number of Hewlett-Packard (HP) disks in the set, it was presumed that an
Figure 1: Sample of the disks by different manufacturers evaluated in the study

The HP system could have been used and further investigation into the part numbers was tried. The sheets, with some of the disks, list information formatted as

<table>
<thead>
<tr>
<th>NAME</th>
<th>PRO</th>
<th>TYPE</th>
<th>REC/FILE</th>
<th>BYTES/REC</th>
<th>ADDRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>H8</td>
<td>0</td>
<td>1</td>
<td>148</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ALT-A DATA 1 1188 0/1/0

This we determined was from a HP CAT command, which is interpreted as probably from an HP 984535 series. H is the drive type, 8 the controller select code, 0 the drive address, and 1 the unit address. The first entry, in this example, indicates a large data file of 1188 bytes, file type unknown.

Extracting data from 5.25-inch and 3.5-inch floppy disks, which were used more recently and more commonly, is readily achievable as it is still possible to find systems which feature a 3.5-inch drive and systems which have at least a floppy controller present. At a computer centre it is still often possible to find all components necessary in order to get working hardware set up to read DOS formatted disks (MFM recording format, [12]), which was the most common format.

The case for 8-inch disks is more complicated. Eight-inch floppy disks were generally out of use by the beginning/mid-1990s and were not a common part of typical office or personal computer systems at that time; thus, they never obtained the interest of a wider enthusiast scene as did the home computers of the same era. The 8-inch disks have a completely different form factor and different connectors for both data and power supply. Thus, there is no standard support for them in most operating systems, and very little general information on the Internet.

2 HARDWARE REQUIREMENTS

The status of the research data was not entirely clear and there was some indication that the disks might contain private (clinical) information on living persons. Thus, the disks could not be sent off-site. As no service company or expert who was knowledgeable about these disks could be found within the vicinity, we started our study with the selection and acquisition of the hardware required to capture the bits.

The initial step was to find an appropriate machine which could both connect an 8-inch floppy drive and host a recent operating system to handle reading the data and transfer it to current media. Two machines were evaluated: An older HP AMD CPU office desktop machine (from approximately 2008) and a more recent DELL Optiplex system, both still featuring a floppy controller. The BIOS of the HP allows setting a single floppy to either 3.5-inch 1.44 Mbyte or 5.25-inch 1.2 Mbyte. The DELL BIOS made it more difficult to configure for the older floppies, as only the 3.5-inch option could be set. Additionally to make things more challenging, actually attaching a 5.25-inch or 8-inch drive to the DELL machine electrically required a power cable adaptor from the more recent 3.5-inch SATA/HDD connectors to the old peripheral power connector style 4 pin.

As no 8-inch disk drive was to be found in the university, a "new" one had to be bought. The market, in the 1980s, offered a wide range of drives each with their own special features. A half-height drive not requiring AC power was chosen. A Mitsubishi (M2896-63) 8-inch drive in mint condition was acquired for about 240 € from eBay. In addition, a converter for the power supply of the 5/24V of the drive and a floppy disk adapter for 8-inch drives on PC floppy controllers were procured from DBit [5] for a cost only marginally less than the drive. The necessary cables were scavenged from old systems and recycling bins (e.g. a 50-pin ribbon SCSI cable and multiple standard floppy attachment cables) and refitted to be used with the 34- and 50-pin connector of the drive [6] (Figure 3). The power converter either plugged into a spare molex power connector from the PC’s power supply or could be powered from a standard ATX mainboard power input (Figure 2).

Both the 24V power supply and the cable adapter board have power requirements and fit traditional PC power cable sockets. To connect the 8-inch drive without external power supplies, the machine needed to have a 3.5-inch floppy disk power connector and a standard molex power connector (for the 5.25-inch and 8-inch drives, the latter one via the voltage adaptor). Having wired all this, the setup for the 8-inch floppy reading experiment reached an initial working status.

Figure 2: Providing the 8-inch drive with 5V and 24V
Checking the system

As no-one was really familiar with the drive, some preliminary tests were run with standard floppy drives. To create a baseline for the experiments, the two machines were checked with both 3.5-inch and 5.25-inch drives and subsequent floppy reads (1.44 MB floppies for 3.5-inch drive and both 360 KB and 1.2 MB floppies in the 5.25-inch drive). We imaged 5.25-inch DOS formatted floppy disks using the Linux dd_rescue command and Linux mtools. The drives and controllers were confirmed to be working properly.

The BIOS settings in the PC did not need to change as they can be overridden when loading the Linux kernel. The only prerequisite is an enabled floppy controller. The Linux module floppy_ko takes (repeated) parameters in the form of Floppy=0,1 (indicating the physical floppy drive and the type of floppy respectively). Nevertheless, the Linux kernel is not aware of 8-inch drives and has no specific settings for them in the module.

The 8-inch drive was connected and the BIOS of the machine configured to 5.25-inch drive. It could be seen, by the flashing access LED, that the drive reacted to commands issued by the controller, for example, when loading the kernel module or running a program which required access to the drive.

3 PRELIMINARY READING AND WRITING TESTS

To conduct some preliminary experiments on the drive, an unused 8-inch floppy (manufactured by Nashua), of unknown quality, was taken from a display case of obsolete media in the computer centre. This was done to run some reading and writing experiments to see if the floppy could be formatted and read without endangering the original disks. One of the challenges was the unknown characteristics of both the floppies and the drive. For the latter there was a specification sheet which came with the drive. It was not apparent to which degree the controller/drive combination would be able to read floppies, which were created on completely different system(s). The same applied to the differences of the kernel configuration which only officially knows about 5.25-inch drives and has no knowledge about 8-inch drives. Unfortunately, our research revealed that there were multiple systems which used 8-inch drives quite differently with regard to format, capacity and geometry (e.g. just for IBM floppies of that format. [26]).

When doing these first trials, the kernel or tool messages indicated an “active write protection”. The re-configuration of the drive to WP (honour write protect) or NP (not protected) did not change anything in the ongoing formatting tests with fdformat. Mysteriously, after a while the "active write protection" disappeared for an unknown reason.

It was possible to low-level format the test Nashua floppy disk using a rather wide range of settings, but the verification failed for every setting. It is known that these formatting tests run smoothly on 5.25-inch floppy drives by a stepwise movement of the read/write head. However, on the 8-inch drive some settings, especially the higher density ones, produced strange jumps of the head and odd sounds. The DBit controller displayed the same track number on its two-digit seven-segment display as the fdformat gave for its operation (Figure 3).

Finally we did a simple reading test by inserting the disk and executing a directory listing command (ls /dev/fd0 and mkdir a:). All "read" tests of different disks, including those from the archive, failed as track 0 could not be read. This was a first indication that the disks were not of DOS environments origin, possibly complicating the access significantly.

4 USING THE BITCURATOR LINUX ENVIRONMENT

To rule out hardware-related issues such as cabling, the first suboptimal connection – a long cable with an open end which might lead to signal reflections – was shortened and a second edge connector cable was produced. The new one avoided any malfunction of the connector as it was an original 50pin.

A PC was then booted from a USB stick containing the BitCurator suite to avoid any hardware emulation a virtualization layer might introduce. The number of tools in this particular setup was limited to the command line floppy and raw IO tools available with the Linux operating system, such as: getfdprm, setfdprm, fdrawcmd, fdformat, mtools, dd, and dd_rescue.

One of the major challenges was the lack of experience with the 8-inch drives and floppy disks. Thus, trial and error was the approach used. The knowledge needed to be unearthed step-by-step hoping to find the relevant bits and pieces to get everything working in the intended way.

For this task the web site of the Linux floppy disk utilities [25] was consulted as the Linux man (manual) page only explains the command line settings but does not give examples. We reproduced various suggestions provided from the "How to identify an unknown disk using Fdutils" section. "Finding out the number of sides (heads)" helped us to determine that the test disk we used was actually single sided and that two sided formatting and verifying must fail. Further configuration parameters were checked and tested using fdrawcmd and setfdprm, followed by a successful low-level formatting tool. Error messages were checked from both
the tool’s printout and the kernel log file. The need within the wider Linux community for floppy disk tools will diminish further and the website, last updated in May 2005, is in risk of disappearing at some point leaving no or only a few traces.

**Disk Geometry and other Parameters**

For non-familiar users the options for the disk geometry remain a bit opaque, apart from the recording mechanism which is presumed to be MFM and cannot be changed. For example, typically the drive supports 77 cylinders (should equal number of tracks), but the IBM specification for type 2D (printed on the BASF disks) have a special index cylinder and then 74 usable cylinders [26]. While the index cylinder configuration remains the same, the remainder can differ significantly.

The label on the Nashua test disk revealed very little. There is a part number (FD-1D-WP) which we were able to verify is a single-sided double-density floppy disk. The specifications on the original manufacturer’s box listed on the side of the carton are as follows: FD-1D WP-R (the item number), S.S. / D.D. SOFT which translates to single sided, double density, and soft sectors. This explained the failures of the previous low-level formatting verification experiment: A single-sided disk can actually be formatted as double sided, but reading from the empty side will fail. Changing the parameters to single sided with `setfdmprm` progressed our attempts at verification.

The `fdrawcmd` can issue raw floppy controller commands in order to discover the data transfer rate, the sector numbering scheme, and the number of sides on a disk. We successfully ran this on the test disk to confirm the geometry. Unfortunately, we were not successful when trying the target disks. We had to rule out possible problems such as:

- The disks were not formatted at all. Even though these were backup disks, as implied by the labels, that is no guarantee that they had ever been verified.
- The disks may have been improperly stored before they were transferred into the university archive, thus losing most of the magnetic charge.
- The drive simply cannot interpret the recording format.

5 **USING AN OSCILLOSCOPE**

To identify if the disks contained any recorded information at all, an oscilloscope was connected to the analogue amplifier behind the magnetic reading head. A GW Instek GDS-1022 oscilloscope was linked to the disk drive at AD3 and AD4 on the read circuit (Figure 4). The signal was set to normal level 2.8V on TP-AD3 and 2.75V on TP-AD4. Signal peaks for both channels were 800Vpp.

The rotational speed of the 8-inch floppy drive is rather slow compared to modern equipment and the recording density is very low. Even with just 26 sectors of 128 Bytes, a substantial number of changes in the magnetic flux are required to represent the data. It was hoped that it would be possible to at least distinguish between an unformatted disk (or demagnetized disk due to bit rot) and a disk containing data: white noise in the first case and some visible, repeated changes in the second.

To trigger readings, the low-level program `fdrawcmd` was run with different options on several disks [25]. Initially, we focused on track 0 and started with the Nashua disk, using `fdrawcmd` to produce read signals, which produced a visible pattern of different frequencies on the oscilloscope. In further rounds, we looked for similar patterns on track 0 of the other disks. The picture obtained was quite different and looked more like white noise compared to the measurements on the Nashua.

Changing to a higher track number produced, for both classes of disks, patterns which were definitely visible. This could explain why we were not able to read anything with the methods we had been applying up to that point.

The oscilloscope setup had some limitations for the purpose of data interpretation. We were unable to store longer sequences of signal readings but could only freeze the content of the screen (Figure 5). For further stream interpretation a logic analyzer would be useful, e.g. a Bitscope Micro with DSO Data Recorder software [1].

The readings from the disks did not imply that they were completely empty. We had confirmed that our hardware setup of the 8-inch drive and the adapters were working, and that the disks seemed to contain something, our next step was to employ some more advanced approaches, such as forensic floppy controllers.

6 **KRYOFLUX AND CATWEASEL**

Having established that the delivered disks were not empty, different approaches were chosen. These circumvented the use of the standard floppy controller to allow a more direct access to the medium [13]. The KryoFlux and Catweasel floppy controllers are add-in cards which were designed to read disks at a low-level approach and circumvent most of the built-in PC floppy disk controller logic.

As the University of Freiburg does not have either of these devices, we asked for help from our colleagues at the Computerspielemuseum (Computer Games Museum), Berlin who have both...
The KryoFlux

The KryoFlux card is built into a tower case containing a power supply unit (PSU), a 5.25-inch drive and a 3.5-inch drive. The Mitsubishi drive was connected to the KryoFlux which in turn was connected to a Dell laptop running Windows XP. The 50-pin data cable was connected to the fdadap floppy disk adapter, then the 34-pin to the KryoFlux. The KryoFlux was connected via USB cable to the laptop. The 5V/24V power supply to the disk drive was provided as previously described.

For our initial test of our setup we used our test disk, which failed after reading only the first track. As this test disk has consistently failed, we concluded that it is a dud disk. We then tried to calibrate the drive. This was a bad move, as it calibrated only to track 0 and subsequent attempts at imaging also failed.

The Catweasel

Please note that the Catweasel is no longer manufactured as the demand and interest in magnetic floppy media has diminished. We knew of software that worked with Catweasel to image 8-inch disks, so we connected the Catweasel system to our disk drive and tried to run the Catweasel software ImageToo1s 3. However, as there is no configuration for 8-inch disks this did not work out. We then tried cw2dmk [17], as the documentation mentions 8-inch disks and that a variety of formats can be imaged. Unfortunately, this also did not work, returning the error “Failed to detect any drives”. To make sure that the Catweasel device was in good working order, the 5.25-inch drive was reconnected and we successfully imaged a Commodore 64 disk. Retrying the 8-inch drive, cw2dmk failed again with the same error. Later on it was discovered that cw2dmk does not work with Windows XP. Another backwards step!

The KryoFlux Again

After connecting it all up, we powered up according to the KryoFlux manual but the software reported that it did not recognize hardware. This was strange as we did have it working earlier in the day.

Checking everything again, we noticed that the circuitry was getting power from two sources and deduced that this was preventing the KryoFlux from resetting. We then unplugged power to all components except the laptop, rebooted and then plugged in the KryoFlux and then powered on all downstream components. This was the correct way to do it. The KryoFlux recognized the drive and we were able to start imaging.

The imaging was set to output KryoFlux Stream Files, preservation and all other settings to the default values. This meant that 83 tracks were tried on each side of the disk, resulting in “Error reading stream device” from track 77 onwards (with tracks starting at 0).

We managed to get an image for each disk, albeit in so far uninterpreted stream files. The stream data contain two items of logical data, the timing of flux transitions and timing of the index. This is just a very low-level interpretation which does not translate at all into human-readable form without further transformation. Each disk took around 4 to 5 minutes to image.

7 ANALYZING THE MAGNETIC FLUX

Diskettes are magnetic storage media on which small areas are more weakly or more strongly magnetized. The magnetic read heads on a disk drive operate by detecting the changes in the magnetic flux. The encoding specifies how the bit is written to the disk surface, and the disk format specifies the byte sequences to represent the structure of the data, such as the cylinder, head number and sector number.

A disk format is the organization of data on the disk which enables the computer system to recognise and verify the data. The bitstream (flux) is separated into addressable parts – sectors – with data marks and address marks in order to tell different types of information within the sector apart and to perform error checking. The binary information is encoded as a pattern of magnetic flux reversals, that is, changes from “0” to “1”.

Recordings are made on a single track at a time. A track consists of a serial sequence of bits which are all interpreted as 8-bit bytes. To ensure that there are changes in magnetisation, an encoding method is used.

The read electronics must be able to get and maintain bit and byte synchronization. Bit synchronization is achieved with a clock bit, i.e., encodings have data bits separated by clock bits.

In MFM encoding, the first pulse period always contains a clock pulse; the second pulse period may or may not contain a data pulse. If the digital data is a “1”, a data pulse will be present in the second pulse period. But, if the digital data is a “0” then there is no pulse present (see Figure 6).

The KryoFlux stream format [16] is recorded in binary, and records only the changes in the magnetic flux of a given track with timing information and cannot reveal any bits directly. The KryoFlux software suite contains stream analyzers for a wide range of different recording encodings for a variety of systems, e.g. FM, MFM and so forth. It was unclear on which system the floppies
were actually written, so all options for different encodings were tried and all failed.

In the meantime, we had identified the drive type as an HP9895A and researched the drive specifications from the resources at the HP Computer Museum [3] and The HP 9845 Project [14]. It was discovered that this disk drive could read and write both Frequency Modulated (FM) and Modified Modified Frequency Modulated (MMFM or M2FM [12]) encodings. The KryoFlux software had not recognized an FM encoding, so we surmised that the encoding was possibly M2FM, with Least Byte First and Least Bit First. While these sites have a wealth of information, documentation of the M2FM format is missing.

Creating Logic Analyzer Software

A small Python script was written to reproduce the recording patterns to be analyzed and visualized in logic analyzer software. A major challenge was to derive the clock signal from the recording as all possible encodings are self-clocked (Figure 6). While the low-level formatting creates quite clean recording patterns the actual data written to sectors often implies tiny distortions (jitter) as it is impossible for the mechanics of the drive to perfectly sync to the clock. This step was really helpful as we were now able to scan along the stream of each track. Our first interpretations were:

- The recording format appeared to be FM (and not M2FM or MFM as we expected to see).
- The low-level format of the system produced clearly visible synchronization patterns at regular distances. One followed by a short block of data, the next one followed by a long block. These patterns were visible for every track.
- This was interpreted as possibly a kind of an information header to the sector that follows, which presumably contains the actual data.
- Counting the zeros – which are easier to count, especially in an empty sector – in the graphical representation we estimated a sector size of 128 Bytes. This contradicts the specifications for the HP drive which has 256-Byte sectors.
- Each information block seemed to contain data (a few non-zeros) which is different for each block. This information seems to be created during the low-level formatting of the disk.
- Some data blocks contain information (significant number of non-zeros).
- If the data block contains information, it was presumably put there later (after the low level formatting). There are tiny distortions visible at the end of the information block. It is nearly impossible to perfectly synchronize with the original recording [13]. This creates tiny distortions which are clearly visible (in the cases evaluated). They are additional indicators for data written.

As some of these results contradicted the expected results, e.g. recording format and sector size, continued analysis was necessary.

Statistical analyses on the KryoFlux streams established the numbers for each length of signal (ground truth). There were visible clusters and only a few outliers (magnitudes less than the number in clusters). The distortions in sector writing are some of them. To reproduce data from the stream requires a correct clock signal, which is challenging as a binary “1” is a changing phase in the middle of a cycle (and then no additional clock signal is inserted); otherwise a clock is reproduced from mandatory phase changes of consecutive binary zeros.

# snippets of Python code for KryoFlux stream analysis

```python
# distance of Data or Clock bits in M2FM multiples of 24
def genBitList() :
    for i in xrange( 300 ) :
        s = '1'
        i-=2
        while i > 24 :
            s += '0'
            i-=2
        i+=16:
            s += '0'
        bitList.append(s)

# extract a byte 'Clock' and 'Data' of a byte
def pDataTakt(header, bits) :
    Clock = 0
    Data = 0
    for i in xrange(8) :
        Clock *= 2
        Data *= 2
        if bits[(i)*2] == '1' : Clock |= 0x1
        if bits[(i)*2+1] == '1' : Data |= 0x1
    Clock = revOrder(T)
    Data = revOrder(D)
    return revBitOrder(Data), revBitOrder(Clock)

# extract a complete sektor from bits
def pDataTaktData( header, bits ) :
    C = []
    T = 0
    D = 0
    c0 = 'c1 = '  
    for i in xrange(0,len(bits),16) :
        C.append(bits[i:i+16])
        Sektor = []
        for i in xrange( len(bits)/2 ) :
            T = 2
            D = 2
            if bits[(i)*2] == '1' : T |= 0x1
            if bits[(i)*2+1] == '1' : D |= 0x1
            if (i&8) == 7 :
                c0 = B(c1)
                c1 = B(chr(revOrder(D)))
                Sektor.append(revOrder(D))
    ...
encoding possibility and put out a call for help on Internet fora and to other universities.

Our next step was to create an actual bit stream for each data sector and read the sector information. The main challenge in this lay in the reassembling of the bitstream into bytes.

Reassembling the Bitstream

At a well-timed moment, we received responses to our call for help which provided us with more information for a decoder and confirmed the recording format as M2FM. With this knowledge, we started decoding the header information (short “sectors” after a synchronisation pattern) from the KryoFlux output.

<table>
<thead>
<tr>
<th>Track</th>
<th>Sector</th>
<th>Track (or cylinder) number</th>
<th>Sector information</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>00</td>
<td>1F</td>
<td>0x70</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>1F</td>
<td>0x70</td>
</tr>
</tbody>
</table>

Figure 7: Extract of bitstream showing identification mark and track number

Sector information is not easily retrieved, as a track recording does not necessarily start with sector “0” and usually contains more than a single rotation of the platter. This information is contained in the short block after a sequence of 32 binary ones. We did a visual count in the representation from the logic analyzer software (see Figure 7). Directly after the pattern follows the identification mark where we found the 0x70 hex address mark (after bit swapping and reading from right to left) and then the track address information.

The track (or cylinder) number 31 (0x1F hex) was displayed in the next byte. Both bytes produced expected results which validated the approach. Having established the proper byte boundaries, the flux stream interpreter could now be completed. The simple structure of the disks simplified the task, as no complex hierarchical filesystem structure had to be taken into account to reassemble the files.

The proof for a properly working script was the reproduction of file names and occupied sectors which could be successfully compared to the sheets available in some of the disk sleeves.

ETTX...<-------- | TEXT...<-------- | FILENAME   |
OPEN.R=/<-------- | POWER.</<-------- | TEXT 68 - 61 |
OR.M...<-------- | ROM...<-------- | POWER 61 - 108 |
RBTAS1.K...<    | BRATIS...K... | ROM 925 - 1115 |
NHFQDL[.w...<   | INFHDL[.w...< | BRATIS 277 - 352 |
HCBKEL[.w...<  | CHKBEL[.w...< | INFHDL 1115 - 1234 |
...              |              | CHKBEL 465 - 557 |

During the different experiments, more facts about this floppy disk format were gathered. The disks contain an unusual number of sectors compared to the typical structure of IBM- and PC-style disks: 30 sectors instead of the expected 26. These sectors are interleaved and track numbers are not consecutive. The flux readings, created by the KryoFlux tools, contain in their track readings approximately five rotations of the platter. This offered the opportunity to compare different readings of the same sector to each other, and help eliminate errors due to inconsistent timing (jitter).

8 CONSOLIDATION

After three months of research, as well as trial and error experimentation, we received confirmation that the machine was an HP9845 with an external floppy drive, the HP9895A disk drive, attached. The disks have 77 tracks or cylinders, and contain 30 sectors with an interleaving factor of 7. The machine was marketed for scientific purposes and was popular in the university till the mid-1980s before the IBM PCs were introduced. It offered a wide range of peripheral connections and general programmed input/output (PIO).

Many of the floppy disks contained in the set appeared to have mostly data of clinical trials on them, judging from the file names and brief descriptions on the disk labels. At this point, we still did not know how these data were organized and if there was text contained within the files. The reason for looking for text files was that text was encoded in ASCII with each character taking one byte of storage, while the numeric data was encoded with different byte sizes dependent on whether they were full-precision numbers, short-precision numbers or integers. One of the disks was labelled “Publications”, so this was selected to test our flux stream decoder. An extract of the result of decoding one track is in Figure 8, which shows the hexadecimal value of each byte, the ASCII character, and the byte-swapped version.

9 IMPLICATIONS

Preservation projects that we and others have undertaken have shown us that legacy hardware is still a significant part of the digital universe [7, 8, 13, 23]. These projects may be in the form of complete systems or (removable) storage media. The last 30 years produced a plethora of magnetic, optical and more recently electronic removable and fixed media including audio and video cassettes, cartridges, floppy disks, ZIP disks, Syquest drives, hard disks, and so forth [11, 19], in a variety of formats. Many of them were deployed as archival media, especially the cheaper ones. While floppy disks and optical media such as CDROM, DVD (or Blu-ray Disks) might be more common, other items including ZIP, Syquest, Jaz, LS-120, SparQ, Orb or MicroDrive might surface in the legacies delivered to special collection libraries or digital arts institutions. More recently, electronic media like CompactFlash cards, memory sticks or multimedia cards will complement the already wide selection.

The drives to handle certain removable media have begun to disappear from the typical second-hand markets, e.g. it is much more difficult to obtain a (confirmed working) 3.25-inch drive than...
it was just five years ago. In short Successful [media] migration diminishes as the age of the medium, or hardware necessary to read it, increases. While most studies focuses [sic] on the longevity of the medium, no doubt fuelled by manufacturing marketing, the true risk lies in the scarcity of hardware necessary to read these formats [10]. Compared to the ease of low-level analysis of 8-inch media, the challenges will increase for most of the recent much more compact and integrated storage peripherals. Some devices became much more proprietary and were manufactured by just one company. These developments increased the number of different standards, but narrowed the amount of different media available for a particular device. Increasing capacities require more advanced technology to squeeze data onto increasingly smaller media.

For many smaller institutions it does not necessarily make sense to maintain a full collection of possible readers and peripherals. For a range of once-popular formats, business models are well established like media transfers from various magnetic tapes. Additionally, there is a selection of transfer peripherals available as listed in the addendum of [13]. Nevertheless, some implementations, especially for USB, might be incomplete or erroneous.1

One strategy, especially for digital art collectors and memory institutions, is to buy spare older equipment for backup or replacement parts, filling their storerooms and taking precious space without the guarantee of usefulness when required [2]. But, as our studies confirmed, there are certain gaps in the service domain challenging memory institutions, such as Freiburg, Archives New Zealand and others seeking to read old 8-inch disks.

**Bridging the Technical Gap**

A digital object is created in a physical and system environment [22]: the former with specific hardware components, the latter with a specific operating system version and utilities. Specialised software, such as digital art and video games, depend on particular hardware for interaction or rendering. Peripherals and controllers need hardware interfaces which become rare as computer technology evolves. While the development of system emulators is seen as a solution to some of these issues, original hardware is needed to prove that an emulator is working properly.

Preservation of born-digital objects necessitates links between past, present and future hardware. A great deal of this can be handled via peripherals. As demonstrated in this 8-inch floppy disk case, challenges are faced at different stages and technical levels. While some of these challenges are unique to each specific type of medium, others follow a pattern, and these will surely repeat for other storage media, such as the following:

- A peripheral to read a particular storage medium must be powered in a specific way. Previously, the voltage and power requirements were much higher than today. 24V and alternating current (AC) connections are now uncommon. In addition, the old types of connectors are no longer manufactured and their documentation of the pin assignment does not typically come with the device.
- To transfer the data to new media, a peripheral or complete computer system is connected to a current system. In many cases, the interfaces do not exist or have changed significantly. An 8-inch drive had a 50-pin connector; for the later 5.25-inch and 3.5-inch drives, these were replaced with 34-pin connectors.
- Parallel connections were superseded by high-speed serial connections as seen in the transition of IDE to SATA.
- The Universal Serial Bus (USB) standardised the connection of commonly used computer peripherals to personal computers, both to supply electric power and to communicate. Thus, it is possible to connect a RS232 serial cable or IDE and SATA cables to USB via an interface adapter. However, the voltage supplied by the USB may not be high enough for older serial connections, and some IDE adaptors do not comply with all aspects of the specification standard making it difficult to properly image e.g. an early laptop IDE hard drive.
- Lack of technical knowledge about connectors, unusual sockets, and plugs cause difficulties: incorrect wiring can destroy well functioning, precious hardware.
- Small embedded components or a System-on-a-Chip (SoC) can replace complete controllers, nonetheless at some point a particular cable or power supply will still be required, as well as the knowledge and expertise on how to connect it.
- Finally, some “glue component” like a firmware or driver might be required to make the peripheral accessible from the operating system.

**Hardware and Documentation Repositories**

The need for preserving hardware, software and documentation has been raised for decades. In 1992, the Commission on Preservation and Access [15] recommended that Computer museum operators can be urged to maintain software as well as hardware, and to be able to operate old programs for purposes of translation. The same call was made by McDonough et al. in 2010 [18] and Rieger et al. in 2015 [21]. As Buskirk [2] put it ... equally urgent is burgeoning obsolescence that leaves orphaned media with no equipment on which to be played or software that is incompatible with current platforms. This aim has not yet been realised.

Dissemination of preservation techniques and workflows is well established for tasks to be undertaken after a usable digital image is acquired, and for popular formats, e.g. 8mm film or VHS to DVD, but little is available for older or rarer artefacts. Much of what is available is in the hands of hobbyists, enthusiasts and digital forensic scientists [16]. Although many hobbyists provide information online, these sites are often not updated and if interest in the topic wanes, the effort needed to maintain the site may outweigh the use for the site owner(s). During this project, the website on which we found the information for the Nashua disk disappeared. The National and State Libraries of Australasia [24] report that inhibitors to data transfer include

- no in-house equipment and equipment is not easy to source
- lacking the appropriate drives
- no internal expertise or staff that have suitable skill set
- no robust workflows for disk imaging

This paper has concentrated on 8-inch disks in M2FM format; data retrieval methods for BTOS/CTOS formatted disks at Archives New Zealand and others seeking to read old 8-inch disks.
Zealand and the University of Freiburg are documented in [23]. Substantial personnel effort and time is invested in setting up the equipment for these projects, as well as financial costs. Often, the equipment is used for a single project, meaning that the return of investment is low, and the acquired knowledge is lost over time. There are many computer museums around the world, however their primary mission is to preserve and display computer systems, not the preservation of born-digital objects. There is a growing need for there to be centres for the transfer of born-digital artefacts to current media which are hubs of expertise and resources, who have a collection of obsolete computers, peripherals, software and manuals, such as the National Library of Australia, for their digital collection, and the Computer Archaeology Laboratory at Flinders University, who undertake research into recovering data from obsolete media.

Technical documentation is a vital resource which is also rapidly disappearing, hardcopy manuals and handbooks are seldomly digitised for online access, especially for devices that are currently perceived by the general public to be out of use.

10 CONCLUSION

Preservation of data from obsolete digital media is not a straightforward exercise. While many floppy disk formats have been studied and have solutions available for them, there are still many which were widely used, that have not yet been deciphered. This project took a lot of patience, a lot of following hints and hunches to track down relevant information, and experimental analysis. We were also aided enormously by the generosity of strangers.

Confirming the disk format and model of the original system took quite a while, until we finally made contact with someone who was involved with the scrapping of the system. Various different levels of forensic science came into play: physical, logical, and higher level information were all needed to identify the system, the drive, the disk format so that we could finally extract some text.

Extracting other data content is much more complicated. We now know that mostly scientific analysis work was carried out on the system, with a large amount of the data being used in databases, spreadsheets and custom-written analysis software. It is difficult to interpret the non-textual data without the context or the software, and we cannot tell if we have the right numeric values, whereas it is easier for ASCII text. We pretty much ended at the point of the 5.25-inch floppy disk recovery study being in need of the original software of the system that data was created with [23]. Additionally, it could be helpful to directly link KryoFlux images to special stream readers built into emulators that interpret the streams directly as disk drive sources, such as in the WinUAE Amiga emulator.

As seen in Figure 8, there are undecipherable characters within the text, these could be word processor formatting codes for specific styles or characteristics of the document, or something else entirely.

Our experiences have reinforced the importance of and need for repositories of knowledge and resources for the initial steps of digital preservation – the creation of usable digital images – which should parallel the sharing of knowledge that has evolved around the handling and future-proofing of artefacts held in galleries, libraries, archives and museums. Especially smaller institutions, lacking the resources for such special-purpose departments in-house, would benefit significantly from these repositories. Such activities must be complemented with modified archival workflows to prevent items falling into limbo between becoming out of use within the institution and being handed over for archiving. Research data management will help to mitigate such challenges in the science domain.

REFERENCES

ACKNOWLEDGMENTS

Many thanks to
Mick Crouch (Archives New Zealand) and Euan Cochrane (Yale University Library) for advice and moral support,
Eric Smith for information about HP devices,
Christian Corti (Computer Center University of Stuttgart) for assistance with analyzing the M2FM encoding,
Konrad Meier (Computer Center University of Freiburg) for providing the oscilloscope and helping with the configuration of the experiment and interpretation of the readout, the Freiburg Emulation-as-a-Service team to provide support and hardware, and finally Jeannette Vollmer to proof read the paper.
ePADD: Computational Analysis Software Enabling Screening, Browsing, and Access for Email Collections

J. Schneider
Stanford University
557 Escondido Mall
Stanford, CA 94305
USA
josh.schneider@stanford.edu

P. Chan
Stanford University
557 Escondido Mall
Stanford, CA 94305
USA
pchan3@stanford.edu

G. Edwards
Stanford University
557 Escondido Mall
Stanford, CA 94305
USA
gedwards@stanford.edu

S. Hangal
Ashoka University
Plot No. 2
Rajiv Gandhi Education City
PIN 131028
India
hangal@ashoka.edu.in

ABSTRACT

ePADD is free and open-source computational analysis software facilitating screening, browsing, and access for historically and culturally significant email collections. The software incorporates techniques from computer science and computational linguistics, including natural language processing, named entity recognition, and other statistical machine learning-associated processes. In this paper, we explain how these processes enable ePADD to support the appraisal, processing, discovery, and delivery of email held by archival repositories and other memory institutions, filling an important role in the preservation of these materials.

CCS CONCEPTS
• Computing Methodologies → Artificial Intelligence; Natural language processing • Computing Methodologies → Machine Learning • Information Systems → World Wide Web; Web applications; Internet communications tools; Email

KEYWORDS
Acquisition, Archival appraisal, Archival processing, Archives, Descriptive metadata, Email, Named entity recognition, Natural language processing, Privacy, Redaction, Screening, Web access

1 ePADD PHASE 2

ePADD Phase 2 began on November 1, 2015 and will end on October 31, 2018. Funded through an US Institute of Museum and Library Services (IMLS) National Leadership Grant for Libraries, Stanford University Libraries, with partners University of Illinois Urbana-Champaign, Harvard University, University of California, Irvine, and Metropolitan New York Library Council, are advancing the formation of a national digital platform by further developing ePADD: free and open-source computational analysis software that allows individuals and institutions to appraise, process, and provide access to email of potential historical or cultural value [2]. During this grant period, Stanford University Libraries and grant partners will continue to improve the program’s scalability, usability, and feature set [1, 8].

2 RESEARCH VALUE OF EMAIL

Email offers singular insight into and evidence of a person’s self-expression, as well as records of transactions, collaborations, and networks [7, 9, 10]. Email communications of prominent individuals, including politicians, writers, scientists, and scholars, reveal their professional and personal actions, decisions, and creative output, as well as their relationships within society and communities [6]. The appeal of email collections therefore extends beyond historians to all manner of researchers, journalists, and the general public seeking to obtain insight into individuals and their transactions.

3 ACCESS CHALLENGES FOR EMAIL

3.1 Screening Email

A major challenge that many institutions face when trying to provide access to born-digital collection materials is the need to ensure that creator and third-party privacy, and copyright, are protected [4]. Email archives can include hundreds of thousands or even millions of messages; the challenge of screening email
collections is compounded when considered at this scale. As manual review is prohibitively time-consuming, institutions require a mechanism to help automate the process of screening for sensitive, confidential, and legally protected information.

### 3.2 Providing Access to Email’s Intellectual Content

Traditional email browsers typically provide only limited search capabilities, and do not permit browsing of correspondents or named persons or organizations within the email archives. They also do not allow browsing of image attachments or support other types of visualization. These limitations make a traditional browser an imperfect tool for staff of archival repositories and other memory institutions wishing to review or describe email archives, and for researchers who wish to access and study them.

### 3.3 Supporting Discovery of Email Collections

Current tools to support archival description and discovery for email are limited in their ability to convey the intellectual contents of the archive to a researcher. Email in archival collections has traditionally been promoted and made discoverable online through scant description in library catalog records and archival finding aids, that provide little detail to assist researchers in learning the identity of the principal correspondents, or the named entities discussed.

### 4 ePADD SYSTEM ARCHITECTURE

#### 4.1 Appraisal

**Appraisal** provides donors, curators, and archivists with a toolset to review and manage an email archive prior to accessioning it to a repository. ePADD can gather email from multiple sources, including mail stored in MBOX format or transferred by IMAP connection. Upon ingest, ePADD de-duplicates messages, resolves correspondent names from the address book, and extracts fine-grained entities using a custom NLP toolkit. These functionalities and others enable users to determine the relevance and importance of email messages, identify and flag confidential, restricted, or legally-protected information, and impose access restrictions prior to transfer.

#### 4.2 Processing

**Processing** is designed for an archivist to further perform all functions included in the Appraisal module, including scanning for confidential, restricted, or legally-protected information, as well as other tasks that prepare the archive for discovery by and delivery to end users, such as reconciliation of correspondents and extracted entities with established authority records (see Fig. 1).

#### 4.3 Discovery

**Discovery** is designed to run under a standalone web server, and allows researchers to browse and search a redacted email collection prior to physically traveling to a repository’s reading room to access the full corpus [5]. Only metadata from the processed email archive is published online.

#### 4.4 Delivery

**Delivery** provides users with access to the full contents of the unrestricted portions of a processed email archive, including attachments, from a managed workstation in a repository’s reading room.

### 5 FUNCTIONALITIES

#### 5.1 Named Entity Resolution / Fine-Grained Entity Type Browsing

ePADD uses a custom fine-grained named entity recognizer/classifier that recognizes categories of entities bootstrapped from DBpedia. These include persons, organizations, locations, government entities, political parties, companies, universities, diseases, and awards. ePADD learns from these categories and is also able to recognize likely entities it has not come across before.

#### 5.2 Name Resolution / Correspondent Browsing

ePADD resolves names and email addresses associated with a single correspondent, improving browsing and visualization. All decisions can be manually overridden using a dedicated interface. Mailing lists can similarly be tagged and optionally consolidated using this functionality. Resolved correspondent names can be browsed and graphed alphabetically or by volume of messages exchanged with the email account holder.
5.3 Lexicon Search

ePADD includes tiered thematic keyword searches geared towards broad analysis of a variety of email collections, including lexicons to identify categories of sensitive correspondence. These lexicons can be edited and tuned, or the user can create all new lexicons to suit their research goals.

5.4 Advanced Search

ePADD includes an advanced search interface enabling sophisticated search queries. For instance, users can perform a search for messages containing entities from the disease entity category, or terms from the sensitive lexicon, and further limit this search by mandating that the search should exclude results from a mailing list. In this way a user can create a narrow search for potentially sensitive information to embargo for a specific period of time or to not transfer to a repository.

5.5 Query Generator

ePADD includes a query generator to aid in comparative entity analysis between the archive and any other textual corpus. Matching entities are highlighted and link to message results.

5.6 Redacted View of Messages

ePADD provides an optional Discovery module, intended to provide improved discovery of the archive online via a public web server. This module redacts all content other than message dates, correspondents (local-part of email address), and named entities, in order to protect creator and third-person privacy and copyright (see Fig. 2).

5.7 Bulk Actions and Annotation

ePADD allows the user to apply actions (including marking messages as reviewed, fit for transfer, or fit for embargo) and annotations to sets of messages meeting user-defined criteria, including all messages associated with a given correspondent, all messages from a given date range, all messages containing certain keywords or named entities in the subject or message fields, or some combination of the above.

5.8 Additional Functionalities

ePADD’s additional functionality includes features intended to further support screening for sensitive, confidential, or legally protected materials, as well as features intended to support user access to the intellectual content of the messages. These functionalities include: regular expression search, account and folder-level browsing, built-in visualization tools, and image attachment browsing (see Fig. 3).

6 FUTURE DEVELOPMENT

Future development during this grant period includes advancing ePADD’s support for restriction and derestriction of materials; continuing to optimize ePADD performance at scale; and development of cross-collection search and browsing. We also plan to develop support for exporting message headers as GraphML for network visualization, and correspondents and fine-grained named entities as linked open data. We will also continue to promote ePADD’s ability to support diverse community workflows and institutional requirements.

Long term, we hope to explore development of cross-institution discovery capabilities and the creation of a web service to federate search and browsing across all content that has been processed through ePADD worldwide, in order to streamline the discovery of this content for research.

Figure 2: Redacted message view in the Harrison Studio papers - Email Series, Stanford University, ePADD Discovery module, 2017 (Version 3.0).

Figure 3: Image attachment browsing in the Harrison Studio papers - Email Series, Stanford University, ePADD Processing module, 2017 (Version 3.0).
ACKNOWLEDGMENTS

ePADD development is managed by Stanford University’s Department of Special Collections & University Archives, part of Stanford University Libraries [3]. The ePADD development team is composed of Glynn Edwards, Peter Chan, Josh Schneider, and Sudheendra Hangal. Development work is carried out in collaboration with partners at Harvard University, the Metropolitan New York Library Council (METRO), University of Illinois at Urbana-Champaign, and University of California, Irvine. Funding for current ePADD development is provided through an Institute of Museum & Library Studies (IMLS) National Leadership Grant (NLG) for Libraries. Development for the initial 2015 release of ePADD was primarily funded by the National Historical Publications and Records Commission (NHPRC).

REFERENCES

Challenges in Preserving Augmented Reality Games: 
A Case Study of Ingress and Pokémon GO

Jin Ha Lee
University of Washington
Mary Gates Hall, Suite 370
Seattle, WA 98195
jinhalee@uw.edu

Stephen Keating
University of Washington
Mary Gates Hall, Suite 370
Seattle, WA 98195
keatings@uw.edu

Travis Windleharth
University of Washington
Mary Gates Hall, Suite 370
Seattle, WA 98195
travisw@uw.edu

ABSTRACT
Video games, as an evolving medium, present specific challenges to preservation, and with the introduction of newer technologies, these difficulties are ever increasing. Augmented Reality Games (ARGs) present challenges unique to the technologies implemented within them, blending the physical world with the virtual world, the socializing factors embedded in play, and the increasingly network-distributed nature of their content. We examine two popular ARGs, Ingress and Pokémon GO, as case studies and illustrate the additional challenges ARGs bring to preserving video games. Boundaries in ARGs question the traditional understanding of gameplay, while preserving changes in virtual and physical space present challenges to ARG representativeness and authenticity. Metaplay illustrates behavior based on non-deterministic situations, often as a result of real world changes. We also review different approaches suggested in prior literature for preserving video games, focusing on preserving the context and history of the game by collecting relevant artifacts and documentation, and explain why implementing these approaches could be challenging for ARGs.

KEYWORDS
Augmented reality games, Video games, Ingress, Pokémon GO, Digital preservation, Preservation, ARG

1 INTRODUCTION
Video games are a complex and multifaceted medium which present many organizational and preservation challenges. The pressing need for preserving digital games was articulated in a white paper published by the Game Preservation Special Interest Group within the International Game Developers Association -- digital games are an important part of our cultural heritage and a failure to take action to preserve them now may result in a significant loss of content and the contexts in which they are presented, within but a few short decades [9]. However, preserving video games is not a simple task, as noted in prior literature. For instance, the final report from the “Preserving Virtual Worlds” project led by McDonough et al. [8] presents a number of issues pertaining to preserving games and virtual worlds, such as the obsolescence of game hardware/software, scarcity of media, third party dependencies, complex and proprietary code, difficulty in verifying object authenticity, intellectual property rights issues, problems in determining significant properties, and the importance of preserving context. Additionally, McDonough [7] and Lee, Clarke, and Perti [5] also point out limitations of existing game metadata and difficulties in collecting useful metadata due to the nature of games and their unique publication process. Multiple versions, relationships with other non-game objects, modifications and additional content, and complicated intellectual property situations are but a handful of issues important to video game preservation. Barwick, Deamley, and Muir [3] also examine the same problem from the practitioners’ perspectives by interviewing representatives from three different museums/archives (Computerspiele Museum, Strong National Museum of Play, and The National Videogame Archive) with extensive video game collections, examining the current status of digital game preservation. They also note the legal challenges in the long-term preservation of games as well as the difficulties associated with presenting games to the museum audience, and preserving the cultural significance of the game and the game’s relation to its cultural context.

In addition to the multitudes of challenges of preserving digital games, McDonough et al. [8] raise an important question in preserving specific types of virtual worlds or games such as Second Life or Massively Multiplayer Online Role-Playing Games (MMORPGs) like World of Warcraft. In these environments, “the value and meaning of a virtual world is primarily derived from the actions and interactions of its players (p. 29)” and experiencing the world outside of their prime time would mean that the player’s interactivity would be more akin to an archaeological study, attempting to envision how the virtual world might have looked. Winget [12] elaborates on the challenges of preserving MMORPGs, including technical dependencies, representation, and collection development. She argues that in addition to solving technical and legal issues, other challenges such as determining the nature of primary and secondary artifacts, modeling interactivity, and collecting contextual information, need to be considered as well.

In our work, we aim to expand the discussion on video game preservation by examining Augmented Reality Games (ARGs). ARGs are generally understood as games that blend the virtual world and real world via gameplay, although some scholars argue for a more precise definition. For instance, Azuma [1] specifies three characteristics that defines ARGs: “1) combines real and virtual, 2) interactive in real time, and 3) registered in 3-D” (p.2). Generally, the term ARG would include a game like Pokémon
GO, which may or may not meet the third criterion depending on the use of the AR feature in the game. In either case, mapping physical space virtually is an integral piece of defining the ARG experience and creates unique preservation challenges due to the demands of location-based play.

The unprecedented success of Pokémon GO in 2016 has resulted in an increased interest in ARGs, including academic perspectives examining the benefits and drawbacks of these games, impacts on real-life player behaviors, emergent issues regarding safety and privacy, as well as interest from game developers wanting to understand how Pokémon GO was able to captivate such a massive and diverse audience. Preserving ARGs will undoubtedly be an important task for conveying the cultural heritage of paradigmatic play (i.e., play specific to a time and place) and how such games have impacted our current society. ARGs amplify existing challenges in preserving digital games by physically mapping context-sensitive and social events in the real world. Here, we specifically discuss two examples of ARGs as our case studies to illustrate some of these challenges. We will conclude with a discussion of the impact of ARGs on various game preservation approaches.

2 Two Example ARGs

2.1 Ingress

Ingress was developed by Niantic, a spinoff company from Google, and was released in 2014. In 2015, Niantic reported that they had over 7 million players. The game uses Google Maps as a base for overlaying game elements onto real-world locations. In Ingress, there are two factions (Enlightened vs. Resistance) that players can choose from. The main goal of the game is to take control of “portals” which are mapped to various real-world locations such as landmarks, public places like parks, or local businesses. Capturing these portals is done by placing a resonator at a portal which marks it based on team affiliation: green for Enlightened and blue for Resistance. If players come across portals already controlled by the opposite team, they can remove opposition control by using in-game weapons like bursters or ultra-straikes. Players aim to take control of as many portals as possible and control an area by deploying “fields” connecting three portals into a triangle. These game control actions can happen at varying scales; connecting portals highly concentrated in a small local area is called “microfielding”, while a large-scale operation which can cover multiple states or even countries is called a “megafied”. The larger fielding endeavors require many participants and tight coordination, and are referred to as “field operations” (“field ops”). Field ops typically involve anywhere from a dozen to hundreds of players engaging with the game in real time. There is an elaborate out-of-game narrative that establishes relationships between player actions and story events. Niantic also hosts global community events called “anomalies” where player teams compete to take control of portals in a local area, and “shard events”, where players must coordinate the capture and movement of in-game items called “shards” to certain target portals to “score” them.

![Figure 1. Screenshot of Ingress (left) and Pokémon GO (right)](image)

2.2 Pokémon GO

Pokémon GO is undoubtedly the most popular and well-known ARG of all time. Niantic recently announced that the game has been downloaded over 650 million times and has made over one billion dollars in revenue. In Pokémon GO, the players’ objective is to locate and catch virtual creatures called Pokémon that appear in real-world locations. Pokémon was already an extremely popular media franchise from Japan and has multi-generational appeal to users across the world. The overarching goal of the Pokémon franchise is to “catch’em all”; collect all the Pokémon, train, and battle with them. In the game Pokémon GO, Pokémon randomly spawn in different real-world locations, as well as at locations called “nests” where a substantial number of a specific kind of Pokémon spawn. Players can also use incense and lures to attract Pokémon to a particular location. Incense can be used by individual players to lure Pokémon to the current location of the player. Lures are used at PokéStops to attract Pokémon to that location and benefit multiple players who are close by. PokéStops are mapped onto real-world locations just as in Ingress. In fact, Niantic used some of the portals in Ingress to populate PokéStops in Pokémon GO. Players can receive resources like balls (to capture Pokémon) and berries (to help improve some aspects of the catch experience, such as increasing catch rate or restricting Pokémon movement) by interacting with PokéStops. There are three different teams (i.e., Instinct, Mystic, and Valor) and players can battle with players from other teams at Gyms using their captured Pokémon in order to take control of them. Gyms, just like PokéStops, represent particular locations in the real world.

---


3 Issues and Challenges in Preserving ARGs

In this section, several issues and challenges for preserving ARGs focusing on the aforementioned games will be presented, including 1) determining the boundaries of the game in ARGs and dealing with subsequent preservation scalability issues, 2) recording changes in physical and virtual spaces while preserving a game’s representativeness and authenticity, and 3) preserving meta-play, social gameplay, and interactive contexts.

3.1 Boundaries of the Game and Scalability

One of the foremost challenges in preserving ARGs is establishing the boundary of the game object. Does the reflection of real world elements in gameplay warrant preserving real world data not represented in the virtual world? Does this mean, for games like Ingress or Pokémon GO, that we need to save the Google Maps data used to create them? One may argue that preserving the object itself, in this case, the digital game, should remain the primary goal in game preservation. However, according to Niantic has made it clear that real-world interactions of players in Pokémon GO are the foundation of their games: as Niantic CEO John Hanke says, “From the very beginning, our games were about encouraging people to go outside and see interesting places.” If this is indeed the case, completely disregarding the representation of real-world data in ARG preservation does not seem ideal, as it may be impossible to replicate intended or meaningful player experience, at least as it is described by the developers. However, even if we were to accept that real-world data should be part of the object we preserve, it is unclear how much and what kind of data we would be preserving. Is the current version of Google Maps data enough? What kinds of legal challenges exist in obtaining and preserving such data? Does it need to include a visual representation or description of each of the real-world locations beyond what is provided by the game itself? How will this information be obtained for all the real-world locations across the globe? Additionally, how would we go about replicating gameplay that involves real-world data and movement between locations (e.g., walking to hatch eggs in Pokémon GO; linking portals or creating a field in Ingress) in ARGs to recreate a similar experience for players in a museum setting?

As Winget [12], Newman [11], and Kaltman [4] argue, the scalability of data preservation for video games gets more difficult as more content distribution becomes network-based. ARGs are entirely network-based and their particular scale is often determined by their players. As more players are added to a game such as Pokémon GO, the scale of data becomes increasingly dense and as a result, difficult to meaningfully track, describe, or emulate. Based on recent estimates from Android Central and Business Insider, an average Pokémon GO play session uses about 20MB of data. For a single user, that figure may seem rather benign, but if we scale that up to the actual Pokémon GO user base of roughly 20 million active daily users of the game, that is roughly 400 terabytes of data per day. The amount of information necessary to account for the representative game experience of just an average day for Pokémon GO is extraordinary, and these are the information needs for most network-distributed games, including ARGs. The result of such needs is a product that is likely impossible to preserve in its entirety at scale. Rethinking what or how to preserve becomes necessary for ARGs without a significant revolution in the quality, quantity and availability of data storage. Pokémon GO is but the start of what is likely to be a larger trend, where franchises attract large audiences and players create large amounts of data, making it increasingly difficult to track pertinent and meaningful information about these games.

3.2 Changes in Physical and Virtual Spaces

ARGs are played in the real world, with specific geospatial points in the real world corresponding to an analogous location in the virtual world. As mentioned in Section 2, portals in Ingress and PokéStops/Gyms in Pokémon GO correspond to real world locations such as monuments, public art, and businesses. However, changes often occur in the real world (e.g., construction removing a business) or the virtual game world (e.g., a new version or update removes PokéStops), that do not get updated accurately in its counterpart.

For instance, Pokémon GO or Ingress players looking to orient themselves to a particular place in a virtual environment may find that while the location exists in the game, the actual object associated with it has since disappeared (e.g., mural, natural object, business). On the University of Washington campus, players may look for the Washington Elm tree reflected as a portal in Ingress or PokéStop in Pokémon GO, but will not be able to find it in the real world, as the tree has been removed from campus. Similarly, a quarter mile away a chocolatier is still shown as a portal in Ingress, yet the business is now closed. These virtually or physically missing elements leave a legacy of sorts, marking a space’s historical past. As ARGs age, elements within them intended to reflect places and objects in the real world disappear as the physical or virtual game environment change.

Such environmental mismatch could certainly become an issue as gameplay is affected by changes in an ARG’s physical and virtual game environment. Furthermore, descriptive information about the game, as a result, will make less sense with the different versions of the game due to discrepancies in physical and virtual game environments. For example, if someone wants to experience an aspect of Pokémon GO, such as PokéStops that were particularly contentious, they need a version of the game from before these PokéStops were removed due to requests from individuals or organizations owning the associated locations. However, much of the external information about the game, such as popular media coverage, or user-generated discussion, tend to not specify which version of the game they are referring to, sometimes making it nearly impossible to track down a specific

---


version containing relevant location information. In ARGs, version control is uniquely challenging due to the necessity of keeping track of multiple environments and their relationships.

3.3 Metaplay, Social Gameplay, and Context

To fully understand the dynamics and gameplay of ARGs, it is necessary to understand how people interact with one another in both physical and virtual environments. The major ARGs that have been released share several challenges with that of preserving other virtual worlds, such as those in MMORPGs. These include preserving the players as an intrinsic part of a game’s historical context, and recording the changes in MMOs and their players over time [2].

ARGs present several additional challenges to preserving gameplay and dynamics, including gameplay taking place out in the real world with players competing, cooperating, and reacting to one another in real space. As a result, there is a strong and unique social dynamic among players of the game that takes place in the real world. In MMORPGs, one might argue that it is possible to at least preserve part of the non-social gameplay, such as PvP (Player vs Player) actions that can be played by a solo player. In games like Ingress, the kind of game interactions one can have is severely limited if you do not account for social play (play with other players). PvP (Player vs Player) actions are integral to the gameplay. Once you take control of the portals around you, there is essentially no game unless players from other teams come and take them down. Even in Pokémon GO, where social action may be perceived as less important than Ingress, luring PokéStops to attract players is a common behavior.

Also, while MMORPGs require players to play together to fully experience the game, and various communities emerge and social gatherings happen, because some ARGs require players to interact with people in close proximity in the real world, a number of additional issues emerge. In the case of Ingress, because it is inherently a social game that forces players to interact with other people in real-life, a large amount of information is shared among players regarding who they should work with, based on their personalities, mobility, and level of commitment to the game. Players who have shown dangerous behavior (e.g., stalking, physical confrontation) are actively watched and avoided via third-party tools that allow people to track the location information of all players. High-profile players from the opposing team are identified and any of their unusual movement is reported back to the community to alert the possibility of a field operation being deployed. None of this information about the players is represented anywhere within the game itself. Furthermore, much of the gameplay regarding Ingress occurs outside of it. Examples include expanding and maintaining local player networks, collecting intel about players on the other team, mapping routes for a mega-field operation, and so on. These types of metaplay are not something recorded anywhere within the game itself, yet remain a critical component of the play. Ultimately, the problem is that interactions happening within a digital game only make up part of the overall play experience of an ARG.

It is also difficult to preserve temporal factors in the physical world that impact the game state. The nature of ARGs present unique challenges in preserving real world factors that impact and alter gameplay, particularly as it pertains to location access in the game world. For instance, in a traditional online game, access to in-game areas is dependent upon the developers changing the game code and in-game environments, or in some instances, player behavior. With augmented reality games, any number of factors in the real world can change access to physical areas, and by extension the game state, even though those real-world changes are not reflected in the game’s world. For instance, winter snows make ARG play in some areas seasonal, and the timing and duration of play is dependent on the weather, not player behavior or the game. Additionally, some public spaces once available for play can become blocked by construction, road closures, new rules and regulations, or other access changes. Some players utilize such tactics to their advantage. For example, Ingress players will make strategic use of portals in amusement parks that close for a season, or claim portals high in the mountains just before snow storms hit. Some Pokémon GO players report locating and claiming gyms that are in less accessible areas in the real world so they can have access to the secure, steady supply of resources owning an inaccessible gym provides.

Understanding such temporal conditions builds the foundation allowing for strategic play, and establishes a form of competitive advantage conferred by knowledge of both the game and how it is affected by temporal conditions, such as the seasons. Yet the question remains as to what the best way might be to preserve the temporal, real-world, events frequently occurring in ARGs.

4 How Do We Preserve ARGs?

In prior literature on video game preservation, varying solutions have been posited by scholars, responding to the numerous challenges that exist for preserving the game and gameplay experience. Lowood [6] calls for initiatives to build game performance archives containing documentation of gameplay and relevant artifacts of participatory culture as well as archives of artifacts and documentation representing the history of game design. Murphy [10], also noting the challenges in preserving MMORPGs, suggests several approaches for better preserving the context of games, such as ethnographic writing and video documentation of gameplay, as well as preserving crowdsourced contributions from game players. While these approaches indeed offer a better model of preserving games in general, for ARGs, we also anticipate additional challenges in implementing them.

Ethnographic documentation has specific limitations due to the amount of information volunteered. In the case of Ingress, certain information is withheld due to the value and impact it can have while playing the game. High-value information is often protected with the same confidence as company secrets, whereby competitive advantage is lost if certain information becomes a matter of public, distributed record. For instance, information such as strategies for successfully running a mega-field operation in Ingress are often intentionally not recorded and instead only
orally disseminated to players. Additionally, such high-value information would also require deep integration of an ethnographer into the game’s social context in order to escape the problematic scenario of being seen as an outsider and thus not being privy to such scenarios.

Video documentation of gameplay would be able to sufficiently record the gameplay itself, but again in the case of ARGs, much of the play experience is outside the game itself. The planning of social events and strategic contexts in relation to the game, or metaplay, often makes up a great deal of ARGs, particularly competitive games like Ingress. The result is that even with video documentation of the gameplay, it would not provide much context for what makes competition within the game compelling or meaningful to many of the players.

Preserving crowdsourced contributions can also be a challenging task, due to the large amounts of misinformation, disinformation, trolling, and fake news based on speculation or false information⁶. These stories often happen as a result of a game’s sudden popularity, and nowhere has this been truer than in Pokémon GO. In Pokémon GO, there are realms of intentionally fabricated information online about what is possible or not possible within the game, such as whether or not certain Pokémon are in the game (e.g., Ditto), or whether it is possible to hatch region-specific Pokémon, or acquire Gen 2 Starter Pokémon. Additionally, there are numerous stories recounting different players undergoing physical harm, others making money off the game, and many varying pieces of difficult-to-verify information about the game, much of which is also related to acquiring a competitive advantage within the game.

Historical account generation is also difficult due to the fact that many ARGs take on a shared narrative, whereby the game changes based on player interaction with the game world. As a result, for a game like Ingress, whose narrative is constructed based on the competition ongoing within the game, an official story is being continuously generated but does not definitively exist until after certain events occur. As an example, an anomaly in Ingress often has the two teams in the game compete over a certain area, and the story branches based on how players perform in that event. This means even though it is known that the game developers often account for multiple different outcomes, the only official story is the one that gets told after the event. The developer often does not intend to act as a moderator of the game world and as a result it is often up to the players to self-report their own historical accounts, thus making an authoritative historical account difficult. Player bases change and shift over time, and with that fluctuating player base is an evolving story, making it particularly difficult to pin down an accounting of events.

As ARGs are still a relatively new phenomenon, many of the questions we asked here have yet to be answered. However, it is also the case that early ARGs, such as Parallel Kingdom, have already been shut down with information about the dynamics and culture of the game only remaining in blog posts or forum discussions, demonstrating that the loss of ARGs is already occurring. Perhaps what first needs to happen is to better understand what the “significant properties” are when it comes to ARGs. McDonough et al. [8] explains the challenge of trying to preserve games without knowing what portion of them will be considered significant by future scholars. This does seem to be the case for ARGs we examined; ultimately, what makes Ingress, Ingress? What does it mean for players to play Pokémon GO? Is it the discovery of real-world locations, some social aspect of play, a renewed concern for personal health, the metaplay, or something else? We have yet to understand what playing ARGs means to current users, let alone future scholars. Our future research will involve a detailed investigation of player behaviors in ARGs, in an effort to elicit user attitudes toward ARG gameplay, and to garner insight into better preservation strategies for these games.

REFERENCES


veraPDF: open source PDF/A validation through pragmatic partnership

R. McGuinness  
Open Preservation Foundation  
c/o The British Library, Boston Spa, Wetherby, United Kingdom  
becky@openpreservation.org

C. Wilson  
Open Preservation Foundation  
c/o The British Library, Boston Spa, Wetherby, United Kingdom  
carl@openpreservation.org

D. Johnson  
Association for Digital Document Standards e.V.  
PDF Association, Neue Kantstrasse 14  
14057 Berlin, Germany  
duff.johnson@pdfa.org

B. Doubrov  
Dual Lab spr.  
Clos du Parnasse 12C, Brussels, Belgium  
boris.doubrov@duallab.com

ABSTRACT
veraPDF is an open source industry-supported PDF/A conformance checker developed by the members of the veraPDF consortium. The software validates all current parts and conformance levels of ISO 19005 (PDF/A[1]). This paper makes the case for open source format validation and describes the project’s approach to building and testing the software. It also explores how the unique partnership between cultural heritage organisations and PDF industry has created an active open source community.

KEYWORDS
Validation, conformance checker, open source, digital preservation, file formats, PDF/A, standards

1 THE PREFORMA CHALLENGE
veraPDF is one of three conformance checkers developed with funding from the PREFORMA (PREervation FORMAts for culture information/e-archives) project. PREFORMA’s aim is to empower cultural heritage institutions to gain control over the technical properties of preservation files. To achieve this they issued a call for tender [2], alongside a challenge brief [3] for the development of open source conformance checkers for document (PDF/A), audio-visual (Matroska, LPCM, FFV1), and image formats (TIFF). In addition to the delivering software, the three ‘suppliers’ i.e. the company or consortium selected to carry out the work, were directed to “establish a healthy ecosystem around an open source ‘reference’ implementation of specific file formats [3]”.

1 http://verapdf.org/  
2 http://www.preforma-project.eu/  
3 http://verapdf.org/

veraPDF comprises four components:

- An implementation checker, which validates all parts and conformance levels of the PDF/A specifications;
- A policy checker, which allows users to implement additional custom checks to enforce institutional policy with respect to the format;
- A reporter, which processes the results, producing reports, both human readable and machine parsable; and
- A metadata fixer, which repairs metadata in files based on conformance with the standard.

PREFORMA also requested the development of an integrating shell capable of controlling all the PREFORMA-project conformance checkers, and potentially others that implement the same architecture and APIs.

The schedule for the conformance checker suppliers was divided into three phases:

1. A four month competitive design phase where suppliers were funded to submit their designs. At this early stage PREFORMA selected two potential suppliers for each format. PREFORMA evaluated the six proposals and chose one supplier for each of the conformance checkers to develop prototype software.

2. A twenty month prototyping phase starting in April 2015. Only the three selected suppliers progressed to this phase, which also prescribed a two month design review process that took place at the end of 2015. Following a year of development, the final version 1.0 prototypes were released in December 2016.

3. A six month testing phase that is ongoing at the time of writing. During this phase PREFORMA carries out acceptance testing of the delivered prototypes.
2.1 Open Source for Sustainability

Adoption of open source software for digital preservation in cultural heritage organisations is high. A survey \(^3\) conducted by the OPF (Open Preservation Foundation) \(^4\) in 2015 showed that 88% of responding institutions used open source software. Cultural heritage institutions are increasingly both using and contributing to the development of open source tools \[^{[4]}\].

PREFORMA did not aim to just develop software, they are just as concerned with establishing active open source communities around the conformance checker projects. The project required suppliers to maintain an up-to-date, public source code repository, and upload monthly software releases to PREFORMA’s open source portal \(^5\).

PREFORMA’s intention was to ensure that the source code, software and documentation remained available beyond the lifetime of the project. This strategy aligns with the OPF’s approach to sustaining open source digital preservation software as described in the OPF software maturity model \(^{12}\). The OPF has experience of both sustaining open source projects including PLANETS \(^{7}\) and SCAPE \(^{8}\), as well as providing stewardship for standalone tools such as JHOVE \(^{9}\).

Building on its foundations in both the archival and PDF software industry communities, the veraPDF consortium has successfully established a dialog between the PDF industry and cultural heritage organisations through digital preservation and industry events and webinars, an online and social media presence, and the open development approach.

2.1.1 Licensing. PREFORMA mandated that each conformance checker be dual licensed under the GNU General Public License version 2.0 \(^{10}\) or later and the Mozilla Public License version 2.0 \(^{11}\) or later. All other project outputs such as test data sets and documentation were to be licensed under CC-BY-4.0. \(^{12}\) The veraPDF test corpus is licensed and is freely reusable for testing PDF/A validators.

3 SELECTING FORMATS FOR VALIDATION

Identifying file formats and determining the extent to which individual files conform to appropriate specifications is an essential step in many digital preservation workflows. It demands both detailed understanding of a file’s technical properties, and the format specification. However, cultural heritage organisations do not have the resources to employ in-house, specialist knowledge for every file format which they are obliged to preserve.

According to Artefactual Systems \(^{13}\), developers of open source digital preservation system Archivematica \(^{14}\), “one of the most influential factors in selecting preservation formats is community adoption \(^{15}\).” PDF/A was one of the file formats selected by PREFORMA because the specification is:

- Complete: final definitive documentation is available;
- Open and accessible: available to anyone, free of charge, or for a one-off nominal fee, so it can be reused without any limitations; and
- Widely used by cultural heritage institutions and national archives \(^{16}\).

While there are commercial PDF/A validators available no complete or authoritative open source implementations existed.

4 PARTNERING WITH INDUSTRY

The veraPDF consortium brings together a network of stakeholders with complementary perspectives. Led by the OPF and the PDF Association \(^{17}\), the industry trade group for developers of PDF software, with partners Dual Lab \(^{18}\), DPC \(^{19}\) and KEEP SOLUTIONS \(^{20}\), the veraPDF consortium combines digital preservation and cultural heritage expertise with document industry backing and comprehensive access to the PDF/A standardization process.

4.1 The veraPDF Approach

In order to fully understand the format specification, the veraPDF consortium decided to create a test corpus covering all parts and conformance levels of the PDF/A standards. The idea was that creating test files would highlight any lack of understanding on the veraPDF consortium’s part or reveal ambiguities in the standards.

When there is a problem interpreting the specification the veraPDF consortium brings the issue to the PDF Association’s PDF Validation Technical Working Group (TWG), comprised of PDF technology experts. It was established to analyse PDF validation issues in a transparent fashion and also connect veraPDF to the ISO committee responsible for PDF/A.

The ISO working group responsible for PDF/A (ISO TC171 SC2 WG5) \(^{21}\) had previously decided that the existing ISO specifications for archival PDF (PDF/A-1, PDF/A-2 and PDF/A-3) would not be revised, even in the light of ambiguities

---

\(^{13}\) https://www.artefactual.com/

\(^{14}\) https://www.archivematica.org/en/

\(^{15}\) https://blogs.loc.gov/thesignal/2012/10/archivematica-and-the-open-source-mindset-for-digital-preservation-systems/


\(^{17}\) http://www.pdfa.org

\(^{18}\) http://www.duallab.com/

\(^{19}\) http://www.dpconline.org/

\(^{20}\) https://www.keep.pt/en

\(^{21}\) https://www.iso.org/committee/337930.html
uncovered by the veraPDF consortium’s project. This is understandable as changing a specification undermines any software developed against the previous version. Instead the PDF Validation TWG asked WG5 to review the ambiguities uncovered during the development of veraPDF in the context of the development of a PDF Association Technical Note. PDF Association Technical Notes have a good track-record of adoption by the industry. The development of veraPDF has also directly influenced the standardisation process, with several issues raised leading to enhancements in a forthcoming new part for PDF/A. As such, the PDF Validation TWG was able to influence the development of next-generation PDF/A specifications.

veraPDF has thus served as a vehicle enabling cultural heritage organisations to directly influence the standardisation process, benefitting both the digital preservation community and industry.

5 ESTABLISHING GROUND TRUTH

The consortium’s aim when creating test data was to produce a comprehensive ground truth [5] corpus for the PDF/A standards. veraPDF’s developers carefully examined each clause in the standards, and developed a formal grammar to describe the requirements in a machine-readable fashion. They then produced validation rules with an accompanying programmatic test for each requirement. PASS and FAIL corpus files were created to test the validator’s functionality. Where unable to create a rule, file or test, or when 3rd party validators were observed to disagree, the issue was raised with the TWG. The process is illustrated in Figure 1.

![Figure 1: veraPDF test corpus development process](image)

The veraPDF test corpus comprises over 1,500 PDF files. All veraPDF development and release software is tested against the corpus and the results are published online. The veraPDF developers also test against the existing PDF/A test corpora produced by the PDF Association:

- Isartor PDF/A-1b test suite
- The BFO PDF/A-2 test suite

In addition to resolving problems interpreting the standards, the PDF Validation TWG reviewed and approved the final set of validation rules and test files.

OPF and DPC also performed real world testing with their respective members. Focusing on reliability, performance and usability rather than validator functionality, this testing provided valuable feedback for bug fixes and optimisations.

KEEP SOLUTIONS carried out software reviews and tested external integrations by incorporating veraPDF into RODA, their digital repository solution.

6 VERAPDF TODAY AND FUTURE PLANS

At the time of writing the veraPDF software is available as version 1.6. Two distinct implementations are available. Early prototypes of veraPDF used Apache PDFBox as its PDF parser, and to implement the veraPDF validation model. Although the veraPDF consortium was aware this was incompatible with PREFORMA’s licensing requirements the consortium needed a parser, and obtained dispensation to use PDFBox to test its rule-based approach to validation.

Version 0.26 marked the first dual release in which veraPDF launched its own, purpose-built PDF parser, (the “greenfield” parser), alongside the PDFBox implementation. The greenfield version fully meets the dual licensing requirements with minimal external dependencies.

As well as enabling cultural heritage organisations to ensure their files comply to the PDF/A standards, veraPDF also provides the means to enforce institutional policies. The veraPDF policy checker allows the user to create their own acceptance criteria (“policy”) using XML Schematron syntax, enabling organisations to enforce restrictions in line with local policy, but beyond those stated in the PDF/A standards.

---

22 https://github.com/veraPDF/veraPDF-corpus
23 https://tests.verapdf.org/
6.1 Policy Example

Many cultural heritage organisations will not accept PDF/A-3 documents because PDF/A-3 allows the attachment of arbitrary file formats to the archival PDF. This runs contrary to the common approach of limiting the number of accepted formats to minimize preservation risk. The veraPDF policy checker can help organisations identify and analyse files attached to PDF/A-3 documents to ensure they meet acceptance criteria.

6.2 The Future

The middle of 2017 marks a pivotal moment for the veraPDF project. As PREFORMA’s test phase comes to an end, so does the governance and funding they have provided. veraPDF will become an independent, open source project. The OPF, PDF Validation TWG and Dual Lab will continue to address issues raised by users as well as testing and incorporating community contributions. Beyond this, future development will require funding.

The veraPDF consortium is developing plans to establish a community-led steering group, and explore open source business models 31 such as annual subscription or sponsorship, grant funding and consultancy for software maintenance and to extend the software’s functionality.

Unaddressed so far is the fact that PDF/A represents a small fraction of the document files cultural heritage organisations ingest; the vast majority are simply PDF (ISO 32000) files.

Developing a conformance checker for the PDF format as a whole is a huge undertaking, requiring many man-years of effort and resources. However, as the veraPDF validation model is format-neutral it can be extended to cover all aspects of PDF as well as PDF subset standards such as the forthcoming PDF/A-next specification, PDF/X (print), PDF/E (engineering), PDF/UA (universal accessibility), related specifications, and even third-party standards.

7 CONCLUSIONS

Cultural heritage organisations cannot be expected to provide expertise across all the formats they are responsible for preserving. Active dialog and collaboration with industry sectors who have relevant specialist expertise [5] is essential to the preservation of digital heritage, and the veraPDF project has proven that such collaborations are possible.

Open source software has properties that offer a compelling case for its use in preservation systems. Source code availability, publication of quality assurance results and the opportunity to play a role in developing and testing software can provide organisations with confidence that is difficult to establish when dealing with a commercial vendor and proprietary solutions.

The work described in this paper shows that collaborations between the cultural heritage sector and industry can be effective in building software that meets the needs of the digital preservation community and incorporates the specialist expertise of other sectors.

ACKNOWLEDGMENTS

The funding for this work was awarded to veraPDF by the PREFORMA project, a Pre-Commercial Procurement (PCP) co-funded by the European Commission under its FP7-ICT Programme under grant agreement no 619568. It runs from 1 January 2014 - 31 December 2017.

REFERENCES


A Dutch approach in constructing a network of nationwide facilities for digital preservation together

Joost VAN DER NAT
National Coalition Digital Preservation
P.O. Box 17 097, 1001 JB Amsterdam
The Netherlands
+j31 5204 8788
jvdat@ncdd.nl

Marcel RAS
National Coalition Digital Preservation
P.O. Box 90 407, 2509 LK Den Haag
The Netherlands
+31 14 77 7671
marcel.ras@kb.nl

ABSTRACT
Cross-domain collaboration lies at the heart of approach of the Dutch Coalition for Digital Preservation (NCDD) [1]. Sharing responsibilities, knowledge and infrastructures is an absolute necessity for the realisation of high-quality, effective and efficient digital archive- and information management.

In November 2016, the NCDD research on the construction of a national shared infrastructure for Digital Preservation was rewarded the Digital Preservation Award 2016 in the category Research and Innovation [2]. According to the judges the research report presents an outstanding model to help memory institutes to share facilities and create a distributed, nationwide infrastructure network for Digital Preservation.

This model describes a network of collaborative facilities as a backbone for sharing facilities within the Netherlands. It describes all the elements of an infrastructure for digital preservation. Not only the technical parts of it, but also its organisational, legal and financial parts. Describing them as elements, building-blocks, helps us to understand them and to get a better picture of what is already available on the one hand, and the demand on the other hand. One should notice that this model is based on the most desirable infrastructure. It describes the broad picture and enables us to create a roadmap towards the desirable situation. As always, the proof of the pudding is in the eating. This paper describes the theory behind the model and the steps to put the model into practice in an operational sense.

KEYWORDS

1 INTRODUCTION
Cross-domain collaboration lies at the heart of approach of the Dutch National Coalition for Digital Preservation (NCDD) [1] for the construction of a national shared infrastructure for Digital Preservation. It is key to realizing high-quality, effective and efficient digital information management.

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

Following on a national survey [3], the NCDD in 2010 formulated a strategic agenda [4]. This agenda consisted of a description of the major steps to be taken on a national level in the Netherlands in order to address the issues described in the survey. The strategy is centred on four themes:

1. knowledge-sharing;
2. development of a scalable and usable infrastructure for long-term management of digital information;
3. cost management; and
4. development of co-ordination in collection development policies.

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

2 NATIONAL COLLABORATION
Under the motto “Joining forces for our digital memory”, a research project was started in 2014 which was commissioned and financed by the Ministry of Education, Culture and Science. This project had the purpose to find out what level of differentiation between the domains offers the best balance for efficiency. Without collaboration, inefficiencies loom, while individual institutes continue to expand their digital archives and may be reinventing the same wheel over and over again. The project’s objective was and is to avoid duplication of work, and to avoid wasting time, money, and energy. Economies of scale make it easier for the many smaller Dutch institutes to profit from available facilities, services, and expertise as well. Policy makers can now ponder the question “The same for less money, or more for the same money?”.

The result of the project [5] was a model, which will be described in the next paragraphs. Next steps were taken within the national collaborative framework of the Dutch Digital Heritage Network (Netwerk Digitaal Erfgoed, NDE) [6]. This partnership has the intention of developing a network of nationwide facilities and services for improving the visibility, usability, and sustainability of digital heritage. The network was established on the initiative of the Ministry of Education, Culture and Science and consists of a
number of large organizations occupying key positions in the field of digital heritage.

The NDE has developed a three-pronged strategy covering Visible, Usable and Sustainable Digital Heritage, respectively. A work package has been established for each of these aspects, outlining the projects necessary to achieve its central goals. The work package on Sustainable Digital Heritage is led by the NCDD. Following on the results of the NCDD research project, a project within the sustainable digital heritage work package was constructed. This project built on the constructed model and developed a catalogue of existing service of all elements of the infrastructure for digital preservation. Also, case studies were carried out in order to get a better picture of the demand for services. Case studies should bring Supply and Demand together.

3 THE RESEARCH

With this research, the NCDD has presented a practical and down-to-earth framework for all parties involved, providing a clear picture of what such nationwide infrastructure might look like, including its scalability, and various growth models, culminating in the steps we need to take to achieve this, in both the short and long term.

3.1 Desk Research

Based upon extensive desk research [7 – 20] a model was developed for the catch-all term “infrastructure for digital preservation”.

In Figure 1 the constituent elements allow for an overall look at all the required elements and its relations. This shows that the necessary elements are not just IT-elements, but also involve knowledge, policies, quality control, training and many more organisational aspects.

The central part of the model consists of the OAIS functional entities. These elements are the heart of every preservation system and are the core of the processes of organisations, when concerning themselves with digital preservation. As shown in the figure above, these OAIS-functions are part of the core of the model.

Based on this high-level model we developed a “Business – IT stack” for digital preservation, with Building Blocks representing all elements we consider to be the part and parcel of the digital preservation environment.

In figure 2 all these elements are depicted as Building Blocks. From IT-elements as storage facilities, to standards, training, R&D, and semantics. Any organisation, or collaborative effort, obliged to ensure long-term access to digital information should have (most of) these elements organised. The model then focuses on those elements which could be shared within a network or a collaborative effort. The dark blue Building Blocks are the ones that are potentially shareable. The grey ones seem by definition organisation specific.

Each building Block can be seen as a Service. Services can be grouped together in Facilities [20].

3.2 Fieldwork

Extensive interviews with representatives of a number of large institutes [21], among them NCDD coalition partners, provided insight in the state of the current infrastructure and helped to sketch out a common vision for a desirable future situation.

Current facilities, often controlled by individual institutes, were discussed in detail, as well as the possibilities to share facilities and services with third parties. These interviews gave us a detailed picture of the possibilities for collaboration as well as domain specific services needing an individual approach.

As an example: for one group of elements, the “Support of the primary process of digital preservation”, the results of the interviews look like this:

The black – grey bars refer to the current situation, coloured bars refer to the preferable future situation.
E.g.: the dataset to the left: “Preservation Watch” is presently done mainly at the organisation level, whereas it is seen as a service that should be done at national level.

For all groups of elements of the defined infrastructure a concrete picture emerges of which infrastructural elements could be shared, which is a key factor in creating an organisational and technical infrastructure at the national level for sustained access without unnecessary overlaps. It might be better to speak of a network of nationwide facilities, as large parts of this national infrastructure have already been realised. An entirely new infrastructure from scratch is not necessary. Instead, existing facilities can be shared and opened up to third parties. In this way, a network of facilities will gradually evolve and materialise.

3.3 Scenario’s

Applying the outcomes of the fieldwork to the Building Block model, the following representation was developed:

Fig 4 Building Blocks as entities to be shared or organisation / domain specific

Having the shareable elements defined, scenarios were developed in order to set out the lines for collaboration. The Building Blocks are grouped in three sets. The preferred scenario is depicted as follows:

Fig 5 Preferred scenario for developing a National Infrastructure for LTDP

Technical infrastructure (A) and Preservation Services (B) can run parallel. Technical infrastructure (A) is conditional for Distributed applications for Digital Preservation (C).

3.4 As is and To be

Developments for implementation are already in progress. For example: the first steps for the Technical Infrastructure are being made with the development of four Dutch “Governmental Data Centres”. For Dutch Archives, the National Archive is deploying its Facility the e-Depot. And the NCDD is developing Services for Digital preservation in its present projects.

Many (semi)commercial solutions already exist (e.g. LOCKSS, Preservica, Archivematica to name just a few). They can be considered to offer a Facility with several embedded Services.

Our model would be instrumental to describe a desired (To be) situation of Services. It would then be possible to analyse the gap between existing and desired Services. This would be the foundation for a policy to develop new, and adapt existing Services.

4 MATCHING SUPPLY AND DEMAND

The project carried out within the framework of the Digital Heritage Network has its focus on bringing the model into practical use. Therefore, it was important to have a clear view on supply and demand. Having a model and having services in place is by no means a certainty for usage of the model and the existing services. Three things are necessary for this:

- an overview of tools and services in place, to be used as collaborative elements in a network: a catalogue of services;
- a common articulation of the demand for services; what do we expect from collaborative services?
- a specific articulation of the demand on an operational level; what types of objects need to be preserved under which mandates.

As these three elements where defined, case studies were done to create the connections between supply and demand based on the building blocks of the model. The case studies are the “proof-of-concepts” for the model. Based on the outcomes of the case studies the model could be adapted. The main goal is to bring together the Demand for Services for long term digital preservation with the Supply.

Fig 6 Supply and Demand for Facilities
4.1 Defining Supply: Catalogue of services
A first version of a Catalogue of services has been built. In this Catalogue, a list of Services is made available to smaller organizations who are in need of Services for long term digital preservation. These services are mainly the digital archives services in place at the large scale Cultural Heritage Institutes like the National Library, The National Archives and the Dutch Institute for Sound and Vision. Examples are:

a) the Dutch National Archive is deploying its “e-Depot” in a two-tier solution whereby 11 regional “Historical Centres” make use of a certified repository. They, in turn, can service archives of provinces, water authorities, local municipalities and even private collections;

b) the Dutch Institute for Sound and Vision can serve as the dedicated repository for audio-visual objects in the Netherlands;

c) the NCDD recently launched a Wiki (Dutch only) [22]. Its first topic is “Preservation Policy”.

4.2 Articulating Demand: definition
As explained above, the model for a network of distributed facilities for long-term preservation has different objectives. It helps to map the Dutch digital preservation infrastructure in order to create a nationwide infrastructure. It also is a theoretical model which helps us to define the elements of a digital preservation infrastructure. Even more, it models all elements of an infrastructure, not only the technical parts. Alongside this model organisations will be able to overview their needs and their own organisation. With the model, we were able to create a Catalogue of existing services in the Netherlands. At the flip-side of the coin, the model can help defining the demand for services. Cultural Heritage organisations having digital collections and, being aware of the fact that they need to be preserved for the long-term, can use the model to find out what they should organise. They can use the Catalogue to find out whether these services are available within their own domain and serving their needs. The model will help them define their demand.

4.3 Articulating Demand: operational level
Next step is within the operational level. Making demand more practical in knowing what exactly an organisation needs. Is there an urgent need for preservation storage, or development of a preservation policy? Should digital objects be migrated to another file-format or is there a need for an emulation-service?

Handing over digital collections to an e-depot service brings many questions, beginning with organising the mandate, knowing what types of digital objects you will hand-over and what are the volumes of it? It is therefore crucial to assist small(er) organisations in the articulation of their operational needs. To help organisations in this, a knowledge-driven Decision Support System is being developed.

5. THREE CASE STUDIES
As mentioned above (see paragraph 4) case studies will bridge the gap between the theoretical model and the operational Supply and Demand. Three examples of case studies are:

5.1 Frisian museums
The Dutch province of Friesland has some 40 museums, with Frisian cultural heritage collections. Within the project of Virtual Fryslân, financed by the Provincial government of Friesland, a large deal of these collections is being digitized. The project runs from 2015 until 2018 and it should digitize about 55% of the collections of the Frisian museums, which is about 450,000 objects. However, within the project no obligations for long-term preservation of these digitised objects were described. This turned out to be a perfect case study to find out how to map the demand of the Frisian museums to existing supply of facilities. Therefore, we first needed to have facts about the digitised collections (types of objects, file formats, storage facilities, volumes, etc). A survey was held under 30 participating museums. Some of the outcomes:

a) total volume is 1,25 Terabyte;

b) 2/3 of the museums preserve their collection in a sustainable way;

c) 57% stores the collection within their premises;

d) 19% do not have a back up facility.

e) 10 different file formats of stored digital objects:

Very interesting outcome of the survey is that 70% of the participating museums does have a policy or agreements for long-term preservation of the digitised collections. However, the financial means to preserve the collections for the long-term are low and will diminish in the forthcoming years. So there is an urgent need for a collaborative facility to be used by the Frisian museums as a group. As we now have the operational demand mapped out, we can start finding solutions within the network. Finding a supplier will be much easier now, as there is a clear view on what is needed.

The report of this study is to be published soon.

5.2 “Videcedings” of the Dutch Parliament
The Proceedings of the Dutch Parliament are by law to be archived with the Dutch National Archive (NA). These were, for centuries, paper documents. After the years stated in the Archival Law they are to be transferred to the NA. But with digitisation comes innovation, also in these long-standing traditions. Nowadays video recordings are made as well. These recordings need to be archived by the NA, conforming the law. However, the NA currently neither has the expertise, nor the (technical) infrastructure to do so for long term digital preservation. There are two ways to deal with this issue. Either the NA has to build up expertise and redevelop its infrastructure in order to meet the obligations it has. Or the approach could be to start collaboration with an institute which does have expertise and infrastructure in place already. In the Netherlands this is the Dutch Institute for Sound and Vision. The NA and Sound and Vision started a pilot project in which they clarified all aspects of this possible collaboration. Main point is that the NA does have the legal obligation and will always remain responsible for long term availability of the video notes. But the task
to carry out this responsibility can be sourced out. The pilot consists of the clarification of the organisational-, legal-, financial- and technical issues involved. It will result in a business model and firm specifications of service level agreements. This is an example of how the theoretic model can be executed in real life.

5.3 “De Digitale Stad”

De Digitale Stad (DDS; The Digital City) is the oldest Dutch virtual community and played an important role in the Internet history of Amsterdam and the Netherlands. For the first-time internet was (free) accessible to general public in the Netherlands. DDS is an important historical source for the early years of the Internet culture in the Netherlands. The virtual city and its inhabitants produced objects, ideas and traditions in new digital forms such as web pages, newsgroups, chat, audio and video. This social network proved to be very successful and had 140,000 inhabitants in 2000. But in 2001 it was taken offline.

Ten years later, in 2011, the Amsterdam Museum started the project “re-DDS”, the reconstruction of DDS [23]. Not only to tell and show the story of this unique internet historical monument, but also –and more important- to raise awareness about the risk of the loss of our digital heritage. This was the beginning of a case study in web archaeology: how to excavate, reconstruct, preserve and sustainably store born-digital data to make it accessible to the future generations.

The excavation has taken place, which resulted in reconstructions of two of the three phases of DDS [24]. Most parts of the environments are uncovered, but the next step is to revive the environments and put them in display at the Museum. And then the final step has to be taken, bringing the excavated data into a digital archive for long-term preservation. Here is where the collaborative services model comes in. As DDS is on the demand-side of it, a project has started in which different scenarios are worked-out for bringing the data from the excavation into the e-depot. Alongside three scenario’s all aspects of preservation of the DDS-data are described (technical-, legal-, organisational- and financial) using the model for a distributed network of facilities.

6. ONGOING AND FUTURE WORK

The presentation of a national strategy and the establishment of three Work Programs are an important development, which brings many existing initiatives and plans together. This is a start of an integrated approach for access to and preservation of Dutch digital heritage. The timing is perfect as there is a growing community of professionals involved in digital preservation. The level of knowledge exchange and the willingness to collaborate is growing too.

The model of Building Blocks has served as the basis for defining projects, and will continue to do so for future developments. Our challenge is now to bridge the gap between Theory and Practice even further. Supply and Demand have to be brought together. We are now developing a model to formally describe Services.

7. ACKNOWLEDGEMENTS

The Dutch ministry of Education, Science and Culture financed the research with a grant via the Network Digital Heritage (NDE) to the Dutch National Coalition for Digital Preservation (NCDD). Contributions to the research were made by many, amongst which representatives of (in alphabetical order): Beeld en Geluid, Data Archiving and Networked Services (DANS), International Institute for Social History (IISH), Koninklijke Bibliotheek, LIMA (platform for sustainable access to media art), Nationaal Archief, National Museum voor Wereldeculturen, Het Nieuwe Instituut, Rijksdienst voor het Cultureel Erfgoed (RCE), Rijksmuseum, Stadsarchief Amsterdam.

8. REFERENCES

[3] [3]
[4] [4]
[6] [6]
[7] [7]
[8] [8]
[9] [9]
[10] [10]
[12] [12]
[13] [13]
[14] [14]
[15] [15]
[16] [16]
[17] [17]
[18] [18]
[19] [19]
[20] [20]
[21] [21]
[22] [22]
[23] [23]
[24] [24]
BagIt Fixer-Upper
Scaling BagIt Tools to Manage the Ingest of Petabytes of Digitization Work

Nick Krabbenhoeft
New York Public Library
34-11 Thompson Street
Long Island City, New York 11101
nickkrabbenhoeft@nypl.org

ABSTRACT
The New York Public Library has created over 1.5 PB of files from digitizing over 50,000 audio and video items for the long-term preservation of their content. This paper details the Library’s usage of the BagIt File Packaging Format during Quality Assurance and Audit Submissions functions as defined by OAIS. It also discusses extensions of the bagit-python library in order repair bags that do not pass those functions.

Working with thousandsofterabytes storedinhundredsof thousands of bags requires that our approaches to ingest scale appropriately. Common changes to bags such as the accidental creation of system files in bags or purposeful edits of metadata files will invalidate the entire bag. Noting and responding to these errors is critical for improving workflows, but manual response is impossible. Using the bagit-python library, NYPL has created tools to selectively clean system files from bag directories and manifests, update or add checksums, and create event logs of repairs.

KEYWORDS
Quality Assurance, Audit Submissions, Scalability, Fixity, BagIt, Python, Digitization

1 INTRODUCTION
Before an archive commits to the long-term preservation of a submission, it must verify that it has received expected information. The OAIS Reference Model defines two functions for checking content during ingest: Quality Assurance and Audit Submissions.[3]

Quality Assurance is performed after a producer has transferred a SIP to the archive’s1 Receive Submission function. It is used to “identify any file transfer or media read/write errors.” After passing Quality Assurance, the Receive Submission function transfers the SIP to the Generate AIP function. From there, the SIP and/or the

Figure 1: Diagram of the subset of the OAIS functions discussed in this paper. Dotted boxes denote functional entities; grey boxes denote functions, and text next to arrows denotes information provided by one function to another.

AIP can be submitted to the Audit Submissions function in order to “verify that submissions (SIP or AIP) meet the specifications of the Submission Agreement.” If a SIP or AIP fails the audit, either the producer can appeal the audit or the archive can request a new SIP. Restated, Quality Assurance checks that the structure of the SIP was not damaged in transit, and Audit Submissions checks that the intellectual content of the SIP and/or AIP matches expectations. See Figure 1 for a visualization of these relationships.

It is the archive’s responsibility to define how Quality Assurance and Audit Submission functions are performed. The text of the OAIS includes suggestions for methods such as using system logs and/or hash functions for quality assurance (QA) and review committees for audits, but these are non-exclusive. In implementing these functions at NYPL, two questions have shaped implementation decisions.

(1) How well do the QA and auditing methods scale?
(2) How can information generated by QA and auditing methods improve workflows?
To make any QA method scalable, the required fixity information must be stored in a machine-readable format. For hashes, there are three types of fixity information.

**Hashes** comparing stored cryptographic hashes against ones generated from payload files

**0xum** comparing a stored count and total size of payload files against newly generated numbers

**Completeness** comparing a stored list of payload filepaths against those currently in the payload

The bag consists of a data directory that contains the payload and several metadata files that record the above fixity information and other metadata. For NYPL, the payload comprises the preservation master files, metadata files, and service copies from a single digitized object.

### 2.1 Types of Fixity

The gold standard in fixity information of digital files is the cryptographic hash. Because generating a hash for an object after a transfer and comparing it to a previously known value can detect most bit-level losses, it makes sense as the mechanism for checking the technical accuracy of a file transfer. Recommendations for hash comparisons are common in digital preservation documents from the OAIS Standard to the US Government’s Federal Agencies Digitization Guidelines Initiative.[4]

However, hash calculation is I/O-bound. Every bit of a object must be read to calculate a hash.

There are other types of fixity information that are much faster to generate. For example, if a group of digital objects have not changed, then metadata describing the group like the total number of files, the total size of the files, and the list of filepaths, should remain fixed as well.

Unfortunately, these properties are not as sensitive to change as a cryptographic hash. A bit flip would not affect any of these values. On the other hand, these coarse measures of fixity are far more responsive to coarse modifications such as deleted files, accidentally renamed files, and renamed files.

While these other types of fixity information can not replace hash validation in a Quality Assurance function, using them prior to hashing can immensely speed up the discovery of common errors.

#### 2.2 Benefits of BagIt for Quality Assurance

To make any QA method scalable, the required fixity information must be stored in a machine-readable format. For hashes, there are two widely used formats:

- **manifest files** listing the relative path and hash for a directory of files in a single CSV
- **sidecar files** storing an md5 hash for example.mov in example.mov.md5

These common formats, and the many tools that can read them, make hash validation relatively easy to adopt as a quality assurance method, but for other types of fixity information, there are very few common formats.

A single organization may specify a standard way to store other types of fixity, and even create software to generate and check this method. But, without community adoption, these methods are fragile at best and lost opportunity costs at worst.

The BagIt File Packaging Format defines a standard format for storing the total size and total number of files in a payload, known as the 0xum (total-file-size.total-file-count). It also places strict requirements on the hash manifest. The path for every file in the payload must be represented in a hash manifest. As a result, it is possible to check that a bag is complete, that its payload contains only expected files, no more no less.

In terms of commonality, the BagIt format is both an IETF RFC and an open source specification on GitHub. Tools to create and validate bags have been implemented in multiple languages, including Java[10], Python[12], and Ruby[8], making these additional fixity checks widely available. Additionally, most implementations are open-source, allowing users to further customize and adapt the tools as needed. At NYPL, the Python, Ruby, and Java implementations have been adopted by different processes in the Library.

### 2.3 Disadvantages of BagIt for Quality Assurance

Before discussing BagIt’s usage with the NYPL workflow, it is important to highlight a few difficulties that can result from adopting BagIt.

First, the strictness of completeness. The manifests of a valid bag must list every file in the payload directory. This requirement conflicts with the behavior of operating systems that zealously create hidden system files to store usability data such as preferred file sorting order, file thumbnails, and indexing information. Browsing a payload directory after bagging often causes these files to appear, which renders the manifests incomplete and the bag invalid. Even if the system files existed at the time of bagging, operating systems will silently modify them with new usability information, which renders the hashes inaccurate and the bag invalid.

Second, multiple implementations of a standard can result in multiple interpretations. Ideally, a BagIt tool should create a bag that can be validated by other BagIt tools, but edge cases can make this difficult. For example, Mac systems can create a file named /etc/hosts when written into a manifest, the \r acts as a carriage return, making it impossible for other BagIt tools to parse the manifest correctly. Following the requirement in the BagIt RFC to write filepaths with percent-encoding avoids this problem, but users should be aware that malformed bags can cause validation problems.²

²The development of a conformance suite by Library of Congress staff has made it much easier to test an implementation for common errors. [2]
Using BagIt successfully requires adapting workflows to these challenges.

3 NYPL WORKFLOW

NYPL's ingest workflow can be divided into roughly 4 stages.

1. Bags are received on hard drives from digitization labs and validated.
   - If a bag is not valid, repairs are made when possible or a new transfer is requested.

2. Valid bags are transferred to quality control and audited against the library's published specifications for signal quality, file format, metadata values, and file organization.
   - If a bag does not pass quality control, repairs are made when possible or redigitization is requested.

3. QC'd Bags are transferred to a staging area on network storage and validated after transfer.

4. During ingest to the repository, bags are validated again.

Like most real-world preservation workflows, the NYPL ingest process does not mirror the simplicity of the OAIS model (Figure 1). For example, there is no single Quality Assurance function. Instead, quality assurance is performed each time the bag moves to a different storage medium.3

This workflow also makes allowances for alterations to the SIP. While the Generate AIP function can include file format conversions and metadata gathering or conversion, it does not explicitly include provisions for altering the contents of SIP. However, despite efforts to produce comprehensive specifications, bags can fail the quality control processes.

In the standard, failure during the Audit Submissions function can result in either the producer negotiating for a pass or the archive requesting a new submission. No response to failure is discussed for the Quality Assurance function.

In practice, repair is the most preferable action, when appropriate. Whether the specifications were not exact enough or the cause of failure was not severe enough, the cost of repair can be much lower than resubmission.

4 COMMON QA AND AUDIT FAILURES

In addition to signaling problems with a specific bag, QA and audit failures often highlight systemic workflow problems that require a combination of communicating with producers and refining specifications. The following lists the most common causes of QA and audit failures, why they are flagged as failures, and how they are remediated:

3The Inner OAIS-Outer OAIS model presents an interesting approach to the challenge of mapping real-world work to OAIS ideals. [13] It is possible to model the above workflow as a three linked OAIS archives, a receiving archive, which accepts and manages hard drives on behalf of its designated community; the quality control archive, which accepts and manages bags on behalf of its designated community; the repository archive, which accepts and manages digitized objects on behalf of its designated community; the access unit. Each "archive" maintains their own functional entities for Ingest (including a Quality Assurance function), Data Management, Storage, Dissemination, Administration, and Preservation Planning, and their work is coordinated by an Outer Administration and Preservation Planning entities. While full documentation of organizational units at this level of detail may be impractical, the scalability of the OAIS Reference Model to describe large and small "archives" is an interesting theoretical question.

4.1 Invalid Metadata

Submitted metadata is validated against a schema to ensure correct usage of controlled vocabulary and completion of required fields. A manual spot check of a bag's metadata revealed that a vendor had accidentally omitted a set of required fields while updating their workflow. Further investigation showed that this omission was not caught because a typo in an early version of the schema rendered those fields optional, and that several hundred bags shared the same problem.

The missing metadata included fields required for repository ingest and used to create descriptive records. In order to prevent future metadata omissions, the metadata schemas received a thorough review during which a similar mistake was found and fixed. To alleviate the hold on repository ingest and description, tools were developed to repair metadata for all affected bags.

4.2 System Files

As discussed previously, system files can render a bag invalid. The system files themselves do not pose a preservation risk, but their existence indicates that when the bag was QC'd, it was likely mounted with read-write privileges. This exposes the bag to the risk of corruption or deletion through human error. To prevent this possibility, the QC workstation is checked and configured to mount all drives as read-only.

4.3 File Name Changes

Part of the SIP packaging specification includes storing all preservation master files in a directory named PreservationMasters, even if the folder only includes a single file. Digitization engineers have bagged a project and then removed the s from the directory name, PreservationMaster, when they realized it only contained one file. This change causes a bag completeness failure, since the file path listed in the bag manifest no longer exists.

Since the folder name is part of the packaging specifications, it is also built into the repository ingest process. Bags without this directory cannot be ingested to the repository. Addressing this problem required working with engineers to make sure the specifications were written clearly.

4.4 Missing Checksums

The SIP specifications require that metadata files are packaged alongside the preservation master. An early workflow would bag the digitized media and then add the appropriate metadata files to the bag. This change causes a bag completeness error, since the metadata files are not listed in the manifest.

While hashes are most critical to ensure the fixity of preservation master files, we are also interested in ensuring the fixity of metadata files because they contain an audit trail of how the file was created. Again, the solution was to work with the engineers to make sure the specifications were written clearly.

4.5 Missing Files

The most worrying error is a bag completeness validation failing because of missing files. This is the realization of the accidental deletion specter discussed in regards to system files. An audit of our staging environment once revealed a bag that was missing most
of its files. Fortunately, the original item could be re-digitized, but the discovery triggered an account audit of our staging server and review of our tools and procedures in order to reduce the chance of complete loss again.

5 AUTOMATING BAG REPAIR

Except for the last example, each of the above failures can be repaired. Bags remain eligible for ingest as long as it can be shown that a failure did not impact the fixity of the preservation master or mezzanine. However, the interlocking nature of 0xum, completeness, and hash fixity in bags means that any change to a bag’s payload likely requires edits to multiple parts of the bag.

For example:

- adding metadata files to a bag requires adding the path and hash for each file to the bag manifest and updating the 0xum with the new number of files and total size
- removing system files from a bag requires searching for all common system file names in a bag, comparing that list to the paths in the manifest, and deleting only those files not listed in the manifest
- updating all metadata files with missing technical fields requires regenerating hashes for only the metadata files, recording them to the manifests, and then updating the 0xum with the new payload size.

These are relatively simple but very tedious procedures and prone to human error.

In response, NYPL extended the bagit-python library. The tool update_bag.py automates the repair of bags. [6] Using the methods of the Bag class within bagit-python, update_bag.py performs set operations to identify missing or unwanted files, add or update fixity information in the bag manifest and 0xum, and validate the bag.

Performing these repairs at scale also requires automating the creation of an event log to serve as an audit trail. At this time, every update performed by update_bag.py creates a PREMIS event record that is saved as a JSON file within the bag. Future repairs with update_bag.py are appended to the JSON file. For simpler events like an 0xum update, both the old and new 0xum are recorded in the PREMIS event. For more complex events like rewriting the hash manifests, copies of the original manifests are kept. The tool is available as part of the ami-tools package developed by NYPL.

6 DISCUSSION

Returning to the issue of repair and OAIS, this exploration of Quality Assurance and Audit Submissions functions at NYPL has reiterated how the density of OAIS functions and their relationships remains a hurdle to working with the Reference Model. The descriptions of the functions in the Model are carefully constructed to be neither overly prescriptive nor prescriptive. This leaves the responsibility of creating a shared understanding of allowed actions to the OAIS user community.

But the shared understanding remains elusive. What is described here as a metadata repair as part of the Generate AIP function in the Ingest functional entity, might be described as a pre-conditioning action that took place prior to ingest. [9] Or according to a proposal to create a Pre-Ingest functional entity it might be described as a pre-submission action on a Pre-Submission Information Package. [11] It remains unclear if the Reference Model is incapable of filling community needs for a common language to describe workflows or if the community is unwilling to engage with the Reference Model as a common language.

More prosaically, the BagIt format is an interesting match with the Reference Model. Its form is at once reminiscent of the archival box and the OAIS Information Package. Although update_bag.py contains logging features born of quality assurance needs, not long-term storage needs, others in the BagIt community have proposed more rigorous methods to update bags over time. For example, the Restful Bag Server specification proposes the use of version control methods like Git and Mercurial to store a fuller history of changes to the bag. [1] Enriching the container with its own PDI, the bag could mature into a common format of AIP. Already it contains material for ingest into the Digital Preservation Network (DPN) [7] and APTrust [8] and can be both input and output for Archivematica.

At the same time, it would be best if BagIt is one of multiple common AIP formats. For all of its benefits, BagIt is still based on a metaphor of managing the digital world like the physical world. When bags are ingested into DPN’s and APTrust’s cloud storage, they are turned into a tarball so that the entire bag can be treated by a single object by the object-based storage systems employed by S3, Azure, and other storage providers. Regaining access to one file in the bag requires downloading the entire bag and unzipping it.

7 CONCLUSION

BagIt is a mature technology that lends itself well to Quality Assurance functions in OAIS workflows. At NYPL, usage of the BagIt Format is a key component in the audio, video, and film digitization workflow. It allows for consistent, scalable fixity checking. It is also a finicky format that becomes invalid as a result of a litany of small changes or inconsistencies. In the field of digital preservation, this is feature not a bug. Even in the Generating AIP function where changes to the SIP are allowed.

The sensitivity to fixity problems forces the archive to consider how its infrastructure and workflows interact with the material that they manage. Lack of robustness magnifies irregularities in workflows that can effect fixity of submitted packages, and the accessibility of its open-source implementations allow for extension and customizability of these tools to fit very specific workflow needs.

8 REFERENCES

REFERENCES


Modular Pre-Ingest Tool for Diverse Needs of Producers

Kuisma Lehtonen  
CSC – IT Center for Science  
P.O. Box 405, 02101 Espoo  
Finland  
Kuisma.Lehtonen@csc.fi

Pauliina Somerkoski  
CSC – IT Center for Science  
P.O. Box 405, 02101 Espoo  
Finland  
Pauliina.Somerkoski@csc.fi

Juha Törnroos  
CSC – IT Center for Science  
P.O. Box 405, 02101 Espoo  
Finland  
Juha.Tornroos@csc.fi

Mikko Vatanen  
CSC – IT Center for Science  
P.O. Box 405, 02101 Espoo  
Finland  
Mikko.Vatanen@csc.fi

Kimmo Koivunen  
CSC – IT Center for Science  
P.O. Box 405, 02101 Espoo  
Finland  
Kimmo.Koivunen@csc.fi

ABSTRACT
We introduce an open-source pre-ingest tool that assists the generation of Submission Information Packages (SIPs) that are to be submitted to the national digital preservation service in Finland. The pre-ingest tool consists of several independent components that produce the parts of a METS document required by the national preservation service. These components are easy to modify when developing services for different user demands or for different repositories. Users of the tool provide the necessary information as parameters for the tool, which produces the structure and descriptions for the SIP. The pre-ingest tool reduces the need to deeply understand either METS, PREMIS or other metadata formats to be able to preserve digital assets.

CCS CONCEPTS
• Information systems → Information systems applications → Digital libraries and archives • Information systems → Open source software

KEYWORDS
Digital Preservation Tools, Pre-Ingest, Open-Source Software

1 INTRODUCTION
Preparing and ingesting digital assets in an appropriate format to a preservation service can be a demanding task, especially in cases in which the producer is not familiar with the various preservation standards and metadata formats. This can be very time-consuming and therefore a very costly process. Thus, we have developed a pre-ingest tool to make it easier to create Submission Information Packages (SIPs) programatically, which helps our partner organizations (libraries, archives and museums) to ingest digital assets to our national digital preservation service. Our pre-ingest tool decreases the burden of preserving data technically and thus releases time, which can then be used to produce digital assets instead of being wasting on considering how to preserve such assets.

Our national digital preservation repository provides services for preserving the cultural heritage and research data funded by the Ministry of Education and Culture of Finland. Currently, the preservation service is available for libraries, archives, museums and other organizations in Finland preserving cultural heritage by statutory obligation. We will extend our customer base soon with various organizations that have the need to preserve data, mainly nationally produced research data including publications and research methods. Given the diversity of the user needs, the digital assets to be preserved make up a very heterogeneous whole and simultaneously require various and flexible solutions.

Our national digital preservation service, based on the OAIS reference model [1], has been in production since 2015. Currently, we have more than one million Archival Information Packages (AIPs) in preservation, which amounts to more than 100 terabytes. We have defined common national preservation specifications [2], which describe in detail how digital assets should be prepared before ingesting them to the preservation service, including requirements for metadata and file formats. The design, implementation and development of the national digital repository are done in close collaboration with partner organizations. Thus, we have ensured that the user requirements are fulfilled by the pre-ingest tool.

To assist ingesting digital assets, our flexible pre-ingest tool can be used to generate SIPs. The tool produces a METS document containing all the necessary metadata conforming to our national preservation specifications. The tool includes creating descriptive and administrative sections for a METS document, creating a structural map, automatically extracting technical metadata from
files into the PREMIS metadata format, digitally signing the SIP and finally packaging the SIP as a TAR or ZIP package.

The rest of this paper is organized as follows: in Section 2, we give background information regarding the reasons why we developed this tool and how it is related to other pre-ingest tools. Section 3 presents the functionality of our tool. Finally, in Section 4 we conclude the paper and outline our future plans.

2 BACKGROUND

Ingesting digital assets into a preservation system can be a significant burden for some organizations, especially those with insufficiently competent IT staff. Therefore, we needed to simplify the pre-ingest process by providing a software tool for this. Earlier, in 2013, we found [3] that national memory organizations may have problems in creating syntactically and semantically correct SIPS with a valid METS document. Further, some organizations found the process of creating valid SIPS very time-consuming. The most common mistakes were related to the creation of a METS document with missing mandatory attributes or a document that misuses them. In addition, it was somewhat common to have errors in namespace definitions and internal references in the METS document. These somewhat trivial errors (for those having sufficient knowledge of the necessary metadata formats) are easy to prevent by using the pre-ingest tool to produce a valid METS document which contains most of the "low-level" tasks of creating a SIP. Based on these experiences, and with close collaboration with national memory organizations, we defined the functional and non-functional requirements for the pre-ingest tool.

There are various tools for creating SIPS for various repositories. For example, RODA-in [4] is an application for creating SIPS from local files and directories that includes batch-processing features. RODA-in provides an easy-to-use graphical user interface for creating SIPS but only supports BagIt and E-ARK SIP formats. The Rosetta SIP Factory [5] is a tool for creating SIPS suitable for the Rosetta preservation software [6]. We have chosen a more modular structure that constructs the METS document in parts, which gives more flexibility with which to integrate with back-end systems. The DURAARK WorkbenchUI [7] has been developed to assist the pre-ingest of architectural 3D data to the DURAARK system. It contains a graphical user interface and is implemented with Java. It supports BagIt and Rosetta SIP formats. In addition, some other preservation systems include functionalities for preparing the data for ingest, such as Archivematica [8], which supports the BagIt SIP format.

It is essential for us to avoid monolithic or complex workflows. In our needs, modularity, flexibility for modifications and the possibility to integrate and automatize the SIP creation process for partner organizations’ back-end systems are the key issues. The diversity of back-end systems and processes, and the variety of metadata standards used in partner organizations lead to modular implementation of the pre-ingest tool.

As noted in [9], our digital preservation service is designed to receive large amounts of digital assets using a carefully designed and automated validation process at the ingest phase (see [10] for details). However, when the volume of data increases (both in the size and number of SIPS’), it is crucial to be able to also automate the pre-ingest phase. Organizations that need to send a large amount of data may integrate their back-end systems with our national digital preservation system. Our pre-ingest tool can be deployed for this integration work very flexibly.

Our tool aims to create different pre-ingest services for our digital preservation service. Since the tool consists of several independent components, it is relatively easy to modify it or replace some components for different needs, including cases when developing services for different producer needs or for different repositories. Further, pre-ingest services with a GUI can be implemented for organizations that occasionally need to use the pre-ingest tool.

3 FUNCTIONALITY

Our pre-ingest tool is a set of modular software components. These components produce parts of a METS document and eventually produce a SIP that conforms to our national specifications. We have published the tool at GitHub as open-source software [11]. The architecture of the pre-ingest tool is selected to support easy service generation on top of it and to support the integration of a diversity of customer systems into the national digital preservation service while also allowing independent use when full system integration is too heavy. Pre-ingest services that are built on top of the pre-ingest tool employ the components of the tool in an appropriate order. Further, pre-ingest services may implement supplementary functionalities (e.g. sophisticated error handling and detailed reporting) whenever necessary.

An overview of our pre-ingest tool is depicted in Figure 1. To build a SIP conforming our digital preservation specifications, one can perform the following steps in the following order: a–g (where steps a–d can be executed in any particular order). As a result, these steps produce a digitally signed METS document, along with the digital assets composing a complete SIP that the producer can ingest. In what follows, we describe these steps in more detail.

Firstly, the tool generates descriptive metadata (step a). The descriptive metadata must be in a separate file and in a format acceptable in our repository (e.g. MARC, DC, MODS, EAD, EAC-CPF, LIDO, VRA or DDI). The location of metadata is given as an argument, and the tool encapsulates it into the METS descriptive metadata section. Descriptive metadata can describe a single file or a collection of files (e.g. a digitized book consisting a TIFF image per page).

Secondly, the tool generates technical metadata (b) for the files to be ingested. The technical metadata sections of the METS document are generated using an appropriate technical metadata format (e.g. PREMIS, MIX, VideoMD or AudioMD). The tool uses open-source components, for example the well-known JHove tool [12], to extract the necessary technical information from files. Should the technical metadata already be available in the organization’s back-end system(s), the producer can easily modify the tool in order to retrieve this information from the back-end system automatically.
Thirdly, the tool generates rights metadata for digital assets (c). The required information is given as an argument. Rights metadata, like the descriptive metadata, is metadata that cannot be generated automatically in most cases but the information is crucial when disseminating digital assets years, or tens-of-years, later.

The fourth step is the generation of provenance information (d). The provenance section or sections of the METS document are created in PREMIS by giving the necessary information for the events and agents as arguments. PREMIS events and agents can describe the provenance information of a single file or a collection of files.

As noted above, these first four steps (a-d) can be executed in any order as they are independent from each other. The output of these steps, however, has to follow some general rules so that latter steps can produce a valid SIP.

The fifth step requires input from the first four steps in order to generate a structural map and file elements (e) for the METS document. In this step, the tool automatically creates references from the structural map to descriptive metadata and references from file elements (the digital assets to be preserved) to provenance information, technical metadata and rights metadata. Figure 2 depicts the element linking in a METS document in our national METS profile. The structural map is created based on different profiles. By default, the directory structure of the digital assets is described in the structural map. However, for example, some organizations create structural maps based on a structure described using the EAD\(^1\) metadata format.

Lastly, the tool generates the METS document (f) and composes the SIP (g). The tool collects the results of the previous steps and creates a complete METS document. The METS document is then digitally signed with a PKCS\(^7\) signature\(^2\) in order to ensure the integrity and fixity of the SIP during transfer into the preservation service. However, ingesting the SIP into the digital preservation service is not in the scope of the pre-ingest tool.

**4 CONCLUSIONS AND FUTURE WORK**

In Finland, the digital preservation of cultural heritage is enabled by a generalized preservation service for national libraries, archives and museums. The digital preservation service as an organization supports these memory organizations in several activities related to digital preservation. National memory organizations work in close collaboration with the preservation service in order to define national preservation specifications, which describe in detail how digital information shall be gathered in SIPs. In practice, this pre-ingest phase also requires several steps that are common for all organizations utilizing the preservation service. To help partner organizations in the pre-ingest phase, we have introduced a tool which simplifies the process of constructing SIPs. Especially, the pre-ingest tool significantly reduces the need for a deeper understanding of METS and PREMIS metadata formats and automates these common steps in the SIP’s construction. With the

---

1. Encoded Archival Description, [https://www.loc.gov/ead](https://www.loc.gov/ead)
tool partner organizations can decrease the costs of the pre-ingest phase and therefore considerably lower the barriers to initiating the preservation of digital assets with the generalized preservation service.

We have created the pre-ingest tool in close collaboration with partner organizations. The requirements are diverse – one organization may even have multiple systems integrated into the national digital preservation service – so the pre-ingest tool must be modular enough to fulfill the different needs.

The pre-ingest tool is in production in partner organizations in order to integrate their back-end systems with the national digital preservation service. We have received good feedback when testing it with a representative sample of partner organizations. The pre-ingest tool will be used more widely when new partner organizations deploy our digital preservation service.

The pre-ingest tool will be developed further when national preservation specifications are updated. For example, support for new file formats, metadata standards and checksum algorithms can be implemented when needed. In the near future, there will be a need for more sophisticated SIP creation services. For example, the Open Science and Research Initiative (a national initiative for promoting research information availability and open science) has plans to start the digital preservation of research data. The pre-ingest tool can be used as a basis for easy-to-use SIP creation services aimed at different disciplines.

ACKNOWLEDGEMENTS

The authors would like to thank all members of the Digital Preservation team at CSC – IT Center for Science Finland, as well as the National Digital Library support group for their valuable comments and input during the preparation of the national digital preservation service. Lastly, we thank the Ministry of Education and Culture of Finland for funding this project.

REFERENCES


1 http://openscience.fi/
Permanence of the Scholarly Record: Persistent Identification and Digital Preservation – A Roadmap

A. Dappert
The British Library
96 Euston Road, London, NW1 2DB
UK
angela.dappert@bl.uk
orcid.org/0000-0003-2614-6676

A. Farquhar
The British Library
96 Euston Road, London, NW1 2DB
UK
adam.farquhar@bl.uk
orcid.org/0000-0001-5331-6592

ABSTRACT
This paper proposes steps towards a roadmap for improving the integration of two communities that deal with persistence and long-term stewardship of digital content. They are Persistent Identifiers (PIDs) and Digital Preservation. Both disciplines have made significant progress and practical contributions. Yet their approaches are not fully linked and there is considerable potential to integrate their solution space and to improve either of them by learning from the other. It addresses three core issues:

1. How does the long-term digital object life-cycle affect PIDs, the entities they identify, and the metadata that describes them?
2. How can PIDs help long-term preservation?
3. How can long-term preservation help to shape PID best practice and ensure long-term access to the scholarly record?

We also sketch out initial results of our ongoing work along this roadmap.

KEYWORDS
Persistent identifiers, digital preservation, scholarly record, roadmap, distributed collaboration

1 INTRODUCTION
The Persistent Identifier and Digital Preservation communities both address core issues related to persistence and long-term stewardship of digital content. They have made significant progress and practical contributions over the past two decades. In spite of addressing shared challenges, the opportunities for each to benefit from the other’s progress remain largely unrealized. This paper addresses three core questions to provide a roadmap to improve sharing of results:

1. How does the long-term digital object life-cycle affect PIDs, the entities they identify, and the metadata that describes them?
2. How can PIDs help long-term preservation?
3. How can long-term preservation help to shape PID best practice and ensure long-term access to the scholarly record?

While the work described can be applied to any digital material that is worth preserving, we emphasize the scholarly record. It is characterized by complex sets of contributors and long chains of information creation and exchange. These necessitate globally unique persistent identifiers more than many other digital materials.

1.1 Space, Time and Intent
Persistent identifiers (PIDs) play an important role in the scholarly infrastructure. They enable both people and computational agents to reliably identify and link to entities such as articles [e.g. provided by Crossref] or data [e.g. provided by DataCite]. Furthermore, they can be used to identify researchers [e.g. provided by ORCID] or rights-holders living or dead [e.g. provided by ISNI]. PIDs are becoming an essential component in the workflows of funders, researchers, research organizations, data centers, publishers, libraries, and others. As of 2017, tens of millions of PIDs have been assigned to these core scholarly entities through global PID service providers and their partners. Trusted and reliable identifiers associate a resource with a character string. The following PID criteria hold (inspired by [1]):

- A PID is a name, rather than an address.
- PIDs are globally unique.
- PIDs are persistent.5
- PIDs are selective at the right level of granularity.
- PIDs are interlinkable.
- PIDs are interoperable with other identifiers.
- PIDs are designed to last beyond the lifetime of any system or (most) organizations.
- PIDs are globally resolvable as a URI with support for the full range of HTTP including content negotiation.
- PIDs are managed through a sustained committed organization and governance process.
- PIDs come with metadata that describes the resource’s most relevant properties.

PID services mint PIDs on request and provide services such as registration, metadata management, fragment identification, content negotiation, search and discovery, and governance. Content owners manage the content and ensure that content

---

1 https://www.crossref.org/
2 https://www.datacite.org/
3 https://orcid.org/
4 http://www.isni.org/
5 This refers to the PID itself. It does not imply that the content must be persistent at all times. For example, the content may be streamed or be versioned.
location information is kept up-to-date with the PID service. Because both the PID service and the content owners hold metadata that describes the identified resources, PIDs can be indexed and searched. PID service providers, such as DataCite, establish a contractual governance commitment with content owners to ensure long-term stewardship and accessibility of the identified resource.

Leading motivations for PID use lie in the stewardship of the scholarly record. PIDs “improve the ease of locating resources; are actionable on the Web; enable metadata update and corrections without losing the resource’s identity; can integrate legacy naming systems; promote linking and interoperability between services; and reduce confusion among versions of a resource. Widespread uptake of PID e-Infrastructures can accelerate the adoption of Open Science by building trust through seamless discovery of scientific artefacts; clear attribution to contributors; traceable provenance; unambiguous citation in scholarly discourse; supporting reproducibility; and enabling improved metadata quality through linking connected metadata sources.” [2].

PIDs can be applied to publications, data, other research outputs, researchers or other personas, organizations, legal entities, funding instruments, projects or patents, and more. They can also be used to distinguish aspects of an entity such as separate versions, multiple formats, levels of granularity; or of an object, such as its intellectual definition (e.g. a FRBR work or expression), or its rendition consisting of a bitstream, a single file, or a composite set of files.

PIDs and the relationships between them create a connected network of information about the global scholarly record. This is a graph in which the metadata of one PID relates to that of another PID. PIDs can and are being used in the workflows of funders, researchers, research organizations, data centers, publishers, libraries, and others. Most of the discussion has been around their “contemporary” functions in the processing, use and reuse of resources, such as in data center or publisher workflows.

It is important to note that the scholarly record and this graph and the resources within it are used across three dimensions:

- **Space**: semantic linking of PID-identified resources creates the open eScience landscape in which we can globally connect and analyze data and associated metadata. This is emphasized in linked open data environments.
- **Intent**: Resources can be reused and re-purposed in ways that were unanticipated during their creation.
- **Time**: The life-cycle of research and the scholarly record spans centuries. It reaches from the conception of research ideas to reuse of results decades or centuries after their creation.

There has been much work towards creating the connected scholarly record and enabling validation and reuse of research outputs. In contrast, relatively little effort has gone towards ensuring long-term access to the scholarly record. It is this last dimension, “time”, for which this paper outlines a roadmap for future work.

Use cases across time include reserved PIDs for preliminary research outputs; transfer of responsibility for an entity from a creator or publisher to a memory institution; use of PIDs and content resolution after format migrations that are necessitated through obsolescence; handling deleted or lost data; PID creation for data that are created within memory institutions, when large data collections are mined resulting in derivative data sets; and provision of long-term stewardship for identified content.

![Figure 1: PID-related information flow among stakeholders](image)

---

[2] Reference to a source not provided in the text.
We investigate the role that memory institutions and digital preservation practitioners play in the use and preservation of PIDs, their accompanying metadata, and the content they describe. Much of this is done by raising research questions whose answers will help improve both PID use and digital preservation practice.

2 THE SCHOLARLY RECORD LIFE-CYCLE

How does the digital object life-cycle affect PIDs, the entities they identify, and the metadata that describes them? To investigate this we need to look at the following issues:

1. What role should PID stakeholders play in order to ensure long-term preservation?
2. How can one manage the distributed long-term responsibility?
3. How can PIDs support entities that evolve over time?
4. How can we preserve the PID graph as it grows over time as more links are established through incremental improvements, use and reuse?

2.1 Role of stakeholders in ensuring long-term preservation of PID-related information

While PIDs are sometimes created for local use only, they are inherently about global reuse and establishing the connected scholarly record. Therefore, PID-related information is passed along from one organization to another across the life-cycle. They include:

- **PID services**
- Funders
- Research organizations
- Researchers (as producers, consumers and reviewers)
- Publishers
- Data centers
- Institutional information managers
- Citation managers
- Indexing services
- Public content repositories
- Private content repositories
- Libraries and archives

The stakeholders perform different Business Actions on the Information Objects over the life-cycle.

<table>
<thead>
<tr>
<th>Business Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funding</td>
</tr>
<tr>
<td>Research management</td>
</tr>
<tr>
<td>Institutional information management</td>
</tr>
<tr>
<td>Research</td>
</tr>
<tr>
<td>Impact assessment</td>
</tr>
<tr>
<td>Reviewing</td>
</tr>
<tr>
<td>Publishing</td>
</tr>
<tr>
<td>etc.</td>
</tr>
</tbody>
</table>

Each business action consist of a sequence of basic actions that affect the Information Objects in the following ways.

<table>
<thead>
<tr>
<th>Basic Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create</td>
</tr>
<tr>
<td>Update</td>
</tr>
<tr>
<td>Enrich</td>
</tr>
<tr>
<td>Replace</td>
</tr>
<tr>
<td>Delete</td>
</tr>
<tr>
<td>Transform formats</td>
</tr>
<tr>
<td>Disambiguate</td>
</tr>
<tr>
<td>etc.</td>
</tr>
</tbody>
</table>

Figure 1 shows an information network graph for PID-related information flow among the various business actions. Each node represents a class of business actions (e.g. PID service provision or provision of institutional information repository services); a link represents a flow of information objects (PIDs, metadata, or content). The nodes shown correspond to currently active actions; the links correspond to flows that are fairly well established as of 2017. Several actions may be performed by the same actor; e.g. a national library may provide PID services, run a web archive and perform long-term content management. A link implies that there is an information flow between some of the relevant actions covering some of the information objects. Reviewing this representation highlights links that are desirable for long-term preservation of the scholarly record, but may not be sufficient, or even exist, at the moment.

Where links are not well established, it is not possible for an actor to obtain information reliably or without substantial effort. Often links exist, but they do not pass along all of the valuable metadata that is associated with a PID. For example, in order for a library to acquire the information it needs about a dataset, it may be necessary to retrieve one subset from a data center, and another from a PID service provider.

A question is whether the information network has the links that are desirable in practice. For example, currently, there are no well-established links that inform Libraries and Archives about metadata updates in external PID Services, or that enables researchers to export PID-related metadata from Libraries and Archives into Citation Managers, or that pass rights metadata from PID Services to Researchers or from Data Centers to Libraries and Archives.

Even in the case that there are good links in the information network, it is not always clear who is responsible for assuring the long-term usability of an Information Object. To answer this, we need to know how each organization type can best contribute to
the assurance for long-term maintenance of the scholarly record; determine where in the life cycle it is easiest or most effective to create PIDs; determine where PIDs should be enriched with metadata; and when to perform other essential Actions.  

2.1.2 Longevity of organizations  
The stakeholders should best be responsible for each Action. One step in the roadmap is to investigate these rules and which of identified entity and the more doubt there is on the provenance of more contradictory assertions may have accumulated about an access. The later in the life-cycle a stakeholder is positioned, the more trust exists in its continued ability to support the information. The sooner in the life-cycle good housekeeping applies (such as assigning PIDs), the easier it is to avoid violations of the information on the way. The later in the life-cycle a stakeholder is positioned, the more doubt there is on the provenance of metadata.

One step in the roadmap is to investigate these rules and which of the stakeholders should best be responsible for each Action / Information Object pair. For example, what metadata should be created by PID service providers to support long-term preservation? What content or metadata format transformations should data centers perform to support long-term preservation?

2.1.3 Handover  
There are many pragmatic questions that must be addressed to ensure that the handover of Information Objects between stakeholders is effective. These include:

- **Who takes responsibility for minting PIDs?**

- **Where should business-as-usual handover of any information related to the scholarly record be initiated? Should handovers be defined and governed in a systematic manner? How can business models of different organization types define a hand-off best practice, as we are already familiar with from, for example, the hand-over between records management systems to archives? How is responsibility to be transferred technically?**

- **What should happen when preservation trigger events occur? As a proactive example, online digital research repository Figsshare has joined the Digital Preservation Network (DPN). Their announcement states that “Research data made public on Figsshare will be deposited into DPN, a dark archive that preserves scholarship for future generations. Figsshare users can guarantee that long-term access to their scholarly resources will be protected in the event of any type of change in administrative or physical institutional environments.” Similarly, some journal publishers subscribe to the CLOCKSS dark archiving system to protect against the case that they may no longer be able to make their outputs available. A more reactive example of a sustainable governance migration is the handover of the PURL PID service management from OCLC to the Internet Archive, which has a declared long-term business model. In the event of organizational failure, there should be mechanisms in place for both the identifiers and the identified entities with a governance structure and a method to offer resolution and access services.**

- **Who should be responsible for aggregating the distributed scholarly record? PID services collect metadata on identified and related entities. This may give the impression that they guarantee the long-term availability of the scholarly record. But PID services have limited scope; their mandate only extends to the persistence of the identifiers, discovery metadata, and resolution services. For example ORCID’s main goal is to provide PIDs for researchers. ORCID also collects information about these researchers’ scholarly output such as alternative person identifiers, histories, funding, patents, and associated works. But it is not clear that it is ORCID’s responsibility to guarantee this metadata for the long-term. Where is the right scope for each of these PID services?**

Considering the information network and the varying long-term commitment of the stakeholders in the scholarly record, there are two crucial questions to address:

- **How should the handover of Information Objects between stakeholders be governed and managed?**

---

[12] https://www.oclc.org/articles/2769
[16] https://archive.org/
[18] https://orcid.org/
• Conversely, how do we ensure that the PID graph is fully connected where possible so that there are no unintended islands of information?
• What is the role and responsibility of a national library in preservation of globally distributed metadata that is associated with various PIDs across multiple independent providers?
• Metadata associated with content is often deleted upon handover between stakeholders. This may be for very good reasons. For example, when an image is shared on the web, one may wish to remove identifying information to comply with data protection regulations. If there are multiple copies of some content that is identified by a PID, there may be different metadata associated with each copy; how can a digital object consumer identify which copy holds the metadata they need or are entitled to?

2.1.4 PID minting responsibility. Organizations can only mint PIDs if they make a long-term commitment to enable access to the identified content. But there is valuable scholarly content, for example in the form of blogs, that has no dedicated long-term champion. Webarchives in public and private content repositories and in national libraries and archives can provide persistence for some of these digital assets that do not have owners who can commit to their long-term accessibility. An interesting proposal by Zierau et al [3, 4] bases these assets’ persistent identification on the persistent identification of the web archive itself. A PWID consisting of an identifier for the web archive, the harvest date-time, the harvested URL, and the context specification permits persistent global identification of any harvested web content with the guarantee of permanence provided by the webarchive, rather than the content originator.

2.2 Managing distributed long-term responsibility

No single organization today holds even a copy of the full scholarly record including PIDs, associated metadata, and the identified content, much less holds an authoritative copy. While concentration of information can simplify large-scale use, it may increase the risk of large-scale loss. Distribution and redundancy offer a form of resilience and improved availability. Given this decentralization of stewardship, how can one manage the distributed long-term responsibility? How can one avoid discontinuities in modeling and interfaces, to ensure interoperability at the edges of organizations’ scope?

The Scholix framework offers a conceptual model, an information model, information standards and encoding guidelines, and options for exchange protocols toward solving interoperability issues. It is “a high level interoperability framework for exchanging information about the links between scholarly literature and data”. But there is also a need for technical, governance, and coordination solutions, in particular for providing long-term availability of the scientific record without requiring a central uniform repository.

2.3 How can PIDs best be used to support entities that evolve over time?

The scholarly life-cycle involves entities that are evolving over time. There may be changes to metadata or content. When, for example, datasets and software identified by the PID service change or related versions are created, one must track how each version relates to earlier ones. Each version may be identified by a new PID and linked through meaningful relationship types. Whether or not changes establish a substantially different object that deserves the assignment of a new PID depends on the use case that is supported by it, and on the policies that underline the use case. For example, changing the spelling on an author’s name may be considered a minor correction that does not necessitate the assignment of a new PID to a book. There is no use case that would handle the corrected object differently. Even so, one can keep a cumulative trace of any corrections. Adding an author to a book may necessitate the creation of a newly identified object, because it may be necessary to reflect the fact that different copyright assumptions were made before and after the correction.

That is to say, there is a use case that results in different actions on the two identified objects. Therefore, one should distinguish the corrected object from the earlier one through a new PID and one should record the relationship between them, as well as the event and the policy that necessitated the creation of the new PID. As a consequence, it is not the PID service that determines at what level of granularity PIDs should be assigned. PIDs support the clients’ use cases. The PID services have to flexibly accommodate different client policies and use cases. But this also implies that PID service providers have little control over the resulting granularity of the research objects that are identified. It is then the PID minting clients that can negotiate among each other to establish guidelines on policies that both support the implemented use cases and support interoperability and information exchange between different institutions.

18 http://www.scholix.org/
These dynamics don’t only apply to defining what use cases trigger versioning, but also to controlled vocabularies or the granularity at which PIDs are assigned to digital objects. For example, different even types necessitate the creation of a new PID for a derivative dataset. To support long-term management of evolving research data one would want to record the dataset’s provenance by recording the relevant events. These events types are typically defined by a controlled vocabulary, such as “software patch applied”, “time-filter applied”. The PID service can define suggested controlled vocabularies, but the client must be able to use their own personalized vocabulary to meet their individual use cases.

One of the key R&D questions is, therefore, what functionality needs to be provided by PID services to enable their clients to capture the necessary versioning information about evolving entities. Initial discussions can be found in [5].

Memory institutions are practiced in dealing with these questions in the context of their digital repositories. Digital preservation metadata work, as discussed in Section 4 has provided recommendations for how to handle these situations that now can be applied to new contexts, such as PID services.

3 MEMORY INSTITUTIONS - PIDs HELP LONG-TERM PRESERVATION

Using PIDs can improve processes for institutions that need to satisfy a long-term mandate.

3.1 Authority control

Wikipedia [6] states that in “library science, authority control is a process that organizes bibliographic information, for example in library catalogs by using a single, distinct spelling of a name (heading) or a numeric identifier for each topic. … These one-of-a-kind headings or identifiers are applied consistently throughout catalogs which make use of the respective authority file, and are applied for other methods of organizing data such as linkages and cross references.” Authority control supports information management and is shifting toward PID-based rather than string-based solutions. This is a significant change in the information management practices in memory institutions that helps to avoid shortfalls of string-based authority practices, such as spelling errors. For example, using an ISNI for current or historic rights can define suggested controlled vocabularies, but the client must be able to use their own personalized vocabulary to meet their individual use cases.

One of the key R&D questions is, therefore, what functionality needs to be provided by PID services to enable their clients to capture the necessary versioning information about evolving entities. Initial discussions can be found in [5].

Memory institutions are practiced in dealing with these questions in the context of their digital repositories. Digital preservation metadata work, as discussed in Section 4 has provided recommendations for how to handle these situations that now can be applied to new contexts, such as PID services.

3.2 Infrastructure for preservation of the digital scholarly record

Digital preservation is a form of long-term information management. Institutional repositories can provide local storage and archiving for scholarly outputs and they are equipped to manage metadata and some content. Realistically, the ability to preserve the wide array of research object types, such as non-SQL databases, is limited and needs further work. Furthermore, repositories are not equipped to provide content resolution services to replace failed PID services. The practice in PID services can inform improvements in digital preservation services in this regard. And, as mentioned earlier, repositories have limitations in managing PID related metadata when they are not the organization on record and known to the service that minted the PID.

If memory institutions harvest content from the web, PIDs can be very helpful, but still have implementation inconsistencies that prevent effective automated harvesting. We need to improve managing long-term data and artifacts that include PIDs so that they can be used to better streamline digital preservation efforts. For an example, see Van de Sompel, Rosenthal, Nelson’s [7] discussion on eJournal preservation.

3.3 PID use in digital preservation repositories

In digital preservation, identification of digital content is essential. In the contemporary scholarly process, PIDs are used for validation and reuse of research results. Long-term reuse of material that is held by memory institutions is even more challenging, since the material is created and consumed by third parties. Reuse happens over much longer periods of time. As a result assumptions about the environment and context required are less likely to hold. This makes PID use even more important. But by far not all content in digital long-term repositories is identified through PIDs. Most digital repositories deal with a variety of identifiers, most of which lack one or more of the PID criteria outlined above. It is essential to understand how long-term repositories can enhance existing local or transient identifiers within their scope to support the PID criteria.

PREMIS states [8] that “for a given identifier to be usable, it is necessary to know the identifier scheme and the namespace in which it is unique. If a particular repository uses only one type of identifier, the repository would not need to record the scheme in association with each object. The repository would, however, need to know this information and to be able to supply it when exchanging metadata with other repositories.” This requirement only ensures that an entity is identifiable within one repository. To support a global information network, it would require very precise knowledge about when a scheme was applied, how the scheme changed over time, how the versions relate, and so on. As Information Objects flow through the information network, chances are that a long-term repository would not be able to collect the essential information in external schemes to establish identifiability. The question to address is how PID use can support this information flow.
Another aspect of long-term information management is a need for heightened resilience. The longer-lived content is, the more likely it is that parts of the information object may inadvertently be corrupted. For example, we have anecdotally witnessed that the links between PIDs and their associated metadata have been broken through software programming errors so that PIDs could no longer be linked to the content they identified. In order to mitigate this sort of risk it is advisable to not solely rely on PIDs, but to judgmentally enrich them with redundant metadata that would permit a semantic match of entities between distributed long-term systems if the PIDs themselves get corrupted. The questions that arise are, what metadata is best suited to serve this purpose, and how best to ensure synchronization of redundant metadata in multiple places.

3.4 Other archival tasks that can benefit from PID use

Many more entities are managed in memory institutions’ information governance. Ideally there would not only be PID services for research outputs and agents (persons and organizations), but also for funders, grants, laws, patents, software packages, events, etc. Every PID service requires a governance structure, a metadata scheme, and support for information creation and exchange. These are currently missing for many entity types. Memory institutions are now also creating new and derived data sets related to their digital collections through text and data mining, analysis, crowd sourcing, and citizen science. Stable ways of identifying these data sets, their provenance and their contributors need to be implemented.

4 PID SERVICES - LONG-TERM PRESERVATION HELPS TO SHAPE PID PRACTICE

The digital preservation community has developed an array of practices and techniques for long-term information management. How can this help to shape PID best practice and ensure long-term access to the scholarly record? Where are gaps in PID services’ current practice that can be informed by memory institutions’ practice?

A key answer is the metadata that is associated with PIDs and the entities that are identified by them. PREMIS [8] is the de facto metadata standard for long-term access to digital content. It recommends the information about digital content that is very likely needed for long-term use and preservation. Its goal is to ensure the availability, identifiability, integrity, viability, renderability, understandability and authenticity of digital content. PREMIS articulates data modeling and metadata principles that should be adopted early to remove vulnerabilities for digital content and to ensure its long-term usability. But PREMIS does not just address digital preservation. A file can become unreadable because its format has become obsolete over years or because there was a power failure during file transfer today – both situations require similar consideration. Therefore, the approach taken in PREMIS can be helpful for near-term management of digital content.

Figure 1 illustrated how metadata that is associated with the scholarly record is passed through a network of stakeholders’ actions until it ends up in memory institutions for long-term content management. The stakeholders involved in the scholarly information network can benefit from those PREMIS principles. For example, PID services, such as CrossRef or DataCite, are justified in specializing on content types, such as scientific articles, monographs, or data sets. As their portfolios grow to accommodate newly supported types, their data models need to be extended. The first-principles approach taken in PREMIS helps to ensure that data models are both extensible and interoperable. Since PREMIS is expressed as framework and principles it can be implemented in any implementation environment.

Future work should articulate those recommended PREMIS features, perform a gap analysis to understand to what degree they are currently not supported with scholarly stakeholders, what implications this has for the long-term access to the scholarly record, and how these short-comings could be overcome.

An initial set of example recommendations of how stakeholders in the scholarly information network can adopt good practice from PREMIS is discussed in the following.

PREMIS distinguishes Objects, Actors (including, people, organizations and computing components), Rights statements and Events, which are sufficient to support modeling most semantic relationships required. Objects can be described on four separate, declared levels:

- the intellectual description that helps search and find the object;
- representations, which are sets of files that together create one rendition or execution of a digital object;
- component files of each representation;
- and individual bitstreams.

Efforts that model digital objects without distinguishing these levels often end up confusing issues that should be separated. For example, some PID solutions conflate resolution to a landing page that holds descriptive information (similar to the intellectual entity) with resolution to the content itself (which could be a representation, file, or bitstream). This lack of clarity hinders automatic crawls, machine-harvesting and machine-interpretation of parts of the scholarly record. Van de Sompel et al. [9, 10] analyze this situation for web use of PIDs. Adopting the PREMIS model would resolve this challenge. Different users may need to resolve to different levels. For example, a crawl bot may just want to find representations and files; a search indexer may just be interested in intellectual entities and descriptive metadata; someone running an impact analysis may just want to identify the researchers and, from there, link to their research outputs. Because of these varying use cases and goals the underlying conceptualizations need to clearly identify the types of entities of interest. This results in access mechanisms that support these use cases flexibly. PREMIS entities are sufficient to support modeling most semantic relationships required to capture these distinctions. Technical metadata, as specified for PREMIS objects, should be created as early as possible in the information network (e.g.,
including information about file types, checksums, creating software, or computing platform requirements).

The PREMIS model covers derived, dependent or structurally related objects and provides a transparent way of relating them with each other. Each relationship can document the nature of the relationship, and the events and agents that were involved in creating it. As we have seen in Section 2.3, this ability is important to support use cases for many of the actors that rely on PID services.

PREMIS provides two powerful tools for identifying and describing partial and dynamic datasets. The first one is the ability to describe and identify bitstreams (mentioned above). If data fragments can be described as sets of bitstreams, they can be directly identified and described.

Sometimes they are, however, better described by the event that created a dynamically derived subset (e.g., for data sets that are derived from a base data set by applying a filter). This can be addressed by events associated with relationships. The derived dataset can be identified by annotating its relationship to the original data set with an event that captures the selection criteria or selection algorithm. The derived data set is identified, but does not actually need to be instantiated; it can be computed on demand from the original data set. In this way PREMIS modeling can be used to support the implementation of the RDA recommendations for Data Citation of Evolving Data [11] within its basic inherent data model and without the need for creating any extended functionality for special cases.

Relationships also let you relate data sets that are related through a structural hierarchical inclusion relationship on various levels of granularity. PREMIS objects can be used to describe software and hardware or other parts of the computation stack, which are essential components in the scholarly record that are currently not adequately covered by PID services. Adopting PREMIS compatible conceptualizations would make it easier to provide PID support to these object types and to support interoperability among scholarly stakeholders.

Another example is the maintenance of provenance information in the scholarly record. Provenance may contain events from any point in the object life cycle. For example, they might record due diligence activities, or events that transform an object or its metadata. It is important to record these events to determine the degree of authenticity of the object over time.

PREMIS events also encourage linking to event-related agents, ranging from data creators to researchers, curators, to publishers, but also organizations or software agents. This form of modeling does not simply identify the role of an agent with respect to an object, as is currently frequently done, but it specifies in which event this role was taken. This allows for a much more precise and time-linked recording of the agents’ roles.

CrossRef and DataCite are starting to collect and distribute information about events related to PIDs. For example, when an article cites a dataset or when a new version of a dataset is released. Currently, this is not based on a generalizable event model, and it does not extend to provenance events. The responsibility for this sort of information may mainly lie with their clients, but PID service providers should consider collecting it as additional assurance. Again, adopting the PREMIS modeling style for events (or agents or rights) can improve interoperability across the information network and simplify information exchange.

These examples illustrate some of the PREMIS features whose adoption could improve the information exchange in the scholarly information network. Future work may investigate how these principles can best be translated to individual implementation environments.

5 CONCLUSIONS

We set out to investigate how Persistent Identifier services can be extended to long-term information management. When starting to address this, we realized that the number of unsolved issues was substantial and should not be addressed ad hoc. A more systematic analysis of the research and development space was required to combine lessons from the two domains. Consequently, this paper sets out key questions and challenges for a roadmap to improve the alignment between Persistent Identifier and Digital Preservation approaches. Due to the real advances that have been made in each community, this alignment may enable a more deliberate design rather than stepping through ad-hoc improvements. Enhanced aligned services involved in creating and maintaining the scholarly record would support a more complete information flow. The resulting data models would explicitly support long-term preservation of PID resolution services, metadata that captures information about the scholarly record, as well as access to the actual research objects.

ACKNOWLEDGEMENT

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 654039”. We would like to thank our reviewers Herbert Van de Sompel, Ignasi Labastida, and Barbara Sierman for suggesting work in this area.
REFERENCES


Life and Death of an Information Package: Implementing the Lifecycle in a Multi-Purpose Preservation System

Bertrand Caron
Department of Metadata
Bibliothèque nationale de France
Paris, France
bertrand.caron@bnf.fr

Jordan de La Houssaye
Department of Information Technology
Bibliothèque nationale de France
Paris, France
jordan.de-la-houssaye@bnf.fr

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

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

ABSTRACT
This paper aims to explain how the National Library of France (BNF) faced the question of the information lifecycle, during the implementation of its digital preservation system, in the light of the experience it acquired managing and using this system, from a theoretical approach following the OAIS reference model to its implementation. It was understood very early that from a preservation point of view the management of subsequent Versions and Editions of an Information Package was a particularly sensitive problem akin to risks management. However, it was thought at the time that the system should be able to manage the lifecycle of an AIP (Archival Information Package) in a universal way that would be valid whatever the package considered and the context. These ideas quickly proved to be impractical and didn’t resist the test of reality. Some improvements were provided, while taking into account the rest of the digital ecosystem of the BnF.

KEYWORDS

1 INTRODUCTION
Among the challenges an institution has to face when it puts into practice digital preservation, the management of updates of data isn’t the easiest to face. It is necessary to conciliate precautions to avoid the loss of information, requirements for enabling enrichment of data, and financial sustainability of storage. The lifecycle of information has to be considered through the prism of risk management, therefore no turnkey solution is available for those who have to deal with this issue.

The BnF (National Library of France) is assigned with the mission of preserving over the long-term a part of the French cultural heritage, including in its digital forms. It has been building since 2007 its own digital preservation system. The issue of the packages lifecycle had to be taken into consideration.

In this paper we reflect on the history of digital preservation at the BnF in order to account for successive refinements that were made to our original assumptions regarding the lifecycle. In the first part, we detail the way the OAIS reference model was interpreted in order to make implementation choices in the SPAR system. In the second part we detail chronologically how the system originally managed any updates of AIPs, and how new use cases the BnF faced forced us to adapt our preservation system. In the third part we reflect on the way these improvements changed the lifecycle of AIPs in SPAR as well as the broader preservation environment at the BnF.

2 THINKING ABOUT A LIFECYCLE: WHERE, WHY, HOW?

2.1 The Context: SPAR and its Ecosystem
The BnF preservation system is named SPAR (Scalable Preservation and Archiving Repository). It is being built in house since 2007 and is in production since 2010. Its scope is to manage all entities that can be automated through modules corresponding to the OAIS entities. The system includes the management of the storage infrastructure.

In SPAR, the sets of documents to be ingested are processed by tracks and channels (sub-tracks), according to their nature (e.g., digitized books, audiovisual files, web archives, administrative records), their legal framework, and the way the BnF plans to apply preservation strategies. At the present time, SPAR can ingest objects through six tracks: digitized documents and associated files, audiovisual objects, legal deposit (ARC or WARC files), negotiated legal deposit (ebooks and other born-digital documents), administrative records, acquisitions and donations (born-digital documents out of the scope of the legal deposit), and third-party archiving (various kinds of files, from partners outside the institution).

Each track has a leader (named a ‘track manager’), someone in the library, generally a librarian, who is more than a representative. This person is responsible for preserving the collections, and is in charge of conversations with the IT department. The result of the negotiations is formalized, for each channel of the track, in Service Level Agreements (SLAs) that formally rule the interaction between the Producer, the Archive and the Consumer (see Figure 1). More precisely, they define the terms of ingest, preservation and dissemination (e.g., formats accepted, maximum size of
packages, availability of service, number of copies etc). Each SLA is transcribed in XML files that configure the system.

As of March, 2017, more than 7 million Archival Information Packages (AIPs) are preserved in the system, containing about 300 million files, for a volume of 3 Petabytes.

2.2 Original Principles of the Implementation of OAIS Lifecycle Concepts

The SPAR system is strongly inspired, from the beginning, by OAIS [1], and its implementation follows closely the principles of this standard. Therefore, not only the responsibilities, the data model, the entities and their functions were supposed to follow the reference model (part 3 and part 4), but also the practices described in the standard (part 5, Preservation perspectives).

Among the practices that the system had to take into account, was the answer to the question: what to do when an update of a package is proposed for ingestion? Previously to SPAR, updates were managed on a case-by-case basis, and generally the new digitized or digital document replaced the older one, that was erased on the storage capacity or became inaccessible. Updates didn’t happen very often, but we foresaw that over time it would turn into one of the usual cases: quality improvement on the Producer’s initiative (especially in digitization contexts), format migration for preservation purpose... The preservation system had to implement a lifecycle in order to prevent the loss of data and to enable the management of digital document at a large scale.

For this purpose, the OAIS standard offers some valuable concepts. Section 5.1 (and more precisely 5.1.3) distinguishes four types of Digital Migration that can affect the Archival Information Package (AIP) for preservation purposes. Refreshment and Replication are operations that do not affect the Content Information, the Packaging Information or the Preservation Description Information (PDI) of the package; they are directly handled by the Storage layer [6] and are beyond the scope of this paper. Repackaging implies changes to the Packaging Information, and Transformation implies changes to the Content Information and/or to the PDI; both cases enter in the field of the lifecycle of the Information Package2. The standard then specifies what the Archive is supposed to do (5.1.3.5): only in case of a Transformation, the Archive has to create a new AIP Version, i.e. an AIP that “is a candidate to replace the source AIP” (1.7.2). Concerning what to do with the replaced AIP, there is no constraining rule: “The first version of the AIP is referred to as the original AIP and may be retained for verification of information preservation.” (5.1.3.4).

In case of an upgrade or an improvement of the AIP, at the Producer’s initiative, the OAIS standard specifies that “This is not a Digital Migration in that the intent is not to preserve information, but to increase or improve it.” Thus, a new AIP Edition has to be created, but once more the standard leaves the choice to the Archive to retain or not the previous AIP (5.1.3.5). At the BnF, it has been understood that these choices were a matter of risk management regarding data loss.

Because implementation choices also had to be done, the interpretation for an AIP Edition was: increasing or improving information means adding Data Objects (for example: supplement image mode digitization with OCR files) or adding/modifyinng metadata (such as Descriptive Information, PDI...). In every case, there is no necessity to preserve the previous AIP, because there are no risks of losing data, and because the SPAR system includes a lot of functionalities to avoid the ingestion of packages with metadata too poor for preservation. The understanding for an AIP Version was: modifying Data Objects can be thought of as a Transformation, and therefore leads to a new Version of the AIP. In this case, because Data Objects were changed, it has been considered by default that deleting the previous AIP Version was a risk: as the OAIS says that “the new AIP is viewed as a replacement for the source AIP” (5.1.3.5), it would be too big a responsibility for the Archive to remove the source AIP. Therefore a general preservation rule was defined: the first, the latest and the penultimate Version of a package has to be preserved (“0, N-1 & N Rule”, see Figure 2), while a new Edition always removes the latest one.

As OAIS requires (sub-section 4.2.1.4.2), this part of the history of the package is recorded as Provenance Information. Moreover, every instance of the AIP is addressable at the level of the Reference Information, through its identifier: the ARK [5] identifier is systematically suffixed with two qualifiers, one to mark the Version and the other to mark the Edition. Initially, it appears in the form of version0.release8. When an update occurs, the number of version or release, following the case, is increased for the new package; for example, if the first update produces a new version, the qualifiers for the new package are version1.release8. From a preservation point of view, this enables to apply automatically the “0, N-1 & N rule”. From an Access point of view, this enables to establish a simple and global policy: if the Consumer asks for an AIP in absolute (i.e. only the ARK identifier without qualifiers), the system retrieves the last Version and Edition; if he asks for a particular Version or Edition (i.e. specifying qualifiers), the system retrieves the requested Version or Edition if it still exists.

In order to document and preserve PDI, Representation Information and Packaging Information, the BnF initially chose the METS format. A single METS file, called “manifest”, is used as a wrapper for descriptive, technical and provenance metadata. For provenance

1Unless otherwise stated, all the references in this section are from the OAIS.

2Repackaging has not been on the agenda for the BnF yet, given the youth of the system.
metadata, PREMIS was chosen to document operations that affected the Content Information before and after ingestion. Among these operations are particular events occurring in the package lifecycle: SIP creation, ingestion and package update. When a package update is processed, the whole set of events which affects the Information Content from its creation is preserved as the package audit trail.

This general policy, already defined in 2008 [3], can be viewed as a simplification regarding what the reference model offers, but it presents the advantages of enabling an automated implementation.

3 THE LIFECYCLE ACTUAL IMPLEMENTATION AND ITS EVOLUTION

3.1 Initial Implementation

During the initial development of the SPAR system, a decision tree was implemented during the ingestion phase, in order to assess automatically if a package proposed for an update had to lead to a new Version, a new Edition or a failure. A quite complex sequence diagram was specified to describe SPAR’s decision making, which occurs in an internal process of the ingestion phase named “ACT_10” (see Figure 3). This sequence combines the SIP manifest analysis and comparisons between the AIP and the SIP manifest. Some steps were a transcription of a simple rule (for example, if there is no fileSec in the SIP METS manifest, then it is only a metadata modification, and the system has to create a new Edition); some implied complex comparisons, should the SIP contain Data Objects.

Regarding the "0, N-1 & N" rule, it has not been implemented, so that every Version is preserved in the system. As for the Edition, the system is able to really remove the previous one.

As the system is in production since 2010, with the first track “Digitized documents for preservation”, the BnF faced several use cases. That led to reconsidering the whole package lifecycle, not only in the SPAR system but also in the broader environment of digital management in the library.

3.2 A First Exception to the Automated Lifecycle: Requesting Explicitly for a New Version

In the context of mass-digitization, it often happens that a defect is detected after ingestion, or that digitized documents are supplemented with OCR, which causes a reprocessing of the package. It appeared that the automated mechanism was not sufficient to address with absolute confidence the different use cases, because of their complexity and variety. One use case is when, at the same time, the update consists in modifying the existing Data Objects and adding new ones, for example when a book digitized in black and white is newly digitized in color and OCRed. The system was then supposed to create at the same time a new Edition and a new Version, which means that the automated implementation of the lifecycle was not complete. At this point, when this case occurred, a risk existed that the system would create a new Edition and delete
the previous Data Objects, while it was not certain that this was what the Producer wanted.

It was therefore decided that the decision to create or not a new version had to be taken outside of the preservation system, earlier in the workflow. A specific request functionality for a new Version was implemented, coming explicitly from the outer "delivery chain for digitization", that knows the context of the delivery or the update better than the preservation system. This request is submitted to SPAR by means of a PREMIS updateRequest Event, which keeps track of the request in the manifest after the update is processed.

It is interesting to note that in the context of mass digitization at the BnF, every subsequent digitization leads to a new Version of the target AIP. From an OAIS perspective, this suggests that a digitization is a kind of migration, and indeed, it can be argued that digitizing a book for example is a media migration from physical to digital. Consequently, each subsequent digitization can be thought as a new migration, albeit from the same original source, and therefore has to lead to a new Version of the corresponding AIP.

3.3 Requesting a Deletion of the AIP

The second evolution came directly from the integration of Administrative records. Indeed, handling archival processes forces taking into account the elimination of the content of a package. Regulation rules or laws make such destruction of information mandatory. However, since SPAR gives a permanent identifier to each package, it couldn’t just erase the record: the elimination is thought as an update. In SPAR, updating a package requires a re-ingestion, in
order to apply the policy defined in the SLAs (as the OAIS specifies that "AIPs should never be deleted unless allowed as part of an approved policy; there should be no ad-hoc deletions"). Here, this update will replace the actual content with a certificate of elimination, called "tombstone". This file contains the reason for the elimination as well as the requester and the authorizer. Doing so enables an appropriate response in case someone wants to retrieve this package.

Similarly to an update request, this request is submitted to SPAR by means of a PREMIS deletionRequest Event. In fact, the act of elimination is viewed as an update of the package with a new content (the 'tombstone') that will replace the latest one while erasing all the previous Versions of the package. This is the first time the system accepts that the original Version (Version 0) can be deleted. Of course, in order to keep such an operation safe, the permission of doing it must be given in the ingest SLA as well as who can initiate such a task and in which conditions.

At the end of the operations, a unique version is kept, the latest, containing the tombstone but also, in its manifest, all the audit trail of the operations applied to the package. No one can retrieve any previous content but anyone can be informed as to why it has been discarded.

It is worth noting that the result of a deletionRequest is a new Version of the AIP, whose content is the so-called "tombstone". It could be argued that it should be a new Edition, and not a new Version, because its Content Information has been upgraded (albeit not to improve it but to destroy it). However it has been understood as a Transformation (and therefore a Migration) because the Content Information has been altered and the resulting AIP is meant to replace the source AIP.

3.4 Requesting a Deletion with Redirection of the AIP

The previous evolution was fine from an archival perspective but, seen from a preservation one, losing the information content is not acceptable. In the context of the digitization programs, the same analog item might have been digitized by different means (old black and white digitization vs. new high resolution color one) and with different identifiers, generating sorts of duplicates. When keeping both is seen as unneeded, choosing the "best" one cannot be an automated process but is directly related to the management of the collection: the weeding procedure. In order to make it possible, the system was enhanced by adding to the deletionRequest the obligation of specifying another package. During the process of selection, the collection manager generates a tombstone not only stating the reasons for his decision but also providing the alternate package holding the same informational content. During the ingestion of this tombstone, the existence of the substitution package is verified. Then, in case someone wants to access the discarded package, the system will automatically redirect its request to the associated one. Once again the linear lifecycle is here clearly eluded but following a clear and stated decision.

This functionality came from the needs of digitization, but turned out to be useful for other tracks, such as Web legal deposit, to also address the case of a Producer’s mistake that causes a duplicate and that is detected after the ingestion of both packages. If the duplication is intentional, then the Producer can use the same identifier for both packages in order to create a new Edition of the package. In fact, this case matches exactly what OAIS calls “improvement” of information.

3.5 Requesting a New Edition Explicitly

Following the updateRequest, a new functionality was then implemented to address the case of data enrichment, for example when OCR files are added to still images. Here, new data and metadata are delivered simultaneously, and the automatic detection of a new Edition turns out to be very difficult. The explicit decision to create a new Edition (in PREMIS terms: replacementRequest) is now possible on the track manager’s initiative. For consistency reasons, this functionality was implemented on the model of the updateRequest (see 3.2).

3.6 Requesting a Channel Switch

Given that every channel is ruled by SLAs defining ingestion, preservation and dissemination policies (see 2.1), the need for changing those policies for a package implies changing its channel. A channelSwitchRequest has been developed, resulting in a new Edition of the package in the target channel. This functionality can be viewed as a mix of a replacementRequest (as it is an explicit request for a new Edition) and a deletionRequest (as it also requires authorization, explanation and documentation). The Channel Switch is far from being a simple operation performed by an administrator; on the contrary, it means a new ingestion into the system, so it has to be allowed at the same time by the SLAs of the source channel and of the target channel. Switching a package from one channel to another only means moving it and not modifying it, consequently this operation must not imply a doubling of the storage size. That’s why it results in a new Edition of the package.

While working on this matter, it was deemed beneficial to expand the possibility to document such an operation. When the switch is defined, it is now mandatory to indicate either the reference of a BnF internal document or the identifier of a preservation plan also preserved in the system [4], that explains the decision of switching the package or a set of packages from one channel to another. In addition, in an iterative approach, this constraint has been extended to the deletionRequest functionality.

Almost ten years after its first version, the sequence diagram of the "ACT_10" process has been substantially enriched (see Figure 4). Now, the first steps involve detecting one of the four PREMIS Events that explicitly lead to a new Version or a new Edition of the package. Only after these steps begins the analysis from 2007, that has been simplified because some of the steps were no longer useful after implementation of the new functionalities. Thus, the risk that the system creates an unwanted Edition, is significantly reduced.

4 ENRICHMENT AND ENHANCEMENT OF DIGITAL CONTENT: LIFE BEYOND INGESTION

After having produced digital contents for more than twenty years, the BnF had to take into account the fact that its digitization policy is no longer limited to the creation of new digital copies of analog...
Figure 4: Simplified view of the decision tree to process a SIP for update (2017 version)

Material, but also aims at enriching or replacing older digitization and adding new derived products. As a consequence, the package lifecycle, at first designed to be linear, is becoming circular (see Figure 5): Information Packages subjected to enrichment or enhancement have to be disseminated, reprocessed or enriched, and then ingested again. Services and systems which until then had no conscience of their role in digital information preservation (e.g., QA delivery chains, dissemination services, etc.) had to endorse responsibility for actions that would affect the Content Information and its quality. These services were affected by this new principle and had to evolve accordingly in order to ensure that no loss could affect the packages fixity and quality.

4.1 Dissemination
The act of disseminating an IP with the intention of enhancing some of its components or enriching it with new representations consequently appeared to be a critical phase in the package lifecycle. Particularly, the context (date, reason, agents involved) had to be taken into account to determine the systems behavior when the SIP expected to update the corresponding AIP will be submitted. A new PREMIS Event, named disseminationCompletion, has been
chosen to record the dissemination operation, which created a DIP delivered to a Consumer\(^5\). The responsibility for disseminating the package and for recording this PREMIS Event had to be endorsed by services aware of the operation context and goals, that is, dissemination services directly in contact with Consumers.

4.2 Package Delivery

As the preservation system is unaware of the production context, specified by digitization contracts for example, the creation of a SIP for update had to be made before ingestion by services that have the understanding of the reason why an AIP had to be updated. The packageDelivery PREMIS Event that records the transfer of a producer package from a Producer to the BnF mentions the intended use of this producer package (replacement or addition of Data Objects). Only QA delivery chains, combining the intended use and the policy defined by track managers, can determine how to create a SIP and which kind of request (for a new Version, cf. 3.2 or a new Edition, cf. 3.5) should be submitted to SPAR. Hence, merging the Producer package and the corresponding AIP to produce a new SIP for update is performed by the delivery chains, which, unlike the preservation system, have an understanding of the update expected results. At the ingestion phase, SPAR just creates a new Version or Edition by taking all Data Objects from the SIP, though preserving the whole audit trail coming from the AIP.

4.3 Policy about Versions Retainment

As previously mentioned, no policy was globally defined concerning the choice whether to retain or not previous package Versions. Preservation experts considered that this choice had to be made by people particularly aware of digital collections management. In the case of heritage digitization for example, the following policy was adopted:

1. whenever a reprocessing due to quality failure in the course of an ongoing contract generated a new Producer package, delivery chains will request the preservation system for a new Version;
2. whereas, in the case of a new Producer package intended for enhancement of the existing AIP, delivery chains will request a new Edition.

At the end of the digitization contract, a general decision has to be made to define which policy should be adopted for previous

\(^5\)The Consumer of the DIP happens, in this case, to be the Producer of the SIP for update.
Versions, taking into account the causes that led to requesting a new Producer package.

4.4 Managing the Risk of Quality Loss

Though creating AIPs with several SIPs delivered progressively is a common problem that Archives have to face for several years⁶, the iterative process of Data Objects production, based on previously produced Data Objects (OCR files produced from image files, or accessible Daisy files produced from OCR files) became extremely tricky when both production processes had to be managed simultaneously.

Indeed, the risk of quality loss arose because Producer packages intended for updates could be delivered concurrently from different Producers, either to enrich or to enhance the original AIP. The decision was taken to reject a SIP for update intended for enhancement if the corresponding AIP has been disseminated for enrichment.

Moreover, in order to warn contract managers if some case of quality loss should arise, the delivery chains have to compare the last AIP Version or Edition and the Version or Edition of the DIP used to generate the Producer package. Whenever the last AIP Version/Edition is more recent than that of the DIP, QA would raise a warning flag and alert contract managers and track managers.

5 CONCLUSION

In the course of our daily operations, two trends emerged:

First, we tend to explicit more and more our intentions in the SIP: if a new Edition or a new Version is required, we state as much explicitly. It is the business of the track manager to decide if such or such an action to the data should lead to a new Version, a new Edition or an entirely new Package. This trend reached the point where we started to invoke directly the business intent into SPAR.

Second, we now acknowledge that the question of the management of subsequent Versions should be treated track by track, as part of the broader question of preservation policy. This question is once again in the hands of the track manager.

The enrichment of the AIP lifecycle leads to putting our track managers more and more at the center of our daily operations. We foresee the time when this question of the AIP lifecycle will be treated explicitly in the SLAs of our tracks, alongside the treatments associated with business related actions on the packages.

From an OAIS point of view, this trend is perfectly logical. We first needed to build a working system, and therefore we took some shortcuts in our OAIS implementation. 10 years in the making, both the system and the business matured. SPAR is now clearly understood as the technical part of our OAIS, completed by the BnF as an organization dedicated to preserving part of the French cultural heritage.

ACKNOWLEDGMENTS

The authors would like to thank all the members of the SPAR team for their commitment and their invaluable contributions, John Thompson for his proofreading, and Claire Caillet from the CNES as well as Claire Sibille from the SIAF for the fruitful talks about OAIS.

⁶In particular, the CCSDS standard PAIS (Producer-Archive Interface Specification) [2] addresses this need.

REFERENCES

Semi-automated Generation of Linked Data from Unstructured Bibliographic Data for Japanese Historical Rare Books

Natsuko Yoshiga
Graduate School of Science and Engineering,
Saga University
Saga 840-8502, Japan
natsukoy@cc.saga-u.ac.jp

Shin-ichi Tadaki
Graduate School of Science and Engineering,
Saga University
Saga 840-8502, Japan
tadaki@cc.saga-u.ac.jp

ABSTRACT
A large number of bibliographic data and images of Japanese historical rare books have been published on the Web. For constructing structured bibliographic data, such as in a format of Linked Data, the data providers need to extract structured information from notes of the bibliography, which are written in Japanese natural texts with domain-specific terms. These jobs have been usually performed by persons with special knowledge. In this paper, the authors propose a semi-automated method to convert natural texts in real bibliographic data into Linked Data. As a part of the method, a simple ontology of key elements (named entities) in bibliographic data is constructed along general bibliographic rules for historical heritage. The ontology also has a capability to describe relations between a book and its parts. This allows mechanical access to information such as a creator of a cover picture, preface and so on. Finally, a script creates connections from named entities to URIs which describe headings or glossaries provided online by public organizations.

CCS CONCEPTS
•Applied computing →Digital libraries and archives;
•Information systems →Semantic web description languages; Multilingual and cross-lingual retrieval;
•Computing methodologies →Ontology engineering;

KEYWORDS
unstructured data, named entity, Linked Data, ontology, conceptual model, bibliographic description

ACM Reference format:

1 INTRODUCTION
Historical rare books, published before Edo period (until 1868) in Japan, have been collected by various organizations worldwide and those bibliographic data have also been published on the Web. For instance, one of the biggest collections of historical rare books in Japan [1], containing about 555,300 sets of bibliographic data, has been available online since February 2017.

Most of conventional bibliographic data are in tabular form. Some important information for researchers, such as creators of cover pictures and book owners, are written in notes with Japanese natural texts using domain-specific terms. As a result, those data are human readable, but not structured. In other words, those data are not suitable for analyses by software applications.

Recently, some organizations [2, 3] have tried to transform conventional bibliographic data written in Japanese into Linked Data. These attempts require a large amount of cost because Japanese natural texts are converted into structured data by specialists familiar with domain-specific terms.

In this paper, we propose a semi-automated method for transforming bibliographic data, especially texts in notes, into Linked Data form. We call the method as semi-automated because the method partly requires inspection by human for improving accuracy. The method consists of the following steps: 1) extracting named entities which are keywords with semantic meanings in the data, 2) giving ontological structure to the data for showing the structure of a book, and 3) creating connections from the extracted named entities and values to other URIs such as headings created by NDLA [2], VIAF [4], AAT [5], DBpedia [6] and HuTime [7].

The proposed method is applied to an actual online collection: one of digital rare book archives at Saga University Library [8], Ichiba Naojiro Collection. Its bibliographic data (Ichiba) are originally described in tabular form (Table 1). The fields of Ichiba basically obey standard bibliographic rules proposed in [9–11]. Finally, the data are transformed into structured RDF format, which are accessible as a SPARQL endpoint [12].

There are preceding studies on automatic transformation from natural texts into Linked Data [13–15]. In addition to methods for word extraction employed in those studies, we use a novel concept to construct structural information by referring to standard bibliographic rules of historical rare books and conceptual models for cultural heritage resources which are recommended in CIDOC CRM [16], FRBR05 [17], EDM [18] and so on. These conceptual models are employed to represent relations between a book and named entities which are essential keywords for describing the historical rare books.

The proposed model connects a historical rare book with its contributors and establishing processes as structured bibliographic data. Thus, those structured bibliographic data help researchers to study historical rare books through background information. In addition, Linked Data enables the data to be shared through the Web.
### Table 1: An Example of Bibliographic Data in Ichiba (with Translation)

<table>
<thead>
<tr>
<th>Fields</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial number</td>
<td>45</td>
</tr>
<tr>
<td>Genre</td>
<td>5-1-(2)-(6)-(6)</td>
</tr>
<tr>
<td>Genre1</td>
<td>文学 (literature)</td>
</tr>
<tr>
<td>Genre2</td>
<td>国文 (Japanese literature)</td>
</tr>
<tr>
<td>Genre3</td>
<td>小説 (novels)</td>
</tr>
<tr>
<td>Genre4</td>
<td>(no data)</td>
</tr>
<tr>
<td>Genre5</td>
<td>酒造本 (Sharebon)</td>
</tr>
<tr>
<td>Pronunciation of the title</td>
<td>セイロウココロエグサ</td>
</tr>
<tr>
<td>Size</td>
<td>(no data)</td>
</tr>
<tr>
<td>Volume</td>
<td>(no data)</td>
</tr>
<tr>
<td>Authors and editors</td>
<td>(二二) 蓮菜山人作 (Houraisanjin the Second created)</td>
</tr>
<tr>
<td>Date of publication (the Japanese era)</td>
<td>安政四年三月 (Ansei 4, a preface was created)</td>
</tr>
<tr>
<td>Date of publication (the Christian Era)</td>
<td>1857</td>
</tr>
<tr>
<td>Printed or handwriting</td>
<td>仮 (handwriting)</td>
</tr>
<tr>
<td>Notes</td>
<td>安政四年三月 (Ansei 4, a preface was created)</td>
</tr>
<tr>
<td>Seal of ownership</td>
<td>1</td>
</tr>
<tr>
<td>Collections</td>
<td>市場 (Ichiba)</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>(no data)</td>
</tr>
</tbody>
</table>

### Table 2: Named Entity Classes and Examples of Instances in Ichiba

<table>
<thead>
<tr>
<th>Classes</th>
<th>Samples of Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>安政四年 (Ansei 4, 1857), 十月 (October)</td>
</tr>
<tr>
<td>Place</td>
<td>東京 (Edo), 京都 (Kyoto)</td>
</tr>
<tr>
<td>Person</td>
<td>蓮菜山人 (Houraisanjin), 紫文里谷 (Umebori Kokuga)</td>
</tr>
<tr>
<td>Role1</td>
<td>作 (authors), 画 (painters), 編 (editors)</td>
</tr>
<tr>
<td>Title</td>
<td>息子部屋 (Musukobeya), 蓮菜山人作 (Houraisanjin)</td>
</tr>
<tr>
<td>Genre</td>
<td>文学 (literature), 酒造本 (Sharebon)</td>
</tr>
<tr>
<td>TFB2</td>
<td>表紙 (covers), 序 (prefaces), 改 (endnotes)</td>
</tr>
<tr>
<td>Term2</td>
<td>改題本 (a retitling book), 復印 (reprints)</td>
</tr>
</tbody>
</table>

1. Forms of contributions.
2. TFB is a class of words which represent parts in a book. This class name comes from "terminology for bookbinding.
3. A class of technical terms except for ones of TFB and Role.

## 2 NAMED-ENTITY EXTRACTION AND ASSIGNMENT OF URI

Named entities [19] are characteristics of a book, such as publication dates and places, authors, roles, titles, genres and predefined technical terms (Table 2). Some are in standard fields of bibliographic data in tabular formats and others are described as natural texts in notes. The predefined technical terms consist of terms from terminology for bookbinding, types of publishers, and other domain-specific technical terms, which are defined obeying cataloging guidelines for libraries [10, 11]. Those named entities help researchers to explore related information.

For describing historical rare books, it is important to refer to some parts of the target book using a word in terminology for bookbinding. A bibliographic description for a historical rare book usually contains, in addition to its main text, historical information, such as creators of parts of the book, relations to other books, owners, publishers. Such information referring to parts of the book with words in terminology for bookbinding is very important for researchers to study origins and establishing processes of the book. Moreover, some historical rare books lack explicit titles because of damages or loss. In those cases, those books are called with some texts, such as the first sentence of the main text or titles in other parts of the book. In consideration of those backgrounds, descriptions on parts of books with words in terminology for bookbinding are very important.

For extraction of named entities from natural texts, Japanese language has specific difficulties because it is an agglutinative language which has no delimiters between words. For overcoming the difficulties, a morphemic analysis tool called MeCab [20] and pattern search programs are used in the proposed method. In notes of bibliographic data, Japanese texts consist of around 10 words per sentence because the texts in data are required to be in laconic style. The performance of MeCab is improved if a user prepares user dictionaries suitable for target texts. In the proposed method, we prepare a dictionary containing words relevant to historical rare books collected through digital archives and web databases [21].

In the previous experiment on Ichiba [21], 951 sentences in notes from 222 records were analyzed by MeCab. An evaluation tool of MeCab with 4449 prepared correct answers was employed to measure the accuracy of the analysis. The F-measure as an accuracy was 0.863 with the user dictionary, comparing to 0.594 without it. The result shows that use of a user dictionary relevant to historical rare books is very effective. The rest of words which were not properly analyzed need to be corrected by hand.

In addition, the extracted words are classified by categories into named entities. Finally, the words tagged by one of the named entities are assigned to URIs generated by location of the words in the sentence and registered in a database [21].
3 CONSTRUCTION OF INTERNAL DATA STRUCTURE

In section 2, we discuss that some predefined technical terms in bibliographic rules are important to indicate characteristics and structure of historical rare books. The proposed method describes the relations of these terms in OWL [22], where we define an ontology suitable for the data. Ontologies written in OWL can mechanically provide knowledge in domains. To model an ontology, it is necessary to collect classes, properties and terms for constructing relations.

First, basic relations (R1, Figure 1) between Tenseki class, a conceptually top class for historical rare books, and named entities (Table 2) are constructed with Protégé [23].

In detail, we first define TFB (terminology for bookbinding), Role (roles) and Term (other technical terms) classes according to standard Japanese bibliographic glossaries and rules [9–11]. In real operation, ontological relations between TFB, Role and Term classes (R2, Figure 2) are prepared in an original laconic style using symbols for referring to entries (Table 2 in [24]). After completion of R2, a script is employed to translate R2 into OWL format.

TFB, Role and Term classes also have domain-specific relations: variants, inclusions, intersections and other relations shown in Figure 2. For instances, Tenseki class has instances of Title class which relates to MainTitle and TitleOfWork classes (Figure 1 and Figure 2). MainTitle class indicating titles written on outer covers refers to main titles in standard bibliographic descriptions. On the other hand, TitleOfWork class describes titles taken from texts in inner parts of a book, which are used as a title if cover titles are unreadable or lost. TitleOfWork class is described with words in TFB. In Figure 2, “序題”, a title on a preface, belongs to both TitleOfWork and Preface classes. In other words, this relation is described as an intersection between TitleOfWork and Preface in the ontology.

In the similar way of the relation between TitleOfWork class and classes for parts described with words in terminology for bookbinding, Role class describes roles and activities of people in Person class. Term class offers definitions and structural information of technical terms.

Finally, R2 is combined to R1 to supply the detail domain-specific ontology in addition to the basic ontology of Japanese historical rare books. The combined ontology [25] works fine on Protégé.

4 TRANSFORMATION OF REAL BIBLIOGRAPHIC DATA INTO ONTOLOGICALLY VERIFIED LINKED DATA

4.1 Applying Ontology onto Real Data

A bibliographic data, Ichiba is transformed into RDF format, which obeys the ontology defined in section 3. First, a main URI are assigned to the top level of a book. URIs which belong to instances of main fields are directly connected to the main URI.

Then sub URIs assigned to instances of TFB (parts of a book) class are also connected to the main URI. Other URIs for named entities extracted from each record of Ichiba, are assigned to the sub URIs. By this operation, each sub URI, which is a part of a book, is also able to be described with its origins. The model of the relations among the main URI, the sub URIs and named entities are shown in Figure 3.

4.2 Linking Instances of Named Entities to External Headings

To connect instances of named entities to external headings, a script crawls vocabularies on the Web. The script can retrieve candidates of headings relevant to the instances and store them in the prepared database. Actually, NDLA, VIAF, AAT, DBpedia and HuTime are chosen for exploration because these vocabularies have relevant words to Japanese historical rare books. For instance, “安政4年1月” in notes of Table 1 belongs to Date class. It can be automatically accessible to an entry through HuTime. For cases of person names, "（二世）蓬莱山人” in Table 1 can be linked to NDLA and VIAF. As the result of retrieval, you can get three candidates for “蓬莱山人” because descriptions in bibliographic data are often ambiguous.

In Table 3, the results of queries for each type of named entities show that most of words except for Title class can be connected to a URI in external Linked Data services.

Person class, however, is relatively difficult to be linked to external entries, because it is difficult for non-specialists to choose a proper person from candidates. Generally, a name does not uniquely refer to a person. In addition, a person holds several names particularly before Edo period. As a result, we need to infer a proper person by peripheral information relevant to the bibliographic data by hand. This is one of the remaining problems to transform bibliographic data into Linked Data. In the proposed method, after automatically extracting candidates, proper headings especially for person names are chosen from them by hand.

5 DISCUSSIONS

In this section, we discuss the efficiency of the proposed method for other bibliographic data of cultural heritage.

First, the method of extracting named entities using McCab and pattern search programs is common for analyzing Japanese natural texts. Moreover, most of natural texts in bibliographic notes are short and written in similar patterns. For example, in notes of Table 1, named entities in one sentence line up in order of date, person and explanatory complements [21]. Those complements often
include predefined terms corresponding to one of instances in TFB, Role and Term classes. The rules for description in notes' fields, of course, depend on the collection itself. In any case, you can write your pattern search program for your target collection by referring to its description guidelines and real examples in notes.

Second, the proposed ontology has a simple but clear function to convey processes of contribution for a rare book by clarifying relations between a book and its parts. The ontology obeys frameworks of formal ontologies for cultural heritage such as CIDOC CRM, FRBRQ0 and EDM. These ontologies are capable of modeling detail structures in the domain. Particularly, they can express relations between cultural objects and their peripheral events. Accordingly, the proposed ontology can describe historical rare books with peripheral events in general.

Finally, according to the result in Table 3, automated selections of proper headings to each named entity are mostly successful. In TFB, Role and Term classes, sufficiently many words have already been provided by AAT, DBpedia or NDLA in Linked Data. Person class, however, requires much handwork than other classes because it is difficult to precisely choose correct candidates. This
situation is hoped to be improved because international organizations such as VIAF constantly construct public databases of headings on the Web. Credibility of URIs to external information should be validated by specialists of the domain.

In summary, the proposed method in section 1 is applicable to other cultural heritage not only Ichiba. The proposed method should be reviewed from various view points in history and linguistics.

6 CONCLUSIONS

In this paper, we proposed a semi-automated method to convert unstructured bibliographic data of Japanese historical rare books into Linked Data. For this purpose, we extract keywords relating to named entities through natural language processing and automatically assign them to URIs generated by locations of the words. The proposed method generates structured, machine-readable data, which obey a simple ontology under standard description rules. The method was applied to Ichiba and worked well. The method employs general characteristics in bibliographic data of Japanese historical rare books. Therefore, the method can be applied for other collections.

REFERENCES


Diverse Digital Collections Meet Diverse Uses: Applying Natural Language Processing to Born-Digital Primary Sources

Christopher A. Lee
University of North Carolina at Chapel Hill
calleec@unc.edu

Kam Woods
University of North Carolina at Chapel Hill
kamwoods@unc.edu

ABSTRACT
Use of primary sources often focuses on identifying and tracking entities (e.g. people, places, organizations, events) and other values (e.g. dates and times) across documents. There are many existing open-source natural language processing (NLP) tools that can identify and report on named entities, and projects in the digital humanities have previously demonstrated the scholarly value of NLP approaches when working with digitized materials. To date, there has been relatively little adoption of NLP tools for the analysis of born-digital materials by libraries, archives and museums (LAMs). There are a variety of challenges associated with applying NLP tools to born-digital primary source collections, including those forensically acquired from removable media. Many of the challenges relate to the diversity of materials and potential use cases. This paper reports on the BitCurator NLP project, which is developing software for LAMs to extract and expose features in text extracted from such materials. The resulting services and methods can be used by LAM professionals and the users they serve.

CCS CONCEPTS
· Information systems–Digital libraries and archives
· Computing methodologies–Natural language processing
· Security and privacy–Data anonymization and sanitization

ADDITIONAL KEYWORDS AND PHRASES
BitCurator NLP; archival processing; named entities; text processing

INTRODUCTION
A growing body of materials with significant cultural value are “born digital.” Libraries, archives and museums (LAMs) are increasingly called upon to move born-digital materials from their original locations – whether those are networked environments or removable media (e.g. floppy disks, flash drives, CD-ROMs, hard drives) – into more sustainable preservation environments. Information professionals must be prepared to extract digital materials from removable media in ways that reflect the rich metadata and ensure the integrity of the materials. They must also support and mediate appropriate access: allowing users to make sense of materials and understand their context, while also preventing inadvertent disclosure of sensitive data. There has been a significant shift in recent years toward the adoption of digital forensics tools and methods by LAMs in order to meet the above goals. This process has been facilitated by the BitCurator project (2011-2014), funded by the Andrew W. Mellon Foundation, which has packaged and disseminated an open-source software environment1 that allows users to create disk images; extract data and metadata from disks or directories; scan bitstreams for the presence of potentially sensitive data values; characterize the contents of disks; and perform other practical tasks, such as scanning for viruses, finding duplicate files, creating and working with forensically packaged disk images, generating cryptographic hashes, and viewing hexadecimal representations of bitstreams.

The BitCurator Access project (2014-2016), also funded by the Andrew W. Mellon Foundation, has further advanced these activities by investigating mechanisms for providing access to forensically-acquired data. A major product of the project has been BitCurator Access Webtools, which allows users to dynamically navigate filesystems of disk images, as well as searching over the content of many common file types contained within the images.2 The project also created BitCurator Access Redaction Tools to redact strings and byte sequences identified in disk images.3 This includes the ability to overwrite specific strings or regular expression matches, or byte runs that match specific files or directory entries. Members of the BitCurator user community and other interested LAM parties have expressed a need for tools to help in identifying and exploring information based on specific entities (e.g. people, places, organizations, events) of interest to curators and researchers. These needs can be addressed by building existing natural language processing (NLP) and information visualization tools on top of the existing BitCurator environment and BitCurator Access Webtools. This combination of functions can benefit both LAM staff and a variety of end users of digital collections.

RATIONALE

1 https://wiki.bitcurator.net/
2 https://github.com/bitcurator/bitcurator-access-webtools
3 https://github.com/BitCurator/bitcurator-access-redaction
One of the primary motivations for using the BitCurator and BitCurator Access software is to capture and provide access to contextual information. For example, the original filesystem attributes associated with files (e.g. directory paths, timestamps) can be essential to understanding their provenance and original order. There are many other types of contextual information that can be vital to making sense and meaningful use of digital objects. Lee’s “Framework for Contextual Information in Digital Collections” identifies nine classes of contextual entities: object, agent, occurrence, purpose, time, place, form of expression, concept/abstraction and relationship [1]. In a study of reference questions submitted to archives, Duff and Johnson found that most information requests were based on “proper names, dates, places, subject, form, and, occasionally, events when composing their information request” [2]. In their study of genealogists, Duff and Johnson identified information seeking practices that were focused primarily on names, places and time periods [3].

PREVIOUS WORK
There have been several recent initiatives to better exploit specific types of contextual information from within archival descriptions. Much of this work has focused on what Lee’s framework would classify as agents: individuals, families and organizations. Two standardization efforts have focused on characterization of such contextual information: Encoded Archival Context (EAC) [4] and the International Standard Archival Authority Record for Corporate Bodies, Persons, and Families (ISAAR (CPF)) [5]. The Social Networks and Archival Context (SNAC) project has extended the work of EAC-CPF by exploring methods to better create, combine and disseminate name records for persons, families and corporate bodies. While SNAC has been valuable in exploiting and exposing metadata from within human-generated archival descriptions, it has focused on one specific set of contextual entities and it has not addressed the automatic extraction of entities from the content of digital objects themselves.

Open source natural language processing platforms have matured rapidly during the past decade. These include platforms that provide web services and RESTful application programming interfaces (APIs) and integration with industry-standard testing and training corpora. Popular open-source toolkits for natural language processing include OpenNLP, NLTK, Pattern, and spaCy. Some of these platforms have been used in projects specifically targeted at LAMs, but the use cases are often narrow, and none include facilities specifically designed to process content from disk images.

One project that incorporates NLP functionality is ePADD (email Processing, Appraisal, Discovery, and Delivery), developed by Stanford University’s Special Collections and University Archives [7]. The ePADD software allows LAMs to process collections of email by using a customized Named Entity Recognition (NER) engine to identify correspondents within email. The NLP functions developed for ePADD have been customized to the domain of email materials to ensure high-quality results.

Another project with goals closely related to BitCurator NLP was ArchExtract, conducted (2014-2015) at the University of California, Berkeley. Bancroft Library staff worked with a student at the Berkeley iSchool, Janine Heiser, to develop a prototype for applying topic modeling, named entity extraction, and analysis of other common terms to collections of text documents, which Heiser tested against a collection of digitized materials. Topic modeling, in this instance, is a text mining technique used to cluster words identified in the corpus into abstract “topics.” These clusters may include entities extracted using NER, and can be cross-referenced to entities of interest to the researcher. The ArchExtract software, available through GitHub, is a proof of concept and not currently being actively maintained.

In the digital humanities, there have been many years of work on applying NLP to the content of primary sources. Projects in the field often focus on specific areas of NLP, such as NER and topic modeling to provide researchers with meaningful views of the people, organizations, and events described within a formal collection or data gathered from the Web. There is great potential to apply these methods more widely to LAM collections in order to identify and expose the sorts of contextual entities discussed above. Examples of such digital humanities projects include Perseus,5 WordHoard,6 Nora,7 Metadata Offer New Knowledge (MONK),8 and the Software Environment for the Advancement of Scholarly Research (SEASR)9 being used by the HathiTrust Research Center (HTRC).10 SEASR provides a modular analytics approach to improve scholars’ access to digital research materials through a text mining application capable of summarization, an ngram viewer, named entity extraction, and entity-to-graph visualization. This project is no longer active; the last major release was published in 2013.

Other projects have focused on specific types of contextual entities, such as PeriodO (Periods, Organized), which is creating a gazetteer of scholarly assertions about the spatial and temporal extents of historical periods in order to facilitate linking across data sources [8]. These activities have contributed to the available infrastructure for processing texts to facilitate research, but they have been applied predominately to digitized primary sources, rather than to born-digital sources.

BITCURATOR NLP PROJECT
BitCurator NLP (2016-2018), funded by the Andrew W. Mellon Foundation and led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS), is developing and disseminating software that responds directly to the user needs expressed above by identifying, extracting and exposing contextual entities from the wide diversity of born-digital materials that LAMs already hold and continue to receive.

---

1. https://github.com/jprecources/archextract
2. http://www.perseus.tufts.edu/
3. http://wordhoard.northwestern.edu/
5. http://monk.library.illinois.edu/
7. http://www.hathitrust.org/htrc_collections_tools
The BitCurator and BitCurator Access projects have defined and tested support for a digital curation workflow that begins at the point of encountering holdings that reside on removable media – either new acquisitions or materials that are within a repository’s existing holdings – and extends this to the point of interaction with end users, providing and supporting a variety of access mechanisms. BitCurator NLP is extending this earlier work by helping LAMs to identify and explore information based on specific entities (e.g. people, places, organizations, events) of interest to curators and researchers.

Our target use cases differ from previous work in two fundamental ways. First, disk images are internally complex (containing filesystems and sometimes bootable operating systems) and require a significant software dependency stack that is already provided by the BitCurator environment and BitCurator Access Webtools. These include the ability to read, mount and provide access to the contents of various filesystems, as well as extracting, presenting and reporting on files and metadata.

A second distinguishing factor in the application of NLP to disk images in LAM collections is that disks may contain a broad range of file types and data encodings, requiring substantial pre-processing to extract content so that it can be processed by NLP tools and organized into meaningful reports, access points and visualizations. By contrast, most previous applications of NLP methods have focused on more “well-behaved” collections or corpora with more consistent types of content. BitCurator NLP is building from and extending a variety of tools and initiatives discussed above to provide services that LAMs can be run independently or integrate into existing software environments and access portals via simple application programming interfaces (APIs).

Some LAM access systems in use today incorporate NLP to describe the contents of collections, but this technology is often tightly coupled to the platform being used or is applied strictly to file types that tend to share common structures and metadata. An example of this is email, which contains raw text but also significant structured metadata in the header that can be used to cue identification of persons and organizations and describe their relationships. Identifying entities, relationships, and other features of interest by processing open text from heterogeneous collections of files (such as those extracted from a disk image) is inherently “noisier,” as the extracted text will often contain patterns of features (such as persons, places, and organizations) common to a wide range of devices and production environments (e.g., documentation of system files).

BitCurator NLP is exploring approaches that focus on improving the utility of reports produced about the contents of born-digital collections. Using data extracted from open text using NLP tools, along with techniques described in recent digital forensics research to eliminate or deemphasize those that appear to be irrelevant or common to the system rather than the documents themselves (e.g., names and email addresses of developers or organizations that created the software used to produce a given document), the project team will also develop guidelines describing how to apply the tools in ways that support common access and research use cases.

The BitCurator NLP team is ensuring close integration between the existing functionality of the BitCurator environment, BitCurator Access Webtools and the software developed by the BitCurator NLP project. For example, we are increasingly making elements of the BitCurator environment available as self-contained software installers (specifically software packages that may be installed in Ubuntu and Debian Linux environments using existing package managers), so users can selectively install and update them as they find most useful. Institutions could load each of the access tools onto the same machine (or virtual machine) as the one they are using for the initial processing, or they could instead decide to run those tasks in different environments in order to better manage and allocate their resources.

CONCLUSION

Given the numerous projects that have focused on application of NLP to enable scholarship on older primary sources, it is striking how little investigation there has been into use of NLP to the growing volume of more contemporary born-digital materials acquired by LAMs. While these professionals may have access to some tools that apply a subset of NLP techniques (such as topic modeling, NER, and word or phrase frequency analysis) to specific collections or file types, they generally lack software allowing them to execute these techniques in aggregate for large sets of heterogeneous files (including complex files such as disk images). This is particularly important for materials that contain thousands or hundreds of thousands of files, when it is intractable to manually inspect materials to determine which of the files are relevant preservation targets and what relationships hold between the files that could be exposed in archival description. Providing mechanisms that allow LAM professionals to conduct this type of analysis in the existing BitCurator environment and in BitCurator Access Webtools will streamline this process. The BitCurator NLP software is enabling LAMs to incorporate (or improve) basic NLP capabilities within existing access environments (or establish them as a dedicated service) by providing a service layer that reads files from a digital archival store and produces reports for end users on demand.

ACKNOWLEDGEMENTS

This work has been funded by the Andrew W. Mellon Foundation. The BitCurator NLP team is composed of Jacob Hill, Christopher A. Lee, Sunita Misra and Kam Woods. We have also benefited from contributions of the project Advisory Board.

REFERENCES


Operational Pragmatism in Digital Preservation: establishing context-aware minimum viable baselines

Somaya Langley  
Cambridge University Library  
West Road  
Cambridge, United Kingdom CB3 9DR  
szl20@cam.ac.uk

Dr. Anthea Seles  
The National Archives  
Bessant Drive, Kew, Richmond  
Surrey, United Kingdom TW9 4DU  
Anthea.Seles@nationalarchives.gsi.gov.uk

Andrea K Byrne  
Rensselaer Polytechnic Institute  
110 Eighth Street  
Troy, New York, USA 12180  
byrnea5@rpi.edu

Dr. Dinesh Katre  
Centre for Development of Advanced Computing (C-DAC)  
Innovation Park, Panchavati, Pashan  
Pune, India 411 008  
dinesh@cdac.in

Dr. Jones Lukose Ongalo  
International Criminal Court  
Po Box 19519  
The Hague, The Netherlands 2500 CM  
jones.lukose@icc-cpi.int

Bertrand Caron  
Bibliothèque nationale de France  
Quai François Mauriac  
Paris Cedex 13, France 75076  
bertrand.caron@bnf.fr

ABSTRACT

Undertaking active digital preservation, holistically and thoroughly, requires substantial infrastructure and resources. National archives and libraries across the Western world have established, or are working towards maturity in digital preservation (often underpinned by legislative requirements). On the other hand, smaller collectives and companies situated outside of memory institution contexts, as well as organisations in non-Western and developing countries, are struggling with the basics of managing their digital materials. This panel continues the debate within the digital preservation community, critiquing the development of digital preservation practices typically from within positions of privilege. Bringing together individuals from diverse backgrounds, the aim is to establish a variety of 'bare minimum' baselines for digital preservation efforts, while tailoring these to local contexts.

KEYWORDS
digital preservation, operational pragmatism, minimum viable, baselines, local contexts, global views

1 BACKGROUND AND MOTIVATION

Custodianship of digital materials must ensure preservation is enacted in ways that are documented, repeatable and will enable long-term access to digital objects.

There are mechanisms for doing this, such as complying with standards (e.g. METS [13], PREMIS [14] IASA-TC 04 [11] etc.) and undertaking Trusted Digital Repository (TDR) audits and obtaining certification (e.g. Data Seal of Approval (DSA) [7], TRAC [17] and ISO 16363 [12], the Nestor Seal [16] (based on DIN31644 [9]) and DRAMBORA [8]). Other approaches include meeting the NDSA Levels of Preservation [18] or parsimonious preservation [10].

1.1 Unrepresented Voices

While national memory institutions and top-tier universities consider certification, organisations in developing countries struggle with the very basics of accessing suitable people and technology. [20][21][5]. Digital preservation also suffers from a predominance of first-world ‘majority’ views, while other minority voices are either too quiet or remain unheard [2]. The needs of non-Western and developing countries must be considered if digital preservation is to reach beyond Western boundaries [1][6]. What principles or best practices can be relied upon to support everyday work in minimally-resourced institutions or non-institutional contexts?

2 APPROACH

2.1 Foundations and Current Activities

Groundwork has been laid for pragmatic, practical and operational digital preservation through the Digital POWRR project [19] and subsequent ‘digital preservation on a shoestring’ workshops.

Meanwhile, efforts such as recent collaboration on documenting requirements for a Minimum Viable Station for digitising audiovisual materials is taking place [3]. Yet time is critical if we are to enact digital preservation processes without significant data loss [15].

2.2 Program

The panellists come from a wide range of cultural and employment backgrounds. They will address a series of specific provocations including, but not limited to: addressing reported errors from tools, fixity, infrastructure and storage, preconditioning, pre-ingest processes, preservation metadata, scalability (including bi-directional scalability), technical policies and workflows. Audience involvement is encouraged.

The intended outcome is a range of agreed-upon baselines tailored to different cultural, organisational and contextual situations.

3 PANELLISTS

Moderator: Somaya Langley has worked across the arts and culture, broadcast and for collecting institutions as a creative director, digital curation specialist, digital preservation specialist, producer, production manager, sound artist and technical assistant. Organisations she has worked for include the Australian Broadcasting Corporation, the Australian Music Centre, Design & Art Australia Online, the International Society of Contemporary Music (ISCM) World New Music Days, the National Film and Sound Archive of Australia, the National Library of Australia and the State Library of...
New South Wales. She was Co-Director of the 2008 and 2009 Electric fringe festivals, Australia's international festival of electronic arts and culture. She is currently working on a Polonsky Foundation funded project as the Digital Preservation Specialist (Policy and Planning) at Cambridge University Library, UK.

Panellist: Andrea K Byrne has worked with audiovisual archives, research data and government records, and travelled the world to New Zealand and back again. She brought a fresh set of eyes to a nascent, but successful program of born-digital deposits at Archives New Zealand (ANZ) and worked on its first live transfer. She is currently the technology and metadata librarian at Rensselaer Polytechnic Institute, where she is in the process of instituting a digital preservation plan.

Panellist: Bertrand Caron graduated from the École nationale des Chartes in Paris in 2010 as Archiviste paléographe. From 2011 to 2014, he worked as a project manager for heritage digitization at the University of Montpellier. Currently he is the metadata specialist at the Bibliothèque nationale de France (BnF). Since 2015, he has been a member of the METS Editorial Board and the PREMIS Editorial Committee.

Panellist: Dr. Dinesh Katre is Associate Director & Head of Department at the Centre for Development of Advanced Computing (C-DAC). He spearheads the Centre of Excellence for Digital Preservation funded by the Ministry of Electronics and Information Technology, Government of India. He has been instrumental in formally introducing the ISO 16363 standard in India and has participated in the audit process of the National Cultural Audiovisual Archive. In 2010, he charted the scope and roadmap for National Digital Preservation Programme (NDPP) for the Government of India. He was a member of the International Experts Consultative Committee of the UNESCO Memory of the World that drafted the Recommendation concerning the 'Preservation of, Access to, and Use of, Research Data and Government Records'. He spearheaded the Centre of Excellence for Digital Preservation funded by the Ministry of Electronics and Information Technology, Government of India.

Panellist: Dr. Jones Lukose Ongalo is currently the Information Management Officer at the International Criminal Court in The Hague. He is a Senior Information Management Practitioner with over eighteen years of national and international experience in developing and implementing strategies to achieve operational effectiveness and regulatory compliance by leveraging ICT. Key clients and projects include the International Criminal Court (ICC), United Nations International Criminal Tribunal for Rwanda (ICTR), Local & Central Government Uganda, National Water Corporations (Kenya and Uganda), Electrogaz Utilities (Rwanda), Central Bank Uganda, USAID, GTZ and the Government of Jamaica. He holds a BSc in electronic engineering, a MSc in organisational development, a PhD in computer science and an MBA.

Panellist: Dr. Anthea Seles received her doctorate from University College London (2016) and is currently the Digital Records and Transfer Manager at The National Archives (TNA), UK. Her doctoral thesis examined the applicability of Trusted Digital Repository standards in an East African context, which was the Digital Preservation Coalition Winner 2016 for the most distinguished student research in digital preservation. Dr. Seles has worked as a consultant and archivist internationally. She has extensively presented and spoken about practical digital preservation and digital data integrity for accountability and transparency.

4 ACKNOWLEDGEMENTS

Thank you to Michelle Lindlar from the Technische Informationsbibliothek, Ross Spencer from Archives New Zealand, Erin O'Meara from the University of Arizona Libraries and Dr. David S. H. Rosenthal from Stanford University for providing feedback on the panel concept. Thank you also to William Kilbride from the Digital Preservation Coalition. This panel would not be possible without the recent work of Ashley Blewer from the New York Public Library and Elvia Arroyo-Ramirez from Princeton University Library and their collaborators, which was the inspiration for this panel.

Panel discussion will contribute to development of digital preservation strategy for the Polonsky Digital Preservation Project [4].

REFERENCES

OAIS and Distributed Digital Preservation in Practice

An exploration of Danish and other use cases that contributed to the development of the Outer OAIS–Inner OAIS Model for Distributed Digital Preservation

E. Zierau
The Royal Danish Library
P.O. Box 2149
Denmark
elzi@kb.dk

ABSTRACT
The aim of the paper is to illustrate how the distributed aspects of digital preservation can be aligned in practice, with the concepts and principles of the Open Archival Information System (OAIS) Reference Model.

There has been a growing awareness within the digital preservation community of the need for cooperation between organizations to address digital preservation requirements. One common example is that replicas of preservation copies of digital objects need to be independently preserved (e.g., stored, managed, monitored, documented) to ensure that at least one correct replica will survive for as long as needed. Such independence can be achieved through distributed digital preservation that relies upon specific agreements between participating and contributing organizations. The OAIS Reference Model does not address the challenges of distributed digital preservation in detail, though it acknowledges the potential benefits and the options.

A model in form of an extension to the OAIS Reference Model was developed by a Danish bit repository project, as there was an urgent need for such a model. This model has evolved to the Outer OAIS–Inner OAIS (OO-I0) Model through the international project "Framework for Applying the OAIS Reference Model to Distributed Digital Preservation". Previous papers have presented the theoretical basis for the OO-I0 model as a model to describe distributed digital preservation systems in a way that conforms to the OAIS Reference Model, but practical examples of applying the model have been sparse.

This paper provides detailed descriptions of how the need for the OO-I0 model emerged, how it has been used for both design and audit of the Danish bit repository, how we plan to use it for minimal effort ingest, and what other use cases there are for applying the OO-I0 model for distributed digital preservation purposes. This will illustrate how using the OO-I0 model can assist in the analysis of complex digital preservation tasks of a distributed OAIS-conformant repository, where the OO-I0 model provides terminology and contribute to break down analysis and audit questions.

KEYWORDS
OAIS Reference Model, Distributed Digital Preservation, Standards, Audits, Analysis, Collaboration

1 INTRODUCTION
Awareness within the digital preservation community of the need for cooperation between organizations to address fundamental digital preservation requirements are traceable from the early days. One of the first digital preservation community reports from 1996 mentions distribution [14]. Up to the start of the 2010s most community discussions of distributed digital preservation have mainly limited to bit preservation [11,12,15]. However, since then cumulative community experience has demonstrated that collaboration, both technical and organizational, is necessary to address the challenges involved in achieving good practice for digital preservation for more and more aspects of digital preservation [9,17].

The Open Archival Information System (OAIS) Reference Model [4], is an international standard1 that has proven to be extremely useful to the digital preservation community. However, this reference model does not address the challenges of distributed digital preservation in detail. To fill this community need, the Outer OAIS–Inner OAIS (OO-I0) Model provides an extension to the OAIS Reference Model.

The purpose of this paper is to provide detailed examples (primarily based on Danish experiences) that illustrate how the OO-I0 model can be used, and thereby hopefully make it clearer and simpler to use the model to support analysis and audit of complex parts of an OAIS in handling distributed digital preservation in practice.

The OAIS Reference Model guides the development of sustainable digital preservation programs by providing a common vocabulary, an information model, and a high-level digital preservation architecture. The OAIS Reference Model defines an OAIS repository1 as [4] p.1:

"...an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a designated community."

This means that the development and long-term management of a full OAIS repository requires not only technology, but also the skills and practices of people within organizations. At the functional level, an OAIS repository can be understood as a system in which information packages are processed by

1In order to distinguish between the OAIS Reference Model and an OAIS, this paper denotes an OAIS as an OAIS repository, unless it is qualified with Inner or Outer.
individual functions that have been abstracted into higher-level functional entities as depicted in Fig. 1\(^2\).

![Image of OAIS functional entities diagram]

**Figure 1: OAIS functional entities.**

In the remainder of the paper, terms from the OAIS Reference Model and the names of OAIS functions, functional entities, and information packages appear in **Italic** font.

In section 6 on ‘Archive Interoperability’, the OAIS Reference Model briefly acknowledges the potential benefits of and the options for federation and distribution states that [4] (p. 6-2):

“In general one OAIS [repository] is not interoperable with another; however, there are a number of reasons that some level of interoperability may be desirable, motivated for example by Users, Producers or Management”

In relation to distributed digital preservation, the OAIS Reference Model standard also notes that [4] (p. 6-1):

“An OAIS [repository] may be geographically distributed but with all parts under the same Management, for example the Archival Storage Functional Entity could be divided over several separate locations to increase resilience against disaster. In other cases OAIS Archives with separate Managements may wish to co-operate as described below.”

However, to describe the technical and organizational aspects of distributed digital preservation properly, it is necessary to elaborate upon and extend the OAIS Reference Model and that need motivated the design of the OO-IO model.

**The Outer OAIS–Inner OAIS (OO-IO) model** is an overlay enabling the description of distributed digital preservation systems in a way that conforms to the OAIS Reference Model. The main purpose of the Model is to simplify the challenges when several organizations are involved – both organizationally (what needs to be done) and technologically (how it can be done).

The Outer OAIS–Inner OAIS takes its name from the way the model is used. Here, the term ‘Outer’ refers to the OAIS repository at the outermost level, e.g., a repository using a bit preservation system, while the term ‘Inner’ refers to parts of the functional entity of the Outer OAIS that is represented by an individual Inner OAIS, e.g., a bit repository performing bit preservation for the Outer OAIS. In other words; an Outer OAIS refers to an entire OAIS repository implementation that supports distributed digital preservation, including all of its Inner OAIS’s\(^3\).

The starting point for the OO-IO model was a model developed by a Danish bit repository project. An international project team expanded and broadened the use of the Danish model in "Framework for Applying the OAIS Reference Model to Distributed Digital Preservation" [17] that is referred to as the DDP project. In 2013, the DDP project started as a working group to adapt and extend current standards to address distributed digital preservation. The DDP project included representatives from both North American and European organizations that were engaged in distributed digital preservation, e.g., MetaArchive, the Danish BitRepository.org, Chronopolis, Data-PASS, DuraCloud, Internet Archive, UC3 Merritt, and Archivematica. Variations within these cases identified the need to address other OAIS functional entities, in addition to Archival Storage, that require distribution over multiple organizations. Fig. 2 provides an overview of the OO-IO model.

![Image of OO-IO model]

**Figure 2: OO-IO model.**

The OO-IO model builds upon section 6 of the OAIS Reference Model and, in doing so, the elaboration of the OO-IO model explicitly aligns with the current version of the OAIS Reference Model. By design, the OO-IO model specifies an approach for using the OAIS Reference Model to archive interoperability, in the form of distributed digital preservation, which section 6 does not provide.

Distribution as part of digital preservation will always be complex, and the OO-IO model deals with these complexities by deconstructing them into more manageable and simple components. The idea behind the OO-IO model is that each functional entity of an OAIS repository can be described as a complete OAIS repository (an Inner OAIS), that helps an organization (an Outer OAIS) to address that functional entity. Each Inner OAIS is managed as a complete OAIS repository,

---

\(^2\)Fig. 1 corresponds to Figure 4-1 in the OAIS Reference Model [5] with a small deviation. In this paper, eight-angled boxes are used to represent functional entities to distinguish functional entities from functions.

\(^3\)This paper uses OAIS’s as the plural form of an OAIS repository.
though it is dedicated to managing a single functional entity in the Outer OAIS. Using the OO-IO model may only involve one of the Outer functional entities, for example Archival Storage as a bit repository separated and managed by an external organization.

Previously published work on the OO-IO model has focused on theory, with substantiated arguments to demonstrate that the OO-IO model does not violate the principles of OAIS Reference Model. The most recent is the iPRES paper "Supporting the Analysis and Audit of Collaborative OAIS’s - Using an Outer OAIS-Inner OAIS (OO-IO) Model” from 2014 [16]. From a theoretical perspective, the OO-IO model may seem complex and even daunting because distributed digital preservation is a complex topic. This paper aims to make distributed digital preservation more approachable and clear by illustrating the application of the OO-IO model using a set of examples.

The Examples included in this paper span from detailed experience-based examples to potential future use cases. Most of the detailed examples concern preservation storage. Preservation storage is often the use case that prompts organizations to recognize the need for a collaborated effort, and therefore this is where most examples can be found today. However, as the DDP project found, there is much more to distributed digital preservation than preservation storage. Therefore, this paper provides more speculative examples that are based in part on the findings of the DDP project.

The experience-based examples describe how the OO-IO model has been used for the Danish bit repository: First, as basis for definition of terminology, analysis and design of the bit repository. Second, to audit the operating bit repository partly based on the audit and certification standard ISO 16363 [3] that is references the OAIS Reference Model.

Another example reflects the deep immediate interest within the digital preservation community in preliminary preservation storage as part of (pre)ingest. Latest summary of recent discussions can be found on the OAIS Community DPC wiki [1] about pre-ingest. The best poster "Minimal Effort Ingest" from iPRES 2015 addressed this challenge [6], and the topic is also mentioned in the paper on the OO-IO model from 2014 [16]. This example is also about preservation storage, and in spite of the fact that it is not fully implemented, it is included due to the emerging community interest in this case.

Finally, there are short descriptions of distributed digital preservation use cases other than preservation storage and for other OAIS functional entities than Ingest and Archival Storage.

2 EXAMPLE: BIT REPOSITORY DESIGN

This example explains how the need of the OO-IO model emerged, and how the model was used for terminology, analysis and design for a Danish bit repository. The project started in 2009 as a national pre-study project to investigate the possibilities of creating a shared repository providing bit preservation for Danish cultural organizations.

The establishment of the project was based on the acknowledgement of the need for collaborative effort to create independence between replicas (e.g. to avoid loss of data due to fire or local natural disasters), or more formally to achieve proper bit preservation that requires [12].

- Several replicas of data that is being preserved.
- Independence between the replicas.
- Regular bit audit proving that the replicas are identical.

A great inspiration for this was the LOCKSS program [11], where David Rosenthal in several papers pointed to the importance of independence, even at an organizational level to avoid identical procedural errors being performed on a majority of replicas of data and thus causing loss of data.

The participants in this project were the Danish National Archives and the National Libraries. To avoid misunderstandings that could result from different uses of terms in archive and the library sectors, the participants agreed to apply terms from the OAIS Reference Model. This had the intended effect, at least at the point that bit preservation was part of the Archival Storage function entity for each user of a bit repository (illustrated in Fig. 3 as Outer OAIS for a user).

![Figure 3: An Outer OAIS using a bit repository.](image)

However, it also added great confusion because non-preservation IT staff working with storage facilities regarded Archival Storage as representing the bit preservation fully, and therefore in their perspective, bit preservation was solely a question of storage technology. On the other hand, digital preservation staff were convinced that Preservation Planning had to be part of the bit repository as well, and therefore bit preservation could not be limited to Archival Storage. The challenge was that it was hard to distinguish between users of a bit repository (the Outer OAIS’s) and the bit repository itself (as a separate organization). As long as the OAIS repository was seen as one repository, it became unclear whether Preservation Planning would cover both Preservation Planning for the bit repository and the repositories using the bit repository, for example. Furthermore, this approach would conflict with the aim of regarding the bit repository as a repository managed separately from the users’ repository.

To mitigate the confusion, the Danish project created a model where the bit repository was depicted as a complete Inner OAIS within the Outer OAIS Archival Storage functional entity. In this way each of the different cultural organizations represented each their Outer OAIS’s using the same bit repository, which was seen as an Inner OAIS repository including all the OAIS functional entities. To verify that doing so did not violate the OAIS
Reference Model, analysis were made, and the final model was described and published [15]. Fig. 4 depicts this model in terms of OO-IO, the Inner OAIS is called AS Inner OAIS since it represents an Archival Storage. The names of the functional entities are prefixed with “AS-IO” to distinguish them from Outer OAIS functional entities, which are prefixed with “OO”.

The model helped to avoid further confusion when referring to OAIS terms, simply by specifying whether it concerned the Outer OAIS’s (the users) or the Inner OAIS (the bit repository). In this way, everybody could agree that the new Danish bit repository should conform to a full OAIS repository (the same way as LOCKSS does [11]), and that this Inner OAIS would function as OO Archival Storage for each of the Outer OAIS’s representing the cultural organizations.

In this way, the AS-IO functional entities of the bit repository were covering parts belonging to bit preservation and thus moved out of the OO functional entities. For example, the monitoring technology for storage media used for the preservation storage was not part of the OO Preservation Planning but moved into the AS-IO Preservation Planning, while monitoring technology for file formats that exists in the OO remained in the OO Preservation Planning functional entities and thus not a concern needing attention by the shared bit repository. Likewise for the other functional entities.

It turned out that the model was helpful in many other aspects as well. First, to avoid confusion in architectural discussion by using specific references to whether the discussions concern the IO or the OO repositories. Second, to identify which parts were covered in Inner and Outer OAIS not just the functional entities but also information packages and roles. For example IO AIP’s (Archival Information Packages) were simply bit streams with minimal metadata for lookup, and IO Management had to be defined as a separate management dedicated for the bit repository.

The OO-IO model became essential in architectural discussions to analyze the stakeholders’ requirements. The three organizations had very different requirements for bit preservation. One organization had only open non-restricted data, while another had requirements coming from the legal framework in connection with confidential materials [7]. From this analysis, it became clear that a flexible architecture was needed to meet varied requirements at different costs. Additionally, there were requirements for different levels of bit safety, where the number of copies has direct impact on cost. Also specific confidentiality requirements required special setups, which again had possible impact on costs and bit safety.

The requirements analysis feed back to the OAIS-based analysis regarding IO Management. The IO Management consisting of collaborative partners forming the bit repository. The IO Management were to ensure that requirements from Outer OAIS’s were met (i.e. requirements from IO Producers and IO Consumers) Here, the requirements to the Inner OAIS were identified to cover coordination of storing individual replicas on independent platforms services, and coordination of agreements that ensure continuous independence between replicas.

The OO-IO model also assisted in the analysis and definition of the interfaces of the Outer OAIS (OO) and the bit repository as the Inner OAIS. Since several Outer OAIS’s were to use the Inner OAIS, it was essential to have a common understanding and definition of the information needed for the IO Ingest to be able to persistently identify and retrieve ingested bits via IO Access. Another important aspect was definition of audit trails, where the Outer OAIS’s would need audit trail information on the bit level as supplementary information to the Outer OAIS’s audit trails. Digital materials on the logical level (e.g. as a book, where each page has audit trails per replica of bit streams representing the page) would need all audit trail information, including actions performed on the bit level. Thus, a part of the interface would need to include IO audit trail information to be passed as IO report information to the Outer OAIS’s through the IO Administration to the OO Administration.

From the analysis and specifications, the Danish project considered using LOCKSS or similar approaches. However, due to the specific requirement that confidential data must be stored off-line, these options were not possible. Instead the pre-study project initiated a new project to develop an actual implementation of an open source framework to set up a bit repository (bitrepository.org [7]).

Figure 4: The OO-IO model Archival Storage component.

3 EXAMPLE: BIT REPOSITORY AUDIT

After implementation, the OO-IO model was used to lay out audit points relevant for the bit repository. Since the OO-IO model conforms to the OAIS Reference Model, the ISO 16363 standard for OAIS based auditing was used, at least as inspiration. However, distribution aspects of digital preservation are not specifically addressed in ISO 16363. The lack of guidelines for distribution aspects by ISO 16363 has been addressed in several contexts. For example, as an issue raised in many of the interviews with organizations performing distributed digital preservation in the "Framework for Applying OAIS to Distributed Digital Preservation (DDP)“ [17]. Another example is in the paper "Self-assessment of the Digital Repository at the State and University Library, Denmark - a Case
Study [2]. Using the OO-IO model helped identify many additional coordination related issues by defining the Inner OAIS responsibilities and functionality. However, the OO-IO model combined with bit preservation theories helped overcome some of these gaps.

When planning for the audit of the bit repository, the first observation was that the bit repository would only cover some types of digital material. For example, the digital materials resulting from digitization of physically preserved documents were not included. Instead these digital materials could be stored more cheaply with some acceptable risk of loss, because re-digitization is possible. Therefore, these materials are kept within the Outer OAIS and are ingested into the OO Archival Storage via the usual path through OO functions (such as OO Receive data and OO Provide Data) in parallel with AS Inner OAIS covering the bit repository (as depicted in Fig. 5).

This split conforms to the OAIS Reference Model, which allows splitting functions according to needs of an implementation. Consequently, the bit repository audits do not cover digitized versions of physical materials, but only born-digital material or substitution digitization where no loss is acceptable.

In planning of the individual audits, the OO-IO model was used by mapping responsibilities between Inner and Outer OAIS’s and accumulating evidence across multiple OAIS’s to cumulatively demonstrate compliance with digital preservation requirements. This was partly based on the audit and certification standard ISO 16363 and partly on analysis of audit points needed in order to ensure proper bit preservation ensuring continuous independence between replicas and regularly bit integrity checks across replicas. Before outlining the detailed audits, a full picture of Outer OAIS and Inner OAIS’s was required.

In the actual Danish implementation, the bit repository was not established as an individual organization; instead, each institution negotiated contracts for using each other’s facilities for preservation storage of single replicas. Therefore, it became possible to use the OO-IO model in a different way. However, in the Royal Danish Library case, it was an advantage to keep the perspective of the Inner OAIS representing the full bit repository, so results from the analysis in the pre-study project could be used.

In the pre-study project, media technology was assumed to be placed in the AS-IO Preservation Planning. However, at the time of the launch, it became clear that an extra layer of Inner OAIS’s was needed. Replicas of data were in most cases placed on independent types of media in other organizations, where the requisite expertise about specific media existed, and could be monitored. So, all but the in-house replica was manage by a preservation program with a different Management. In accordance with the OAIS Reference Model, several OAIS

![Figure 5](image-url)  
**Figure 5:** The OO-IO model Archival Storage component with parallel OO functions.

![Figure 6](image-url)  
**Figure 6:** Example of a full picture for a bit repository with Inner (Inner) OAIS’s for replica repositories.
repositories with separate Managements can work cooperatively, but they are not the same OAIS repository. Therefore each of the organizations with a replica were viewed as separate OAIS repositories. To handle all audits of individual replicas, the in-house replica was viewed in the same way. Fig. 6 illustrates the chosen perspective in terms of the OO-IO model, where the bit repository is the AS Inner OAIS to the Outer OAIS, and the three units for replicas are Inner OAIS’s to the bit repository, i.e. Inner- Inner OAIS’s to the Outer OAIS. The Inner Inner OAIS’s are here referred to as the Replica IIo’s.

The use of the OO-IO model has, in the Danish case, assisted in maintaining focus within the individual audits. There are a number of people who have multiple roles for parts of the AS Inner OAIS and/or a Replica IIo, but by explicitly specifying the audit target, the audit process has avoided confusion about which role a particular person should assume in a specific context. Furthermore, it has also been easier to spot when overlapping roles were causing challenges to the independence of replicas and security of data. Lessons learned from the audits are provided below.

3.1 Audit of the Inner OAIS
The audit was organized by specifying the roles and responsibilities that the Inner OAIS was to cover (excluding what was covered by the Replica IIo’s). Most of the responsibilities are related to the coordination needed to ensure independence between replicas and uphold integrity across the replicas. One example is that media migration involves a risk of data loss. For example, audit of procedures to avoid simultaneous media migration, and continuous evaluations of whether new media should be included, possibly by including another options of storing replicas, e.g. on molecular DNA. The following describes various findings in the audits of the Inner OAIS.

One of the first audits revealed that one person had write rights to data that would enable him to edit data without leaving obvious traces of the change, losing the original in the process. The bit repository contained three replicas of data. Additionally, the bit repository contained votes (in form of checksums) to be able to point to a correct replica in cases where two replicas have errors. The person in question only had access to one of the three replicas, but in total, he had access to a majority of the votes. Therefore, he could change the copy and the checksums, and thereby make the repository software overwrite the original copies with the changed copy. So, the audit detected and mitigated this potential risk.

A more recent example was that an organization wanted to change their operating system on the servers holding one of the replicas. They wanted to switch from Windows to UNIX. However, another online replica was already on a Linux system. Since Linux and UNIX are strongly related operating systems, the conclusion was to stick to Windows to not jeopardize the principle of independence between operating systems.

A very current example is that the Danish government decided to merge the Royal Library of Denmark (in Copenhagen) and the State & University Library (in Aarhus) into one organization from January 2017 (now called the Danish Royal Library). Before this merger, there was organizational independence between the replicas, but after the merger, this independence disappeared. A solution to this challenge has not yet been found. An additional potential challenge to independence is that the daily operation of servers will be centralized in this new organization.

In relation to Ingest, there were audit points regarding ensuring security in transfer of data to the Replica IIo’s. In this relation there has been audit points about the state of storage for each replica before any receipt of proper ingest can be given. As described in the next section, one finding was that definition of when ingest is acceptable is not that straightforward, since a receipt of reception from a Replica IIo does not necessarily mean that the replica of data has reached its final destination as an independently stored replica.

3.2 Audit of the Inner Inner OAIS’s
The main advantage of regarding Replica IIo’s as individual OAIS’s, is that it became clear what kind of agreement had been created between the AS Inner OAIS and the individual Replica IIo’s, and thus forming the basis for audits. When the audits of the individual Replica IIo’s were prepared, the OAIS Reference Model and the ISO 16363 standard were used to identify the different audit point in relation to requirements. The tricky part was to make sure that all requirements were incorporated, since some of them related back to the requirements from the Outer OAIS’s, such as those regarding the status of bit integrity, confidentiality, access time, and costs.

The specific requirements regarding confidentiality were particularly hard to fulfill. Here, the Replica IIo with storage of data on an off-line tape was considered good for replicas of confidential data. However, challenges appeared already at a preliminary audit conducted before using the bit repository for confidential materials.

The main challenge was that the existing tape installation was tuned for data with no confidentiality issues, where large amounts of data were delivered daily. The new confidential data would only arrive in small portions, possibly monthly. When data arrived at the tape Replica IIo, it was placed on a temporary staging area and would not be written to tape before enough data had arrived to fill up a tape. For the new confidential data, this meant there was a high probability that data could remain in the staging area for up to half a year before it was actually written to a tape and placed off-line. This raised two questions:

One relates to bit integrity. The question was whether the data on a staging area could be considered ingested when it had not been placed on its final media. The problem is that risk analysis for data loss is based on the independence obtained from storage on different media. Given the timeframe of six months, this was not considered acceptable.

The second question is strictly related to confidentiality of the data. Since the Danish Royal Library has a policy that bit
preserved data must not be encrypted, the question was whether the staging area could be considered safe for confidential data. However, the external tape technology supplier had access to the staging area to be able to assist in solving production problems for the large amounts of open data. Although this was a huge advantage and no problem when working with open data, it was unacceptable for the confidential data.

The conclusion from these findings was that the running tape solution was unfit for confidential data. An analysis showed that it would be quite expensive to bring this platform to an acceptable security level for confidential data. Furthermore, both cost and access time would rise unacceptably for the open data, if stored on a secure platform. Therefore, a separate and much smaller installation would have to be created specifically for the confidential data. In future audits, such special installations are regarded as a separate Replica IIOs and thus giving a separate audit.

The rest of the audit points and findings were not of particular interest for the use of the OO-IO model, but merely traditional as audit of a repository, although here limited to the roles and responsibilities of a Replica IIO.

4 EXAMPLE: OO-IO USED FOR INGEST

The need for preservation storage as part of Ingest is a topic that the Royal Danish Library has been working with since the beginning of this decade [6,16]. The motivation is that the library is not able to control what kind of data is ingested, and where capacity and collection issues may result in longer waiting periods before finalizing the OO AIPs. This corresponds to the former mentioned discussions about pre-ingest [1].

Regarding the types of ingested material, there are cases where it is impossible to predict a timeframe for data curation. One example is reception of hard drives from deceased authors to the manuscript collection. Here, there is a long process before the final OO AIPs can be created: Moving data from the hard drive, curation, restructuring, enriching and performing initial preservation actions. Furthermore, this process is complicated by numerous directory structures and file formats.

Regarding collection issues, another example is computer games; Video trailers must be harvested and user guides digitized before the final curation process can take place to create the final OO AIPs.

Regarding capacity issues, there are numerous examples of digitization projects that are brief and produce large amounts of data. For substitution digitization of fragile and/or deteriorating materials, there are no possibilities for re-digitizing the materials, and thus digital preservation of the new digital materials is crucial. In some cases, such digital materials had to be curated before the final OO AIPs could be produced (e.g. interdependent data that needs to be connected) and due to the amount of data produced, the curation period was stretched over a longer period. In other cases, it has been limitations in ingest to the bit repository that delayed the final securement of data.

The bit repository mentioned in the previous sections will be used for this preliminary archiving. Although there is still no formal setup with analysis and audits based on the OO-IO model, it is still an example that is close to be in use in practice.

Preservation storage as part of Ingest can be modelled by the Ingest component of the OO-IO model and is illustrated in Fig. 7. Here the Inner OAIS is named IN Inner OAIS since it represents an Ingest Inner OAIS. It illustrates that the ingested Outer OAIS Submission Information Packages (OO SIPs) are preliminarily archived in the IN Inner OAIS before it is used for generation of Outer OAIS Archival Information Packages (OO AIPs) and parsed to the OO Archival Storage.

![Figure 7: The OO-IO model Ingest component.](image)

As for the Archival Storage case, the IN-IO Archival Storage represents preservation storage, and is therefore modelled as an Inner Inner OAIS (Corresponding to the Inner OAIS described in the previous sections). This is illustrated in Fig. 8, where the main OO Ingest functions are depicted as well, to indicate that there may be materials that will not be covered by the IN Inner OAIS, but take the usual path through the OO Ingest functions.

![Figure 8: The OO-IO model Ingest component with Inner Inner OAIS and parallel Outer OAIS functions.](image)

*Corresponding to Fig. 2, but where only the orange IN Inner OAIS is included.
Note that the *IN Inner OAIS* in Fig. 8 does **not** cover the full **OO Ingest** functional entity. Instead the *IN Inner OAIS* deliver **IN-IO DIPs** (**IN-IO Dissemination Information Packages**) to the **OO Coordinate Update** function, which cover the further delivery to the **OO Archival Storage** and **OO Data Management** functional entities. According to the OAIS Reference Model [4] (p.4-7):

"the Coordinate Updates function is responsible for transferring the AIPs to Archival Storage and the Descriptive Information to Data Management".

In this case, the *IN Inner OAIS* will be responsible for receiving **OO SIPs**. It will then perform quality assurance on and generations of the **IN-IO DIPs**, which will have to include the **OO AIPs** plus **OO Descriptive Information** in order for the **OO Coordinate update** function to coordinate the updates at the **Outer OAIS** level. This is the way that we at the Royal Danish Library expect it to be modelled for the Danish case. However, it should be noted that the final detailed model for a specific case would depend on what is included in the *IN Inner OAIS*. For example, the reception of the **OO SIP** along with quality assurance may be placed at the **Outer OAIS** level. In such a case the *IN Inner OAIS* will only cover the **Generate AIP** function of the **OO Ingest** functional entity.

Terminology is an immediate benefit achieved by using the **OO-IO model Ingest component**. A common confusion when dealing with preliminary archiving is that an **AIP may refer to the preliminary archived AIP** with minimal metadata before processing (e.g., an image of an author’s hard drive) or it may refer to the final **AIPs** in the repository (emails, research for different literary works etc.). Here, the **OO-IO model** can assist in terminology and visualization of what is referred to: **OO AIP** for the final **AIPs** and **IN-IO AIP** for the preliminary archived **AIP**.

As for the **Archival Storage** component of the **OO-IO model**, there are also advantages in analysis and delimitation of what is covered by the **IN Inner OAIS**, and thus how to audit.

In the Danish case, the *Inner OAIS for are Archival Storage* and the *Inner OAIS for the Ingest IO Archival Storage* will to a large extend represent the same bit repository. Therefore, a complete illustration of the Danish preliminary archiving case would need to involve all levels down to the **Replica level** in the same way as illustrated in the previous sections. It should be noted that such an illustration will seem much more complex and voluminous than what has to be audited in practice, since there will be overlap in places where the bit repository and its **Replica IOIs** are the same.

As for the case of **Archival Storage**, all the *Inner OAIS’s might not be represented by separate organizations with different management, but merely be organizations with relevant roles and responsibilities for the *Inner OAIS that it represents*. Still, defining the *Inner OAIS’s has an important role in the breakdown of audits and recognition of roles and responsibilities.*

5 OTHER DDP USE CASES

The examples in the previous sections are solely related to preservation storage with bit preservation. However, as mentioned in the introduction, the DDP project found other cases where distribution was needed for digital preservation. The following includes such examples as well as examples of use cases where the **OO-IO model** can be used for the remaining **OAIS functional entities**.

5.1 Other Ingest Use Cases

In the previous section, there was an example of using the **OO-IO Ingest** component for preservation storage. However, preservation storage is not the only example of distributed digital preservation needed in **Ingest**. In the DDP project there were found other use cases, in particular a need for distributed ingest. Examples of distributed ingest were micro-service-based solutions like UC3-Merritt and Archimaticca who had examples of using the distribution of micro-services to manage many simultaneous loads of ingest processing.

5.2 Use Cases for Data Management

The **Data Management** functional entity covers information, such as catalogs and inventories.

The obvious case related to preservation storage, is when inventory information can be critical, e.g., for collection building or to be able to recreate an operational system quickly, if the catalogs and inventories are corrupted or lost.

Another example is the case of linked data where sources may be spread across several organizations. In these cases, there might be information that has been considered part of the **Data Management** functional entity, but in practice, this information is crucial for future access to **AIPs**. It can also be **Representation Information** that is distributed across organizations: one organization has descriptions of its preserved assets, another organization has the format registry used for the preserved assets, and a third organization has the environment registry used for the preserved assets.

It can be argued that if the OAIS Reference Model is strictly followed, all relevant information should be placed in **OO AIPs**. However, for linked data, it may not be possible. Preservation storage or cooperative solutions may therefore be needed at least for parts of the **Data Management** database, and thus specification of an **Inner OAIS** can be of help.

5.3 Use Cases for Administration

The **Administration** functional entity mainly contains services and functions needed to control the operations on a day-to-day basis, including the **Archival Information Update** function that provides a mechanism for updating the contents of an OAIS repository [4].

One use case is the case of complex and time-consuming updates of data related to crowd sourcing. This is a case of distributed digital preservation in the form of preservation storage. Preservation storage may here be required in the **Archival Information Update** function, before re-ingest. A practical example of collecting and securing updates is distributed collection of corrections in archived Optical Character Recognized texts from the public. Such crowd
sourcing produces a continuous flow of new data that needs preparation, before it can be passed to Ingest. It should be noted that such cases could also be modelled as part of (pre-)Ingest as it has similarities to securing data as described in the previous section 4.

Another use case is for collection and maintenance of community based standards used in Administration. Standards are important when tracing operations on the long term. Community based standards are important to achieve good practice. On the other hand, digital preservation standards are improving over time, and particularly for community-based standards, there needs to be some control of local implementations to ensure persistent references to the actual standard used.

5.4 Use Cases for Preservation Planning

The Preservation Planning functional entity is responsible for planning to ensure that the information stored in the OAIS repository remains accessible and understandable.

The most obvious examples of preservation storage needed for Preservation Planning regards preservation planning information that is shared and maintained by a digital preservation community, such as format registries like PRONOM [10] and authorities for preservation metadata like event types used in preservation metadata [8]. The best current example is the need for establishment of a common software source library for emulators [13]. This is a huge task, and it is reasonable to assume that no single organization can manage it alone.

The OO-IO model can be used for analysis and auditing of included preservation storage or collaborate maintenance and preservation of the information needed in Preservation Planning.

5.5 Use Cases for Access

The Access functional entity provides services to make archived material visible to Consumers. At first glance, this does not seem to involve preservation storage or collaborative efforts. However, an example in a topic of current interest is how access is provided for materials where the emulation preservation strategy is chosen [13]. In such cases, OO AIPs will contain technical metadata about the original environments needed for access, but the setup needed for emulation will change over time according to present technologies.

Setting up emulation based solely on preserved data in the OO AIPs can be very time consuming (months or years) if all parts of the access platforms are lost (e.g. as a result of a natural disaster). Preservation of an access platform can be crucial for access, although the requirements for such a Preservation may be more short term than the actual digital materials.

There are also cases where responsibilities for access components are shared between different organizations in form of collections of components needed for an emulation platform. One example can be found in the paper "Exhibiting Digital Art via Emulation, Boot-to-Emulator with the EML Kiosk System" [3]. If such access platforms are regarded as crucial for an OAIS repository, this would lead to audit of this collaboration.

Both in the case of preserving an access platform and in the case of cooperative maintenance of an access platform, the OO-IO model can be of help in analysis and audits.

6 DISCUSSION AND FURTHER WORK

One of the first questions that comes to mind is whether the OO-IO model can be used in cases where an Inner OAIS does not comply with the OAIS Reference Model (e.g. one of the Replica IIOs used in the bit repository example). The examples provided have shown that using the OO-IO model in breaking down the complexity in distributed digital preservation can help analysis and auditing, without any reference to compliance. Thus, if achieving the benefits described is an objective, the answer will be yes. However, if the objective is certification then the question is not easily answered, but probably something that should be investigated further.

Another question is how the OO-IO model can and should be referred to by the OAIS Reference Model standard revision that is under way in 2017. As distribution is crucial for digital preservation, it is important that the OAIS Reference Model does address this much better than it does today, and the OO-IO model presently seems to be the best way to do so.

In case the OAIS Reference Model refers to the OO-IO model, this in turn leads to a consideration of how the ISO 16363 standard can be extended to cover distributed digital preservation. There are plenty of evidence that the ISO 16363 standard needs to be better to address distribution and coordination. The standard will most likely benefit from the OO-IO model, but will also need to include more bit preservation aspects especially regarding coordination and replica independence. Furthermore, if the ISO 16363 standard uses the OO-IO model, there has to be guidelines on how to evaluate the entire compliancy based on compliancy of all inherited Inner OAIS’s in order to become a certified OAIS repository.

A challenge in using and applying the OO-IO model is the complexity of the cases that detail the roles, functions, interactions, and outcomes of the interoperability between and within OAIS’s that are required to manage distributed digital preservation environments. Therefore, working with the OO-IO model requires a deeper familiarity with and understanding of the workings of the OAIS Reference Model than is required for simpler use cases and implementations. However this complexity is inherent in distributed digital preservation, and not introduced by the OO-IO model. Furthermore, as seen from the examples, it is not an insurmountable task, especially if one step is taken at a time. Taking all steps to begin with will probably be incomprehensible and mind-blowing for most people, as for example introducing OO-IO by starting looking at the OO-IO Ingest component illustrated in Fig. 8, but expanded with Inner Inner Inner OAIS’s for replicas as done for the OO-IO Archival Storage component in Fig. 6. Introduction can also be simplified by use of names for the Inner OAIS’s as done for the Replica IIOs.

In addition to work mentioned that will be needed in relation to standards, further work on the OO-IO model could be to
elaborate more use cases that illustrate and document audit processes of distributed digital preservation. Furthermore, the examples can be elaborated for use for academic and educational purposes.

7 CONCLUSIONS

This paper has presented a number of examples of the OO-IO model used in practice, and how this model supports the application of the OAIS Reference Model in relation to distributed digital preservation. Although the OO-IO model may seem complex at first glance, the examples have hopefully clarified how it can be used to support analysis and audit of complex parts of an OAIS repository.

In summary, this paper has provided detailed examples demonstrating how the Outer OAIS–Inner OAIS (OO-IO) Model supports the specification and audit of collaborative interactions between multiple OAIS repository implementations. The main advantages presented have been:

- Provision of terminology to distinguish between distributed parts and information at different stages.
- A method to conduct detailed analysis of complex distributed digital preservation, by breakdown into manageable components, where interfaces and contents, roles and responsibilities can be defined.
- A method supporting requirements analysis for stakeholders.
- A method to support identification of audit points, roles and responsibilities relevant for auditing.
- A method to enable detailed audits for all organizations involved in distributed digital preservation.

The provided examples and use cases have also demonstrated that the OO-IO model can be used in a flexible way adapted to the actual repository in hand. The flexibility regards:

- whether information is entirely covered by the Outer OAIS or by one or more Inner OAIS’s
- whether the Inner OAIS covers the entire Outer OAIS functional entity or just subsets of functions
- whether the Inner OAIS covers a separate repository with separate Management or an internal organisation within the Outer OAIS,
- whether multiple levels of Inner OAIS’s should be used

Finally, it has been discussed how the OO-IO model may supplement existing standards to address the important and needed distribution aspect of digital preservation.

ACKNOWLEDGMENTS

A huge thanks to Nancy McGovern, who contributed to an earlier paper on the OO-IO model and to this paper to a degree where we considered that she should be a co-author. This paper is primarily based on Danish examples that demonstrate the development and implementation of the model, so we determined that it made sense for me to be the sole author this time. I would like to acknowledge her contribution to both the present paper as well as to previous work on the theoretical parts, where she contributed to the DDP project and co-authored the iPRES 2014 paper [16]. Also a big thanks to Ulla Bøgvad Kejser, who have given useful input to this paper, but also previous work, where Ulla co-authored the paper about the initial OO-IO model [15]. Also thanks to Andrea Goethals from Harvard University for useful input and discussions on the more theoretical parts and to my Danish colleagues Jakob Möesgaard, Bolette Jurik and Asger Blekinge for useful reviews of this paper.

REFERENCES

References to web resources are here provided using the drafted Persistent Web Identifier URL (draft 00: pwind://archive.org:2016-12-12_14.42.12Z) which is not yet functional.

Superb Stewardship of Digital Assets – Developing a strategy for Digital Archiving and Preservation at the University of Notre Dame

H. Hockx-Yu
Office of Information Technologies,
University of Notre Dame
Notre Dame, IN 46556
USA
hyu3@nd.edu

ABSTRACT
This paper describes the effort to develop an institutional strategy for digital assets management for the University of Notre Dame, where information is currently managed within a devolved organisational structure, and archiving and preservation are often overlooked. It advocates a lifecycle approach of digital assets management and recommends a strategy with the goal to embed considerations for archiving and preservation in policies, workflows and technologies across the entire organisation.

CCS CONCEPTS
• Applied computing → Education → Digital libraries and archives

KEYWORDS

1 INTRODUCTION
Just a few years ago, many of the materials universities used for teaching, learning, research and communications existed in physical formats. More and more of this analogue world now presents only as digital artefacts. Much of the learning, teaching and research at today’s universities is done digitally. Digital records and data are now ubiquitous and their proper management and preservation is critical for academic institutions to achieve their overall aspirations.

The University of Notre Dame (UND) is a private research university located adjacent to South Bend, Indiana, United States. It was founded 1842 by Father Edward Sorin, a priest of the Congregation of Holy Cross. UND is ranked #15 among National Universities in 2017 [19] and #12 of the top research universities in 2016 [2]. Catholicism and Catholic intellectual tradition continue to serve as the underlying foundation for the University’s aspirations as a community of scholars and administrators. It is also the University’s point of greatest distinction from many other research institutions [15].

UND’s leadership consists of Fellows, Board of trustees, Deans of colleges and schools and the President’s Leadership Council. As with many higher education institutions, while there is a central process of allocating funds, devolved managerial responsibilities and budgeting are essential elements of the management model. Authority and accountability are with business units, including the autonomy of spending budgets, regardless of the sources: whether it is centrally allocated funds, return on endowment or grants. The organisation and structure of e.g. IT services in this context consist of a hybrid of practices. While the Office of Information Technologies (OIT) has the responsibility of supporting enterprise-wide computing on campus, it merely serves as a technology partner or a service provider. Business units are fully autonomous in deciding whether to use OIT provided services or solutions. Some business units for this reason have their own IT team to ensure support that is more sensitive to local knowledge or requirements. IT infrastructure services, however, is an area where high rate of adoption can be observed and is increasing. There is in general a good collaborative relationship between the OIT and local IT departments.

As digital information profoundly changes learning, teaching and research, it also brings about many challenges, ranging from discovery among a vast amount of information, diversity of formats, and rights management to analytics. The biggest challenge is however organisational, how the University as an organisation in its entirety ensure consistency or achieve economy of scale, being aware that devolution tends to encourage duplication or creating silos.

Since 2009 various initiatives have already taken place at UND to coordinate efforts and address the challenges of digital assets management. They were of different scale and involved different academic and administrative units. A taskforce or project based model has been commonly used to support these efforts, where executive-level committees were formed to oversee the initiatives, who tasked working groups with participants across campus units to document and prioritise requirements and make recommendations. These initiatives have put in place pieces of infrastructure and some processes needed for digital assets
management and placed NUD in a good position to revisit the topic and address the gaps.

Between October 2016 and March 2017, the author undertook a consultation exercise involving a wide range of stakeholders across campus, including interviews and meetings with over 80 individuals from over 20 departments. The ultimate goal was to develop a campus-wide strategy to guide the preservation, access and dissemination of NUD’s digital assets. The findings and recommendations are reported in this paper.

2 DEFINITIONS

In the context of NUD, the term “digital assets management” is often used to refer to the activities related to managing digital information. The use of the word “assets” implies an underlying process of making value judgement. Regardless of the formats, digital assets are records, data and resources, typically owned by the University, regarded as having value and therefore needing to be retained and/or preserved. The implied valuation raises questions with regard to the identification and appraisal of digital information, as well as roles and responsibilities: how does the process of attaching value work and whose responsibility is it?

Three broad, not necessarily mutually exclusive categories have been identified to scope the digital assets landscape at UND:

University Records
The Records Management and Archives Policy [18] defines University Records as “recorded information created or received in the course of conducting University business and kept as evidence of such activity. This definition extends to records in all formats, including but not limited to e-mail, electronic and paper documents, film and print graphics, and audio and video recordings.

Research Data
The Data Retention and Access Policy [9] defines Research Data as “information recorded or customarily recorded in the relevant field, as a result of research. Data include notes, records, slides, photographs, drawings, information stored in electronic and/or computer readable form, reports, publications, correspondence, and summaries, compilations, or derivatives of other data.

Resources for teaching, learning and research
This category of digital assets refers to the wide range of content the University collects, creates, receives as donation, or purchases to support teaching, teaching and research. Examples include the Library collections, donated datasets, electronic databases, electronic journals, also lecture captures and Massive Open Online Courses (MOOCs).

The lifecycle model [1] is a commonly accepted framework for managing digital information and data, which recognises the change of value and frequency of use over time. It also allows the identification of different stages in the lifespan of digital assets from creation to archiving that otherwise might be overlooked. There are different ways of defining the stages of digital assets, dependent on institutional practices and the level of detail. The model outlined in Figure 1 is a generic and broad one, omitting much detail yet applicable to all categories of digital assets created and acquired by UND. The key is to apply this way of thinking systematically and to make sure that strategies and policies are consistent, and the crucial stage of “Archiving and Preservation” is not routinely overlooked.

Figure 1: Digital Assets Lifecycle

For the purpose of the work described in is paper and based on UND’s specific circumstances, the terms “digital archiving” and “digital preservation” are defined as follows:

Digital archiving
The process of acquiring and appraising records and data for long term retention. It also includes all the activities related to providing access to archived data and records.

Digital Preservation
The series of managed activities that ensures ongoing access to digital material, with a focus on guarding against deterioration and technological obsolescence over time.

2 METHODOLOGY

The consultation was conducted by the author, in her role as Program Manager for Digital Product Access and Dissemination, reporting to the UND Vice President for Information Technologies & Chief Information and Digital Officer. The author’s post is jointly supported by the Office of Information Technologies and the Hesburgh Libraries.

The goal of the consultation was to gain understanding of the current state of digital assets management across campus, and to identify the gaps. The findings will inform the development of a campus-wide strategy and a program of work for its implementation.

Over 80 individuals were consulted across campus, mostly through face-to-face meetings and interviews. The interviews were semi-
structured, using predefined as well as open questions. The focus was on obtaining an overview of the current practice and understanding what happens to digital assets throughout the lifecycle: what digital content are created, how they are managed, used, archived and preserved.

A small group of individual researchers, who do not necessarily represent a college or department, were also consulted. Their requirements include end-of-project data management, preservation of personal research material and support for collaborative research. One of these is an anthropologist, whose personal research “archive” spanning over 25 years contains analogue and born-digital material of various formats including human subject data.

When use cases were identified for collaboration, or referral of (otherwise unknown) services, efforts have also been made to bring people together across business units and functional teams. This provided an opportunity to think about future workflows and processes based on real-world problems and already led to better collaboration, potential grants and improved processes and access to some digital assets. Selected use cases are described in Section 4.

The main goal of consultation exercise was to obtain a broad picture of digital asset management at UND and therefore had an internal focus. Benchmarking was an element of the project and served as point of reference, rather than motivation for change. It was mainly based on comparison with similar research institution in the US and in the Midwest, including Stanford University, Yale University, Princeton University, University of Michigan and Indianan University. The National Digital Stewardship Alliance’s digital preservation levels were also used to assess UND’s compliance [8]. This pointed to e.g. the lack of dedicated positions related to digital preservation, and some NDSA level one data protection issues.

Research reported in this paper is based on data collected from a large number of individuals. Caution has been applied to make sure research ethics have been taken into account and the appropriate protocols are being followed. A formal compliance review submission was made to UND’s Research Compliance, and the Institutional Review Board (IRB) determined that the project qualifies as exempt human subjects research, as its nature does not meet the regulatory definition.

3 FINDINGS AND RECOMMENDATIONS

In this section, findings and recommendations that emerged from the consultation exercise are reported. The findings describe the current digital assets management landscape at UND. The recommendations directly respond to the findings and suggest specific interventions or strategies to address the identified gaps and challenges.

3.1 Findings

While previous digital assets management initiatives have delivered solutions or capability, these tended to meet specific requirements and did not necessarily achieve maximum adoption at the institutional level. Many recommendations were not followed through or resources for implementation were not in place. The executive committees and taskforces were charged with understanding requirements and making recommendations, but not with the overall implementation or change management, leading to start-stops over time.

Finding 1:
Previous digital assets management initiatives did not make enough impact at the institutional level, leaving many valid recommendations unimplemented.

Finding 2:
Charging models for services are inconsistent across campus, and can form barriers for consistent lifecycle management of digital assets.

Finding 3:
Intended enterprise solutions are often unknown, not considered or not affordable to SOME departments.

Lifecycle management of digital assets, especially archiving and preservation arrangements, needs to be embedded across the organisation, reflected in institutional policies, information and data governance, IT services management and day to day business processes and workflows. It needs to be taken into account from
the point of creation and considered especially when entering into contracts with 3rd party suppliers. This is currently not the case. Archiving and preservation are routinely overlooked.

**Finding 4:**
*There is generally a strong focus on “now”. Risk of digital assets being lost is not well understood, nor what is required to ensure their continued access.*

The amount of digital content we create far exceeds the amount we are able to keep and preserve. Archiving is as much about keeping as it is about discarding things. Common questions being asked in this context are what do we keep and who should decide on this? There are a number of university-level policies which define retention periods for records and research data, and go a long way in help answer these questions:

<table>
<thead>
<tr>
<th>Table 1: University Policies relevant to Records and Data Retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy Name</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>Data Retention and Access [9]</td>
</tr>
<tr>
<td>or</td>
</tr>
<tr>
<td>or</td>
</tr>
<tr>
<td>Retention of Grant or Contract Financial Records[20]</td>
</tr>
<tr>
<td>or</td>
</tr>
</tbody>
</table>

**Finding 5:**
*There is a lack of general awareness of the University’s data and records retention policies. And it is not evident how these are implemented for assets in digital formats.*

Some departments are not fully aware of the role and expertise of the Hesburgh Libraries and University Archives. Curatorial decisions are made without consulting them; records are created or collected and kept locally (despite having become inactive) without depositing to the Archives for extended period of time. Some of these records are stored on obsolete media or in obsolete formats, which are at risk of being lost without immediate preservation actions.

**Finding 6:**
*University Archives’ responsibility for collecting, maintaining and preserving the official records of the University is not well understood. It is not commonly accepted that this responsibility extends to include digital records.*

Although digital archiving and preservation are not alien concepts, there is a lack of technical skills and practical experience at the institutional level. Without detailed retention schedules and identification of systems of records, archiving requirements are rarely implemented in the many systems and applications that are used to handle digital documents, records and data. Basic bit-level preservation actions including periodical fixity checks are not explicitly or systemically carried out. Digital files are often deposited or transferred using physical media, without essential metadata or checksum-based verification. The Hesburgh Libraries and University Archives, who have the long-standing mission of preserving knowledge and institutional memory, are still in the process of analogue to digital transition. They are working towards active preservation of collection items and systematic collection of institutional records in digital formats, both requiring time, investments and development of relevant skills.

**Finding 7:**
*Some digital assets have already been lost. Some are currently stored on obsolete media or in obsolete formats, not accessible and at risk of being lost, without (immediate) remedy actions.*
Finding 8:
Born-digital University records are a key area of challenge for the University Archives. An up-to-date, function-based retention schedule (as opposed to office-based), and identification of system(s) of records will allow wider application and support automation. However, without the skills, resources and tools required to acquire, appraise, preserve and provide access to these records, it is very challenging for University Archives to fulfil their role.

Research data takes many forms. Archiving and preservation of research data, because it covers a wide range of disciplines with diverse data management, sharing and curation practices, is a complex area where there is no one-size-fits-all solution. Many universities approach the challenges by setting policies, putting in place storage and data repositories and by providing support and guidance for data management - an important element is to meet public access requirements for federally funded research. The ultimate goal is to support the key principle underpinning the scientific discourse: the ability to verify results and claims independently, which is being seriously jeopardised by the disappearance of referenced data in research articles as well as the reliance on computer hardware, software and configurations and source code.

Notre Dame Research, various computing and data centres, as well as the Hesburgh Libraries already offer a wide range of services and support to researchers. A good example is how grants and awards are managed, where the University’s system of record for research administration, Cayuse, plays an important role. We however need to question if individual researchers can be expected to preserve research data reliably in the long term. If yes, what else needs to be put in place to support them fulfil this role.

CurateND is UND’s digital repository committed to providing permanent access to and long term preservation of the University’s Research and Scholarship, including published work and research data. It also serves as the Hesburgh Libraries’ digital library system. CurateND is a relatively new service, only launched in 2015, and the team has so far focused on the technical development, based on identified use cases. CurateND is not solely an open access repository. It contains open access, embargoed, and private content. The positioning of CurateND is to become the system of records for the University’s scholarly outputs, regardless where the data resides.

CurateND has a set of policies defining various aspects of the repository, including a Review and Retention Policy [4]. The policy commits to retaining certain types of material in perpetuity and defines the retention periods and review criteria for the remainder of the content.

It is still early days, the deposit experience with researchers however seems somewhat mixed. While many researchers understand the potential risks of information and data loss, a degree of reluctance has also been observed. This could be due to a number of reasons:

- Different disciplinary practices in data sharing
- Lack of trust in ceding control of data
- Unclear about ownership and rights
- CurateND does not meet specific requirements
- Depositing research data seen as “one more thing to do”
- Lack of awareness of the existence or purpose of CurateND.

Finding 9:
The association between research data and administrative records are not maintained. The dual role of CurateND as the institutional repository and Hesburgh Libraries’ digital library system is intended to optimise use of the technology. However, it does at times create competition in terms of resource as well as priority.

Storage and backup services are crucial to protecting digital assets. A number of enterprise storage options are offered by the OIT, serving different purposes. Some are unlimited and funded centrally; some come with a base entitlement and charge a fee for additional space. Many departments also use locally sourced storage solutions, e.g. RAID, SAN, cloud storage, due to access requirements or budget restrictions. Key observations and feedback specific to storage include:

- Data is often inconveniently spread across different storage locations, without the confidence or confirmation of proper back-up.
- Different storage solutions are not well understood; some departments experience shortage of storage capacity.
- Alternative storage solutions, e.g. network-attached storage (NAS) devices and external disks, are commonly used, not just as short relief. These do not scale cost-effectively, are not intended for long-term use and can lead to data loss.
- The longevity of some storage solutions (e.g. Google Drive) are unclear.
- Individual researchers’ desktops are not consistently backed-up.
- Some block-level storage solution, e.g. the LTO Tape Archival Storage, is not widely known on campus.

Finding 10:
Despite the existence of enterprise storage solutions, much data still seems to be stored on (scattered) direct-attached storage, or locally managed NAS devices, creating data islands and poor data protection.

3.2 Recommendations
The recommendations are summarised in this section, grouped into a number of areas.

The longevity of our digital assets is a campus-wide challenge. While business units can take actions on their own there is a strong
consensus that coordinated strategies and shared resources are preferred over siloed local efforts.

Two particular frameworks are recommended to guide and benchmark UND’s digital archiving and preservation efforts:

- “Levels of Digital Preservation”, developed by the National Digital Stewardship Alliance [8].
- “Research Data Storage: A Framework for Success”, developed by the EDUCAUSE Center for Analysis and Research [5].

3.2.1 Strategy, Policy and Organisation
UND has an excellent track record in managing paper records and physical assets such as buildings. As more and more assets exist in digital formats, we must apply the same rigor demonstrated in planning and management and take the same coordinated approach to managing our digital assets, starting with a recognition of their value and their stewardship as a strategic priority. One of the University’s strategic goals is to “foster the University’s mission through superb stewardship of its human, physical, and financial resources” [12]. It is pertinent to also include the vast amount of digital resources we create, procure and are given, that underpin our academic and administrative activities, and aim to provide a track record of equally excellent stewardship for this new class of University assets.

Recommendation 1: Amend the 4th goal in the University’s Strategic Plan to include “digital resources” as an additional area where superb stewardship is required, so that it reads: foster the University’s mission through superb stewardship of its human, physical, and financial resources.

Recommendation 2: Develop a business model so that the required resources are in place to prioritise and implement the recommendations made in this report.

Stewardship of digital assets is not a one-off event but an ongoing process, operating in tandem with the full range of services supporting digital information environment and the overarching financial and organisational infrastructure. Project or taskforce-based efforts are time-bound, focus on specific outcomes, so may not be appropriate for core activities such as digital stewardship, which requires processes proceeding continually over time. This does not mean that we should not ever run projects and task-forces to solve specific problems. The key is to emphasise programmatic progress so that outcomes delivered by projects and taskforces are synthesised, followed up and embedded in business processes through change management.

Recommendation 3: Move from project and taskforce-driven activities to a fundamental programme of core activities by establishing a permanent organisational fixture (e.g. a person, an office or a team) with campus-wide remit to coordinate policies, procedures, practices, resources and technologies related to digital assets management.

Digital resources are far from being unmanaged at UND. A diverse array of people, technology and processes are in place, forming the environment in which digital resources are created, stored, secured and used. What we do not have is a culture or the practice of thinking beyond now, i.e. what happens to digital resources beyond the stage of active use? What must we do now to ensure we can find and interpret our digital records and research data in the future? These questions cannot be ignored until digital resources are passed into the custody of the Libraries and Archives, nor does the responsibility of digital preservation reside solely with them, only at the very end of digital assets’ lifecycle.

Recommendation 4: Carry out campus-wide effort to identify the process and control points where digital archiving and preservation considerations can be routinely embedded, e.g.

- Campus-level information committees and governance bodies
- Service units creating digital resources and departments commissioning the work
- Procurement including 3rd party data / information creation and hosting
- IT service management framework
- Business and project planning

3.2.2 Library and Archives
Among the three categories of digital assets defined in section 2, resources for teaching, learning and research is a category for which the University Library and Archives have shared responsibility of long-term custodianship. Some archival records have research values and can be used as education resources, and vice versa. As technologies advance and new types of resources are created and added to the category, it is important that the Library and Archives keep abreast of new developments and clearly define roles and responsibilities, to avoid grey area and to ensure none of the assets fall between the cracks.

Discovery has been made of digital assets stored on obsolete media or in obsolete format (see examples in 4.1 and 4.2). Some of the legacy assets are collection items held by the Libraries and Archives, some belong to researchers’ personal archives of research data. They are no longer easily accessible and are at risk of being lost without immediate remedy action. There are also analogue materials stored on physical media which need to be digitised to avoid media deterioration.

Digitisation and format migration routinely take place at the Libraries and Archives. In addition, a number of academic and teaching support units also have old equipment and offer faculty and students a similar service (which are not extensively used and the usage is declining). It makes sense to consolidate these as a campus-level service.
Legacy content is not the sole challenge. As an institution we also need to build the capability and expertise to preserve current and advanced digital resources. Novel classes of born-digital data, information and knowledge are being created, each with its own underlying artefacts needing to be preserved. Complex digital objects such as augmented reality, mobile apps and gigapixel images are increasingly used to support teaching, learning and research. We collectively do not yet know the best methods for archiving and preserving them. We need to take actions to address the challenges, together with peer institutions and the digital preservation community. We need to invest in a permanent team within the organisation who can build the know-how and technical expertise for hands-on and ongoing digital preservation activities.

**Recommendation 5:**
Establish a digitisation and digital preservation centre, based at Hesburgh Libraries, to serve campus-wide needs. The centre will digitise, preserve, research, and develop solutions for preserving advanced digital resources. It will also educate and advise, by recommending and helping to implement common standards and encouraging best practice.

Some institutional records reside in systems and locations external to the University, e.g. on the web and social network platforms, without our active control. This type of records is also ephemeral in nature and can easily disappear - the commonly cited average lifespan of web pages is about one hundred days. Collecting or archiving these records has therefore become a routine activity for many organisations including over 250+ academic institutions in the US, including all the Ivy League schools. Currently UND’s web and social presence are not systematically archived.

**Recommendation 6:**
Hesburgh Libraries and University Archives need to make an informed decision on whether to archive the University’s web and social network presence as institutional records. Also consider the potential of web archiving in helping to build digital collections in areas where we have strong existing physical collections or research interest, e.g. Catholicism, Irish studies.

Born-digital records are the most challenging issue for University Archives. There is a recognition that the records management function in the Archives needs to prioritise and focus accelerated attention on electronic records to ensure the University’s history, its legal and administrative records are documented appropriately. Efforts are already underway to develop function-based retention schedules and to put in place dedicated electronic records management resources.

**Recommendation 7:**
Build the capability for and more rapidly transition to electronic records management. Archives should work closely with the technological partners on campus to identify systems of records and find ways to implement the appropriate retention schedule in these systems. As a minimum, disaster recovery plans must be in place for vital records.

### 3.2.3 Research Data

Archiving and preserving research data is a complex challenge which require collaboration between many campus units and with individual researchers. To progress in this area, the best strategy is to build on the work already done internally by Notre Dame Research, various computing and data centres, Hesburgh Libraries as well as academic departments, and externally by various research and disciplinary communities. A two-pronged approach is recommended: at the institutional level, meeting public access requirements of sponsored research is no-doubt a priority. From individual researcher’s perspective, an immediate task would be to discover and “rescue” research outputs which have become inaccessible due to technology obsolescence – this should one of the services offered by the digitisation and digital preservation centre mentioned in Recommendation 5.

**Recommendation 8:**
Review, update and coordinate the research data retention policies including explicit treatment beyond the stipulated retention periods. Make sure the designated custodians are aware of and are supported in fulfilling their roles. If not already in place, develop the necessary procedures to implement these policies.

**Recommendation 9:**
Consider measures that can help ring-fence the two distinct purposes of CurateND, as UND’s institutional repository and Hesburgh Libraries’ digital library system, to ensure that requirements for both are met. These could be financial, organisational or branding-related (e.g., dedicate more resources to each pipeline of content).

**Recommendation 10:**
Maintain central records of research outputs (publication and data) arising from (at the very least) sponsored research, regardless of whether these are uploaded to CurateND, disciplinary repositories or other repositories required by funders, and link these with the corresponding administrative records held in the Cayuse system.

### 3.2.4 Storage and Cloud Services

Sitting low in the technology stack, storage and backup are the foundation and most important enablers for any archiving and preservation endeavour. Arguably a centralised approach has a lot of advantages. This, for example, will allow identification / inventorisation of digital assets and enable bit-level preservation actions. The findings around storage however points to the need for basic steps to ensure data protection, e.g. by reducing the use of direct-attached devices as a long term storage solution, and ensuring data is backed-up appropriately. We also need to communicate effectively about the purpose and cost structure of
enterprise storage solutions. If departments cannot use enterprise solutions, either because of performance or funding issues, an alternative approach, instead of driving people to implementing local solutions, is to work with departments and find ways to offer a service that DOES meet their requirements.

Just as many other higher education institutions, Notre Dame has widely adopted the Software-as-a-Service (SaaS) approaches to Cloud computing and are transitioning significant portions of our infrastructure to the Cloud [10]. This Cloud First effort is guided by a lifecycle framework, which includes an “end of life” stage of a solution or a service [11]. Explicit treatment of data, when retiring a service or a solution, however will need to be added and defined, based on the appropriate retention schedule, to make sure data ends up in the right places at the end of the lifecycle.

80% of UND’ IT services are expected to move into the Cloud by the end of 2017. We are aware of the risks associated with cloud computing such as information security and regulatory compliance. Moving data to the cloud however also has implications on its longevity and there are a number of specific aspects we need to consider:

Data integrity
How do we manage the integrity of stored content over the duration of the service, which encompasses not only protecting data from unauthorised alteration, but also from bit rot. How can we ensure data is safeguarded in the event of disasters, supplier failure, or a decision to change service provider? In addition, what happens to the data when a service is retired, or if a supplier goes out of business, and in the event of a contractual dispute or termination?

Archiving and preservation
It will take some thought to work out how to archive and preserve (inactive) data sitting in the Cloud, and how these processes integrate with the many applications holding and processing the data in its current lifecycle. The challenges are in fact parallel to doing this on campus. The location of data does not change the nature of archiving and preservation activities. Many vendors already offer cloud services for archiving and preservation, with the ability to integrate with key cloud storage platforms.

Recommendation 11:
Review storage service strategy in light of the findings. Make plans and take actions to address the identified issues. An immediate goal should be reducing the use of direct-attached devices as a long term storage solution, and ensuring data is backed-up appropriately.

Recommendation 12:
Identify long-term preservation risks related to Cloud services and put in place measures to mitigate the risks (e.g. through contractual agreements). Explicitly define treatment of data, based on the appropriate retention schedule, such as when retiring a service in the Cloud.

4 USE CASES
The section offers description of a number of selected use cases, which came to the author’s attention during the consultation exercise. These real world examples support the findings and recommendations described in Section 3. They also illustrate the ongoing nature and urgency of digital archiving and preservation.

4.1 Recordings at risk
An internal report delivered in 2010 by one of the previous digital assets management taskforces mentioned an archive of historically significant sound recordings held by WSND, the FM radio station that is part of UND’s Student Media Group. These recordings are on open reel tapes and stored in an office environment. They include interviews with prominent campus figures such as Father Hesburgh and speeches given on campus by Robert Kennedy and a number of presidents. A recent inventory-check also revealed recordings belonging to the Notre Dame Sophomore Literary Festival (SLF), a student-organised event that invited the top names in literature to give readings, deliver lectures, and engage in panel discussions on campus, particularly in the event’s early years in the 1960s and ‘70s. Twenty-five poets, fiction writers, dramatists, and critics came to Notre Dame to participate in the SLF. Prominent among the participants were Allen Ginsberg, Robert Duncan, Jerzy Kosiński, Tom Stoppard, Ishmael Reed, John Barth, Diane Wakoski, and William Gass.

Approximately 45 hours of SLF readings and lectures on sixty-eight open reel ¼” tapes from 1968, 1971, 1972, and 1979, currently housed at two separate locations, partially by WSND and partially by University Archives, were selected and brought together, which formed the basis of UND’s grant application for the Council on Library and Information Resources’ (CLIR) Recordings at Risk Program [3]. As a result of the project, the tapes housed at WSND will be transferred to the University Archives for optimised storage and better access, and added to the Archives’ online finding aid.

4.2 1966-1967 Sisters Survey
Sister Marie Augusta Neal was professor of sociology at Emmanuel College in Boston, Massachusetts between 1953-1991. She became the Director of the Research Committee of the Conference of Major Superiors of Women’s Institute (CMSW) which conducted the CMSW Sisters’ Survey of 1966-1967, a population attitude survey designed to assess American sisters’ readiness for renewal. With 649 variables and responses from over 130,000 Catholic sisters, it is believed to be “the largest, single, data gathering event ever performed with regard to women religious” [7]. All materials related to the CMSW Sisters’ Survey and Sr Neal’s follow-up studies were donated to UND Archives between 1991 and 1996, including computer data tapes [17].
The Sisters Survey data was initially assembled, processed and stored on IBM EBCDIC format tapes and converted to newer formats in 1996, before being deposited to the UND Archives. Under the custodianship of UND, the survey data was transferred to CDs then to computer hard disk in 1999. The Survey data has fortunately survived the format migrations but has not been used for 18 years since 1996.

The data came to the author recently without any file extension. Running the file through DROID [21], a file profiling tool developed by the UK national Archive did not deliver any useful information. Knowing that we are looking at cross-tabular data, and the possible involvement of statistical software, a few obvious extensions were added. The files could be opened with Excel but contained endless lines of numbers without much meaning so the first challenge is knowing how to truncate the data.

The truncation problem was later resolved by UND’s Economics Librarian, who has a good understanding of survey data – “I just saw the pattern”, he explained.

Things followed smoothly from this point onwards. The dataset has now been reformatted and stored in .dta and .csv formats. The “codebook” was also recreated – our librarian wrote scripts and extracted all the questions and pre-defined responses from PDFs and pulled them together in one document. The dataset is in the best possible format for re-use. We are just dotting the i’s and crossing the t’s, before releasing it publicly.

A key takeaway of the above use case is that just keeping the bits safe is not simply not enough. Active use is the best way for monitoring and detecting digital obsolescence. Materials created by obsolete technology require ongoing care and necessary intervention such as media and format migration. Furthermore, without metadata, in this case the notes, finding aid and scanned codebook, it would be very difficult to interpret or make sense of the Sisters Survey dataset.

4.3 Deposit workflow for University Archives

It came to the author’s attention that the current workflow for depositing or transferring digital files to University Archives relies heavily on the use of physical or portable media such as DVDs, external hard disks and flash drives. An example of this is how UND’s DeBrattlo Performing Arts Center (DPAC) deposit recordings of concerts and performances. Recordings are initially stored on DPAC’s network drive. Once a year these are burnt to DVDs, to make space for new recordings and for deposit to the University Archives. The workflow was set up at a time when network bandwidth was inadequate for transferring large audio files.

A small project took place to improve the workflow and take advantage of current network capacity. The Storage Services team in the OIT has now set up a dedicated S3 bucket to allow individual departments to drop files for the Archives. Those regularly deposit files to the Archives no longer need to copy files to physical media, and instead place the files in the department’s folder on S3. The Archives will move and process the files, and store them eventually in the Spectra T950 Tape Library, UND’s enterprise archival storage service. Archives can also use the same folder to deliver back any archival files requested by the departments for re-use.

Additional workflows being developed and hinge on this new workflow include file naming conventions, formats and metadata requirements, and checksum base file transfer verification.

5 CONCLUSIONS AND NEXT STEPS

Digital Archiving and preservation are important aspects of digital assets management and are yet to be systematically approached at UND. Work reported in this papers constitute steps towards achieving superb stewardship of the University’s assets which now include digital assets.

UND is currently undertaking the Campus Crossroads Project, the largest building initiative in the history of the University [14]. The $400 million project is aimed at maximising the potential of one of the most recognizable and centrally located buildings on campus: Notre Dame Stadium. The project includes a digital media centre with a 2,000-square-foot studio, bringing together currently dispersed media production work across campus. This is an opportunity to develop a common solution for managing video assets, including digital archiving and preservation requirements – a separate project is already underway, involving departments across campus who produce, use and provide custodianship over videos.

Stakeholders who participated in the consultation exercise were invited back to a follow-up workshop in April 2017, where findings and recommendations described in the paper were reported. Facilitated discussions also took place with the goal to arrive at a prioritised list of recommendations, and to identify the key barriers associated with their implementation. The results, based on the MosCoW and Kano [6] data collected via a pre-workshop survey and group discussions during the workshop, pointed to recommendations 1, 2, 5, and 7 as highest priorities, followed by recommendation 11 and 12.

![Figure 2: Prioritisation of Recommendations](image-url)
A further process of prioritisation and business planning process will follow, to design a programme of work to implement the recommendations, which will be constructed upon the three pillars of policy, process and technology. Digital archiving and preservation are central elements of our forward strategy, as they can help us effectively retain, manage and leverage our digital assets.

REFERENCES


ABSTRACT

As part of the national strategy for Cultural Heritage in the Netherlands, under the umbrella of the Network Cultural Heritage (Dutch acronym: NDE), a working group consisting of representative digital preservationists from different large (mainly cultural heritage) organizations worked to contribute to the certification process of their organizations. This paper describes the various activities of the working group that resulted in a well-balanced approach for certification based on (1) a phased approach, (2) a supporting tool for maturity level checking (3) translations of the DSA and DIN/ nestor guidelines into Dutch, (4) lessons learnt based on a survey amongst DSA certificate holders, (5) training materials in Dutch.

KEYWORDS

Certification, DSA, nestor, trust, trusted digital repositories, cultural heritage

1 INTRODUCTION

The growing maturity of digital preservation of the past 20 years is reflected in the need to be able to check whether organizations are the trustworthy custodians of their digital collections. Several tools were developed, from self-assessment tools and risk analysis tools like Drambora [1] and the SPOT model [2], to official certification tools like the Data Seal of Approval (DSA), the DIN 31644 / nestor and the ISO standard 16363-2012. There is a clear need for benchmarking and checklists, not only at libraries and archives as the frontrunners in digital preservation, but also from funding organizations in relation to the growing amount of research data repositories. These standards are used not only for self-assessment and official audits but also for example in European projects like e-ARK to check the maturity of their participants in implementing the e-ARK products [3]. Still, not everyone agrees that international standards like the ISO 16363 are internationally applicable.[4] In the Netherlands many organizations are taking steps in digital preservation. In order to raise the maturity level of these organizations a programme was started to popularize the concept of audit and certification on a national scale.

2 ACTIVITIES OF THE NDE CERTIFICATION WORKING GROUP

2.1 The certification working group

Funding by the Ministry of Education, Culture and Science facilitated the founding of the Network Digital Heritage (Dutch acronym: NDE), in which all the major Dutch heritage organizations with digital collections work together in developing a system of common facilities and services for improving the visibility, usability and sustainability of our digital heritage.[5] A special NDE working group was started for Audit and Certification of digital repositories (NDE-AUDIT wg), to continue the work already started under the Dutch Coalition for Digital Preservation (Dutch acronym: NCDD). This working group ran a nationwide campaign to raise awareness around auditing, created a roadmap for audit and certification of repositories and above all propagated the benefits for organizations.

The participating organizations in the NDE-AUDIT working group, consisting of six large cultural heritage organizations with a variety of digital collections had a different “maturity level” in digital preservation. Some already had acquired the DSA and started preparations for the nestor seal, others were still discussing whether they were ready for certification. Apart from these content holders one participant represented the Cultural Heritage Inspectorate of the Ministry of Education, Culture and Science. Originally it was planned to start at the same moment with the audit and certification processes and to keep each other informed of the various steps taken. But because of the different levels and different time schedules this did not work out. However, the variety in uptake also offered an excellent opportunity for developing a balanced view by combining the insights of the experienced and the less experienced.

To reach a wider audience, knowledge exchange about audit and certification was not restricted to archives, (university) libraries and cultural heritage organizations. We deliberately also contacted...
some suppliers of digital archiving software in use by these very organizations. Although it was acknowledged that these suppliers could not become certified themselves, as the certification is only given to the collections holders in combination with the systems they use to achieve this, several of them were very interested in acting together with their customers, either in consultancy or in being part of the certification effort.

The NDE-AUDIT working group planned a range of activities that will be described in more detail in the following paragraphs. These will contain:

- A phased approach of certification
- Translations of two standards into Dutch
- Further development of maturity checking software: the Scoremodel
- Survey to collect experiences with DSA v2
- Training and communication

2.2 A phased approach of certification

Although the intention to become a certified organization is often described as a goal in institutional policies [6], if the internal challenges are serious and no external pressure for certification exists, the intention might not be realized. A certification process will require resources. The question “Am I ready for a certification process” becomes important, as it is not wise to invest in a time and effort consuming process if you are uncertain whether you can fulfill the basic standards.

At least in Europe the starting point for certification is the 3 level European Framework [7] in which an organization starts with applying for DSA, followed by a nestor seal and finally applying for an official ISO audit based on ISO 16363.

We decided to use this framework but to add two important steps: the Initial Self-Assessment and the Exploratory Phase.

In the past some organizations have started straight away with ISO 16363, amongst them LOCKSS. Apart from the benefits (more about this later) not seldom the feedback is that this certification process was a hard one, and did cost a lot of time that could have been spent on other activities. In the NDE-AUDIT working group’s opinion, adopting the European Framework model and to go for a more phased approach would be more suitable in our organizations, where not seldom digital preservation was only part of the activities and often an activity in competition with the activities related to physical collections. From bottom to top we distinguished the following 5 phases:

Phase 1. Initial Self-Assessment
Phase 2. Exploratory Phase
Phase 3. DSA
Phase 4. DIN
Phase 5. ISO

In the following paragraphs these phases will be described with the expected outcome of the phase.

Phase 1. Initial Self-Assessment

There are several basic tools for an organization to check its maturity level and to get an initial idea of how mature the organization really is with respect to digital preservation. The NDE-AUDIT working group saw this step as an essential one. In many organizations digital preservation is still a niche activity and by undertaking an Initial Self-Assessment we expect that the organization will get a more reality based view on their preservation activities.

An extensive overview of tools can be found in one of the deliverables of the European e-ARK project.[8] One of the tools available for the Dutch audience is the Scoremodel, developed by the Dutch organization DEN and the Flemish organization PACKED, described more in detail in paragraph 2.4.

The outcome of this Initial Self-Assessment depends on the tools used, but in case of using the Scoremodel a report is printed in which the strong and the weak areas are shown in a “spider web score”. The next step for the organization can be to improve the identified weak areas before the organization will be ready for even the basic level of certification (phase 3).

Phase 2. Exploratory phase

As the audit and certification process will take time and resources, it is important to start a process in close harmony with the management of an organization. We introduced a step in the process in which enough detailed information is collected that can be used to support a management decision of “go/no go”. Although it might not be possible to create exact figures about the time needed, at least what need to be identified is:

- Which departments will be involved. These are not only the departments that are directly involved in the preservation process, but also for example the legal department, Human Resources department, Finance &Control department etc.
- Which persons and in which role in these departments need to be available during the audit and certification process?
- The availability of documentation of the preservation activities is the main requirement for all levels of certification, as this documentation is necessary evidence to prove that an organization is compliant. An estimated guess need to be made how much documentation is already available and in which areas one might expect extra effort (and resources) is needed (gap analysis).
- How well is the certification method understood in the organization? Can one “translate” the DSA requirements to the situation in their own organization? Are all requirements clear? The NDE-AUDIT working group did an investigation in which they took the DSA version of 2016 and compared the terminology used in DSA with the terminology used in their own organization (library, archive, AV organization). This offered an interesting overview of different jargon used in various domains and strengthen the case for discussing this before starting the audit and certification process. It is very important that an organization is aware of time needed to discuss the interpretation of the requirements, as this is strongly related to the expected effort needed.

Which part of the digital collection will be chosen as a candidate for the audit and certification process? It is important that the
As no one of the participating partners in the working group was applying for this audit method, we did not investigate the steps. Currently (may 2017) there are no European certification bodies to perform a ISO 16363 audit, in the US ANAB (standards organization) will take care of certifying auditors in the US.[10]  

2.3 Translations of two standards into Dutch

Coincidentally, another initiative was started by Regionaal Historisch Centrum Limburg and funded by Archief2020 and NCDD to translate the nestor DIN 31664 standard into Dutch. Archief 2020 was a 4-year collaboration (from 2012-2016) between archives with the goal to innovate the Dutch archives on all levels with a focus on digital challenges. Motivation behind the translation of the nestor standard was the need of a set of instruments to qualify an (archival) e-Depot as trustworthy and to have an overview of the main ingredients for a trustworthy designed e-Depot. Apart from that, although most Dutch people understand the English language, a Dutch translation would lower the barrier.

Members of the NDE-AUDIT working group helped to contextualize the first draft translation and supported the translator in domain specific jargon. Along this line it was soon decided to translate the text of the Data Seal of Approval as well. Since November 2016 a new version of the Data Seal of Approval was published, in which several requirements were changed to reflect the merge of the original DSA standard with the ICUS World Data System requirements under the heading of “Unified Requirements for Core [originally “basic”] Certification of Trustworthy Data Repositories”.

Although the total set of requirements (16, formally called Guidelines) stayed the same, some major changes were introduced. There are new requirements on “Security”, “Confidentiality/ethics” and “Expert guidance”. These headings already show the still present “scientific data center” view of the DSA. Although the changes are not many, some are significant. For example DSA version 2 originally mentioned explicitly OAIS in guideline 13, where it stated “The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS”. The new Requirements only mentioned OAIS for repositories “with a preservation remit”. [11] These differences between the versions of the DSA standards were one of the arguments to introduce an investigation of the understandability of the standard in the Exploratory Phase. For example: “Confidentiality/ethics” might be less obvious for national libraries as it might be for archives and social science data centers. The translation exercise itself required some tough decisions. First it was agreed not to translate the well-known OAIS terminology into Dutch, as we assumed the people that would use the standards, should also be aware of the common digital preservation terminology. In practice however, several requirements in DSA (and less in the DIN/nestor standard) asked for some contextual information to make it understandable for different domains. For example in the DSA standard the phrase “publication repository” is used but it is not clear whether here a “data publication repository” or a repository with articles is meant. In several cases references
are made to "citations", which refers to a concept from the scientific domain and is less important when certifying national libraries or archives. So some requirements will need additional information to explain the context. The morse so because related to each requirement is a set of documentation that the organization need to prepare for the audit. It is not always clear from the text what kind of documentation is required and as yet no repository is certified against this new standard, there is no evidence from other organizations that acquired the DSA according to these new Requirements. We need to wait how this will work out in practice, but it might have helped if an example list had accompanied the standard (like is done in ISO 16363 with a Self-assessment template).[12]

Both translations of the standards into Dutch are published [13]. Especially the DSA one is now in use in two organizations in the Netherlands that are preparing for this certificate, the National Archive and the International Institute of Social History.

2.4 The maturity checking: the Scoremodel

The Scoremodel [14] is an online tool in the Dutch language, in which questions related to 7 areas of digital preservation need to be answered, varying from mission and policy to ingest, knowledge and organization, access, planning and quality control and storage maintenance. Each question has an extensive explanation describing how it is related to digital preservation. The outcome of this exercise is a report, including a spider diagram which gives a clear overview of the main strong and weak areas related to the 7 topics. In the report for each question the answer is (automatically) analyzed, describing the risk related to the answer, the context and suggested actions. advise.

In the new version of the Scoremodel (to be published in 2017) the results are linked to DSA criteria. In this way users can estimate the readiness for starting the DSA certification process. The Scoremodel is a very helpful tool to get a first impression of the maturity of the organization with respect to digital preservation, as well as giving the organization practical advise. It will take 2 to 4 hours to add the relevant information in the tool, of course depending on available documentation.

2.5 Training and communication

The working group prepared two workshops in which a large group of Dutch and Flemish organisations (50 in total) discussed their preparations for audit and certification according to DSA. Several smaller Dutch organisations were already preparing themselves for DSA certification, some did this in collaboration with their supplier. Apart from that the NDE program started with a virtual learning environment for digital preservation in Dutch with a special chapter on audit and certification. The main training topics are related to having the right arguments to persuade higher management, requirements for documentation, estimating how much time will be involved in the certification process etc.

2.6 Survey to collect experience with DSA

One of the major activities of the NDE-AUDIT working group was designing a survey in order to gather practical information from colleagues who already applied for the DSA Seal. The following chapter will elaborate on this.

3 SURVEY FOR DSA CERTIFIED ORGANIZATIONS

In order to have more sound background information for the Exploratory Phase in which management need to be convinced of the usefulness of the audit and certification exercise and to get a better estimation of the investments and benefits of certification, we wanted to get information from organizations that already had applied for the Data Seal of Approval (pre-2016 version). The working group set up a survey to be held amongst these organizations, under the title “DSA-Experiences: help your peers!” The outcome of the survey could offer us a comprehensive overview of the experience with the certification process, especially as we specifically has some questions related to the first application of the seal. This information we thought important to inform our management about the certification process and to implement the lessons learnt into training and advice for the Dutch organizations’ planning to start the process.

3.1 Outline of the survey and the results

The DSA Guidelines were first published in 2008, and had an updated version in 2010 and one in 2014-2015. As of March 2016 some 50 repositories had obtained the seal. Some organizations have published their findings in meeting with DSA-requirements. [14] Yet, no comprehensive overview is available on the experiences regarding the process and results of certification of all DSA-certified repositories. In an effort to collect such experiences the NDE Audit working group requested permission from the DSA Board to field a survey among all digital repositories that had achieved DSA-certification by March 2016.

Despite the change from Guidelines to Requirements, the basic tenets of core certification efforts will remain unchanged, however, and the NDE-AUDIT working group is convinced that the outcomes of this survey can be of considerable significance to organizations considering or actually preparing for DSA-certification – also after the introduction of the DSA’s revised requirements in September 2016. The survey was sent out to the email addresses of 50 DSA repositories, or organizations that operated such repositories. The DSA Secretariat was confident about the correct delivery to 47 recipients. Of these addresses 18 filled out and sent in the survey. The total results and scores per questions can be found in the official report.[16] In this paper we summarize these findings, without repeating the original answers. For convenience sake, we categorized the answers and identified the following areas:

- Repository characteristics [Q1-3]
- Certifications achieved and planned [Q4-10]
- Trigger for certification [Q11]
- Certification efforts [Q12-13, 19-27]
- DSA comprehensibility [Q14-18]
- Benefits of DSA-certification [Q28-33]
**Repository characteristics**

Motivation to ask these questions: The DSA audit and certification method was initially focused on research data repositories, but the Dutch candidates for DSA certification are a more varied set of organizations, consisting of libraries, archives, digital art collections, data centers, an AV centre etc. To rightly interpret the answers in the survey, an indication of the kind of digital archive would be important, as would be the size of the organizations. All 18 repositories answered this question. The majority of the respondents describe the type of their repository as domain or subject-based, or as institutional, which is representative for the current DSA community: they employ between 1 and 12 fte’s, of which 0.2 to 10 fte’s work primarily on preservation tasks. It is noteworthy that the average number of fte’s that work primarily on preservation stands at 1; available human resources at the repositories are clearly modest in scope, although four repositories [out of 18] report that they employ more than 8 fte’s. Apparently, also repositories with limited human resources have been successful in applying for the seal.

**Certifications achieved and planned**

Motivation to ask these questions: The NDE-AUDIT working group was very much interested in the “certification history” of the DSA repositories. When did they do their first application for DSA? Did they really renew the certification every 3 years? Were they following the “European Framework model” and planning for the next level? And above all, it was essential that all respondents indicated they would report on their experiences during their first application for the seal.

The respondents were asked which version of the DSA the repositories first obtained (2010 version or 2014-2015, between the two versions are small differences). Out of 18 respondents, the majority of 14 obtained the more recent version, while 4 of them obtained the 2010 version. The responses are an indication of the larger uptake of this certification instrument in the DSA domains in recent years. But how many had renewed their seal after having obtained their first DSA-certification? All 18 repositories responded, of which four that had initially obtained the 2010 seal indicated they had indeed renewed their seal; all others had not yet done so. Given the time frame, there was no necessity to do this. Apart from that there is no pressure from DSA to renew, it is expected that the community will draw their own conclusions when an organization has a seal on its website that is very outdated. A further question in the survey explicitly asked about the estimated time investment in the certification process. One could assume that part of the respondents, the need for continuous maintenance and renewal of the seal is an accepted practice. But were they also willing to apply for the extended certification by means of DIN/nestor certification? All 18 respondents answered and most of them (16) had not applied at this level and were not in process of doing so, one of them was investigating DIN/nestor.

Asked whether they were contemplating to apply for the highest level: the external audit on the basis of ISO 16363 no one of the 18 respondents indicated that they had applied or that they were in the process of doing so.

**Trigger for certification**

Motivation to ask this set of questions related to the trigger for certification lied in the fact that the NDE-AUDIT working group consisted of a variety of organizations all with different motivations to get certified. Building trust and showing that an independent organization had recognized them as a trustworthy organization was for some of them important for the suppliers of data to their data repository. In general it was felt that the “big hubs” as they were called, meaning the main content holders in the Netherlands, should get certified to establish trust in general and to be an example for smaller organizations in the Netherlands. Apart from that there was an expectation that in the long run the funding organizations like the Ministry and research funding organizations would require a certification.

From the wide array of the DSA survey respondents’ input, some main considerations can be distilled:

One-third of the answers (7 out of 16) indicate that the repositories were motivated by an existing, inherent recognition of the importance of continuous professionalization and quality assurance in their digital preservation remit. Five answers derive from the repositories’ recognition of the value of the DSA in showcasing their value as a trusted digital repository to stakeholders. Another group of five answers indicate that the repository was already involved in the DSA’s development or acted on an invitation by the DSA leadership. Four additional answers indicate that the repository’s interest was triggered by an internal directive (management) or an external obligation (e.g., condition for partnering in a research infrastructure, funding).

To summarize: while the responses are varied, it is evident that the repositories were mostly triggered by a recognition that DSA-certification is a natural and appropriate instrument in (showcasing) their ongoing professionalization as trustworthy partners for long term digital preservation.

**Certification efforts**

Motivation to ask these set of questions was the fact that one of the goals of the survey was to deliver a rough overview of the effort and time investment that would be useful to plan the Exploratory Phase and the DSA Certification Process, based on real life experience of first time applicants.

One remark should be made here: based on the answers to the explicit question whether they had decided to keep track of the time investment, all answered that they did not do so before applying for the certification, and only a few (4 out of 16) had decided to do this during the certification process. One could assume that part of the answers below are based on time sheets, some on other sources and some on "memory".

The survey explicitly asked about the estimated time investment in getting a first impression of the DSA as a certification instrument.
This is what we would call the Exploratory Phase. The majority (65%) reported an estimated time investment of 10-20 hours; 3 chose the category 0-10 hours and 2 indicated 20-40 hours. A single respondent indicated a larger amount of time, 60 hours or more.

The actual time investments in the first step of the DSA procedure: the internal preparations, took more hours. The majority (60%) reported a time investment of 50-100 or 100-200 hours; 4 chose the category 0-50, while 2 indicated larger investments (1 of 200-300 and 1 of 500 or more).

Followed by the steps of submission and the actual peer review (questions for clarification etc.) 17 repositories responded; 1 skipped this question. The majority (78%) reported a time investment of up to 50 or 50-100 hours; 2 chose the category 100-20, while 2 estimated larger investments (1 of 200-300 and 1 of 500 or more). In a separate question the survey asked the respondents for their biggest challenge in dealing with the peer reviewer’s comments. This question was primarily intended to identify (potentially) problematic aspects of the respondents’ interaction with the peer reviewers. But most answers (8 out of 13) indicated that the respondents had in fact not encountered “challenges” or “problems.”

But did the time investments comply with or perhaps exceeded their expectations? Responses (16 out of 18) were distributed quite evenly through the full range of potential answers. If we leave out the three respondents who reported that they had no preconceived expectations on this issue, the largest subgroup (6) indicate they had underestimated the required investments, a smaller subgroup (4) had correctly estimated this aspect and the smallest group (3) had overestimated the required investments.

To summarize: for developing a first impression of the DSA, most respondents estimated a time investment of 1-20 hours; for preparations for the DSA procedure, most respondents estimated a time investment of 50-100 or 100-200 hours; for the actual certification process, most respondents estimated a time investment of up to 50 or 50-100 hours. None of the respondents had decided beforehand to keep a record of time investments, but four out sixteen decided to do so at a later stage. Some repositories had no preconceived idea of the required investments; the remaining ones varied widely in their evaluation of expected time investments vs. actual investments. The largest group of these (6 out of 13) indicated they had underestimated the required investments. We expected that a potential source for higher time investments than expected might be the interactions with the peer reviewers, but most respondents indicated that they had not experienced problems in this regard.

**DSA comprehensibility**

Motivation to ask these questions was the fact that at the time of the DSA Survey, the new version of the DSA-WDS was not yet officially approved by the DSA Board. But the NDE working group made a comparison between the DSA Guidelines 2014-2015 and the new version as published on the RDA website, as the development of the latest version was an activity of the RDA Repository Audit and Certification DSA–WDS Partnership Working Group.[17] The latest DSA version took a similar approach as ISO 16363 in choosing the main topics, which are now Organizational Infrastructure, Digital Object Management and Technology. Although the total amount of the original guidelines (as it was called in version 2014-2015) are the same amount as the current Requirements, 3 new Requirements appeared, covering “organizational infrastructure,” “expert guidance” and “security”. It is not possible to make a one to one comparison as there is no background information about the motivation behind the changes, but the influence of the already existing WDS Criterions is visible. Based on this comparison it became clear that some terminology used was rather domain specific and not always compatible in other domains. This was the reason to explicitly ask whether DSA was comprehensible to the variety of organizations: how clear and straightforward were the DSA guidelines?

The majority of the respondents (65%; 17 out of 18 responded) rated these aspects of the Guidelines as “adequate-excellent.” A considerably smaller subgroup (4) rated these as “adequate,” and the smallest group of respondents (2) opted for “poor-adequate.” No respondent chose “poor” or “excellent.”

When asked to identify DSA Guidelines they found most difficult to comply with, respondents listed every guideline in separate comments – but most guidelines appeared in those comments only once. The main reason was that they could “not yet” comply with a given guideline because no preservation plan, no written workflow or no policy for acceptance of file formats was in place at that moment. Apparently they improved the required documentation during the certification process.

**Benefits of DSA certification**

The participants in the NDE-AUDIT working group made an earlier attempt to convince Dutch organizations of the benefits of getting certified, and published a leaflet [18], inspired by the benefits mentioned on the DSA website, although they are no longer mentioned on the DSA website:

- stakeholder confidence
- improvements in communication
- improvement in processes
- transparency
- differentiation from others
- awareness raising about digital preservation.

But how did the DSA certified repositories experience this? 17 respondents out of 18 responded. Clearly the benefit of enhanced “transparency” was rated with the highest significance (8 times as “essential” and 8 times as “considerable”). An *internal* benefit received the highest number of ratings as “considerable” (12): “awareness raising about digital preservation.” Two other benefits that were highly rated are of a more *external* nature: “stakeholder confidence” received 9 ratings as “considerable,” and seven respondents rated “differentiation from others” as a “considerable” benefit.

It is noteworthy that one respondent rated the beneficial effects on “stakeholder confidence” as negligible. In a more detailed question we asked for some more details about the impact of the DSA certification on various aspects of their own organization and
repository. Four aspects scored high on “considerable” in 16 out of 18 respondents: “management’s recognition of the value of long-term preservation and sustained availability of digital assets”, “digital preservation policies” and “technical digital preservation practices” and “the capacity to attract data producers”. Also, the impact on the organization’s professional reputation was rated as “considerable” and “essential.” These ratings confirm the areas where DSA-certification intends to have its strongest impact: by imposing structured, professional and community-driven expectations (guidelines or requirements) on the applicants’ policies and work processes, DSA-certification guides and encourages organizations to describe, document, improve and monitor their essential preservation tasks.

Reportedly, the impact was much smaller in the area of “financial planning” and “allocation of financial resources.” Still, the impact on “allocation of staff” was predominantly rated as “satisfactory.” Given these benefits, would the repository also recommend this type of certification to others within their domain? The vast majority of the respondents (17 responded, 1 skipped the question: 88%) answered affirmatively; they were either ‘certainly’ (9) or ‘very likely’ (6) willing to do so. When queried about their willingness to recommend DSA-certification to their peers, the vast majority of the respondents answered affirmatively, and when asked why they would be willing to do so, most comments characterized the DSA as an instrument to buttress quality assurance efforts. Another strand of motivations related to the quality of the seal as a sign of professional maturity to be showcased to stakeholders and colleagues. The majority of the respondents rated the ratio between investments and benefits as “adequate-rewarding” to “rewarding-excellent”. The final question was related to “lessons learned” or other relevant experiences to pass on to future DSA-applicants. To this question the respondents’ input showed a high level of community sentiment. Most provided a balanced review of their overall experience, placing the required levels of investment on at least an equal footing with the perceived and reported benefits. Key words in respondents’ final remarks hovered around the need for thorough documentation and preparation, significant contributions to professional accomplishment and a tangible boost to quality assurance in the respondents’ work.

3.9 Lessons learnt for the NDE Working Group

The Survey offered enough information to propagate this basic certification to most of the organizations if they reserved some time to get acquainted with DSA. In our opinion the Exploratory Phase is a good description of this. The outcomes of the DSA survey gave a significant insight in the DSA certification process. On all aspects we were investigating, the answers either confirmed our view – like the internal benefit of getting more management attention for digital preservation – as well it gave us more insight in the required effort. There were also some lessons to take with us, as for example to register the time investment. We only need to wait now for some experience of organizations applying for the DSA 2016 version as there is no experience with that yet. One of the requirements that worried the partners in the NDE-AUDIT working group most was the level of required documentation, that seemed to be more than in the DSA 2014-2015 version.

Based on the information of the DSA survey and an educated guess of using the Scoremodel, the following time table was created:

**Expected effort according to Dutch model**

<table>
<thead>
<tr>
<th>1</th>
<th>Initial Self-Assessment based on Scoremodel</th>
<th>2 hrs adding information in model (educated guess) and X hours internal discussion (up to organization)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Exploratory Phase</td>
<td>1-20 hrs based on outcome survey</td>
</tr>
<tr>
<td>3</td>
<td>DSA内部 preparations</td>
<td>50-200 hrs based on outcome survey</td>
</tr>
<tr>
<td></td>
<td>DSA submission activity and peer review discussions</td>
<td>50-100 hrs based on outcome survey</td>
</tr>
</tbody>
</table>

4 CONCLUSIONS

Over the course of 3 years a group of Dutch preservationists collaborating in the NCDD and the NDE worked on popularizing audit and certification in an attempt to professionalize digital preservation. In the Netherlands, several large organizations prepared themselves for certification, resulting in 2 partners receiving the DSA pre-2016 Seal, one partner renewed this seal, 1 partner acquired the DIN/nestor seal as the first in the world en currently 2 partners are preparing for the 2017 DSA version. Apart from these partners in the NDE-AUDIT group several smaller organizations and their suppliers in the Netherlands are preparing themselves, supported by the practical support of the outcomes of the NDE-AUDIT group and meeting each other in regular training sessions organized by the same working group, all traveling along the originally designed Roadmap for Audit and Certification.

ACKNOWLEDGMENTS

A special thanks goes to the partners in the NDE-AUDIT Working group: Margriet van Gorse (National Archive), Annemiek de Jong (Institute for Sound and Vision), Valerij Gillissen (DNS), Robert Gillis (DEN), Ronald van der Steen (Cultural Heritage Inspectorate), Madeleine de Smaele (4TU), Jacob Takema (KB) and Marcel Ras (NCDD).

This work was partially supported by the Ministry of Education, Culture and Science under the National Strategy for Cultural Heritage and with investments of the participating organizations in this working group: KB, DANS, NA, B&G, 4TU, DEN, IISG, NCDD.
REFERENCES

11. Guidance for requirement 9, see Core Trustworthy Data Repositories Requirements, https://drive.google.com/file/d/0B4qnUFYMgSc-eDRSTE53bDUwd28/view
13. https://www.scoremodel.org/nl/home
14. See the analysis of the Archaeology Data Service (UK) from 2011: http://www.dcc.ac.uk/resources/case-studies/ads-dsa; and the more recent report by the Finnish Social Science Data Archive from 2015: http://www.datasealofapproval.org/en/assessment/fid-dsa-case-study
Abstracts

Posters and Demos
Asian Film Archive’s journey in preserving social media as cultural knowledge

Natalie Pang
Wee Kim Wee School of Communication and Information
Nanyang Technological University
Singapore
nlspang@ntu.edu.sg

Joshua Ng
Asian Film Archive
Singapore
joshua@asianfilmarchive.org

ABSTRACT
In this poster, we discuss the topic of preserving a cultural institution’s social media for cultural knowledge. We review approaches associated with preserving social media and apply this analysis by examining Asian Film Archive’s attempt at this. We also look at issues and challenges that have surfaced, as well as offer recommendations in response.

Keywords
Social media, preservation, cultural knowledge, cultural institution

1 INTRODUCTION
Social media platforms have been around since the early 2000s. As interactions on social media platforms take place in high volume and velocity, and technologies continue to evolve rapidly, it is even more pressing than before to preserve digital artefacts as well as knowledge generated on social media.

Arguably, it is those in the archives profession [1] that should have the most scholarly interest in this topic, as well as a heightened awareness of the volatile nature of social media and the need to preserve both the digital artefacts and knowledge emerging through interactions on social media. However, as technologies are usually used in context, this may require archivists to have a deep understanding of the contexts in which they are deployed and used.

Archiving content on social media platforms is a complex endeavour. Firstly, there are certain unique interactions, transactions, and content that are a result of distinct features of each platform. Unlike publicly accessible web pages, content on social media platforms are often hidden and visible only to those on an individual’s social network. Of course, this also raises the question about privacy, the ethics of keeping information, and the blurring of boundaries between public and private records.

Secondly, any attempt to capture social media posts as well as its context is technically challenging. Algorithms cannot do so easily, especially for platforms such as Facebook, which changes its appearances and have complex algorithms driving its newsfeeds and what appears on respective timelines. Thirdly, there is also the question of determining a meaningful timeframe for archiving, as the open and participatory nature of many social media platforms implies that anyone who has access to a post can participate at any point via commenting, sharing, liking, and so on.

2 PRESERVING ASIAN FILM ARCHIVE’S SOCIAL MEDIA
Asian Film Archive (AFA) is a non-profit organisation founded in 2005 and based in Singapore. Its mission is to save, share, and explore the art of Asian cinema. Apart from celluloid film reels, videotapes, and digital film files, AFA also preserves handbills, film magazines, behind-the-scenes photos and footages, scripts, storyboards, and other film paraphernalia.

In wanting to be more than a mere repository and a resource library, AFA actively engages with its stakeholders and audiences through contemporary, innovative, and experimental public programmes [2]. With that, AFA aims to attract and maintain audiences’ interest in film heritage. Beyond film screenings, AFA has recently embarked on curating inter-disciplinary programmes, exploring the intersections of film-visual art, film-music, film-literary world, film-theatre, and film-dance. These programmes have significantly increased the amount of public discourse on social media.

AFA launched its website in April 2005 at www.asianfilmarchive.org, which was captured 187 times over the years by the Internet Archive’s Wayback Machine, with the earliest instance on May 30, 2005. Since the majority of the content was mostly static, the sites have been very well preserved and are still accessible at old.asianfilmarchive.org.

On the social media front, AFA primarily uses Facebook (@asianfilmarchive) and Instagram (@asianfilmarchive). It also has a Twitter account at @AFA_Archive, but the discourse is relatively low, except for some small spikes during international conferences.

In trying to preserve as much discourse as possible, AFA explored the use of paid social web real-time monitoring services like
Brand24, Mention, and Social Express (Brandtology). These services provide similar functionalities at a price range within AFA’s budget. Most importantly, they were the only ones with data export functionalities, which is crucial for preservation. However, as these services were not able to provide historical data more than three months old, we explored other tools and services.

Most of the social media archiving tools that AFA has explored are meant for archiving tweets, such as twarc, TAGS, and Storify. This might be because archiving Twitter is relatively straightforward compared to other social media platforms. However, only capturing tweets would represent a small proportion of the public discourse about AFA’s work. This is because, as mentioned, most of the conversations on Twitter that involve AFA occurred when AFA organized the 19th South East Asia-Pacific Audiovisual Archive Association (SEAPAVAA) Conference in 2016, with the hashtag #SEAPAVAA19.

There was also significant chatter when AFA staff attended international programmes, symposiums, and conferences; for example, the Joint Technical Symposium (#JTS2016), International Association of Film Archives (FIAF) Congress (#FIAF72 and #FIAF2016), FIAF Film Restoration Summer School (#FRSS2016), and Association of Moving Image Archivists conference (#AMIA16).

For a while, lentil [3] by NCSU Libraries looked like a viable tool to archive Instagram posts. But in 2016, Instagram changed their public API significantly. As the search for a better tool to preserve Instagram posts continue, manual archiving using Webrecorder seems to be the only recourse.

A case in point: State of Motion, AFA’s annual programme held in conjunction with Singapore Art Week, features a roving bus-tour exhibition that delves into Singapore’s cinematic history through the critical exploration of historical film locations. Art works are commissioned at each film location. Participants are encouraged to use the hashtag #stateofmotionsg when they post pictures on Instagram, and a total of 330 posts were created. Webrecorder was used to record the Instagram tags explore page, and its access is available for public replay at webrecorder.io [4].

On the Facebook front, the biggest challenge in properly archiving Facebook discourse is maintaining privacy, due to non-public posts. Moreover, Facebook users do not always use hashtags. Interactions with AFA photo or video posts may be indicative of sentiments, which is the main reason Facebook posts are archived. Hence, AFA’s Facebook Page data is downloaded periodically as part of our archival effort.

Previously, AFA experimented with a Facebook group as a platform for dialogue and discussions to continue online after an offline event. Nevertheless, there was hardly any discourse in its two years of existence. The group has since been archived (on March 20, 2017), thus freezing all activities on the group. However, the “download all data” feature provided by Facebook only supports Facebook Pages and not Facebook groups. So again, Webrecorder was used to record all the posts in the group.

Table 1 provides a summary of the tools explored to preserve and archive posts on different social media platforms.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Tool/service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>twarc, TAGS, Storify, Mention, Brand24, Webrecorder</td>
</tr>
<tr>
<td>Facebook</td>
<td>Mention, Brand24, Webrecorder, Facebook</td>
</tr>
<tr>
<td>Instagram</td>
<td>Lentil, Mention, Webrecorder</td>
</tr>
</tbody>
</table>

Table 1: Tools considered for each social media platform

3 NEXT STEPS

AFA is currently working on organizing and consolidating the various archived social media and websites, which will be made publicly accessible via an online catalogue.

Besides that, AFA is exploring the more advanced features of Webrecorder. Other tools and services that are also being considered include Archive Social [5], Social Feed Manager (SFM) [6] and DocNow [7].

To capture social media content as cultural knowledge effectively, there needs to be substantial investment and costs set aside to develop a customized and sustainable solution for preserving and archiving social media. The solution does not merely involve technicalities; there also needs to be technical, archival, and institutional know-how.

4 ACKNOWLEDGEMENTS

The authors were inspired to explore this project because of a previous study on the use of Twitter during a social crisis in Singapore [8]. We would like to thank the AFA staff for their support, especially the Executive Director, Karen Chan, for the permission to document this journey.

REFERENCES

ABSTRACT

This poster presents the initial findings of a collaborative practice-based PhD research project at London South Bank University (London, UK) and Rhizome (New York, US). The project analyses challenges associated with presenting and contextualising internet art in an online archive through the lens of user interaction design. Using methods from human computer interaction (HCI) and user experience design, the project proposes a new design framework for the ArtBase, Rhizome’s online archive of internet art.

KEYWORDS

Online archives, internet art, human computer interaction, user interaction design

ACM Reference format:


1 INTRODUCTION

Changes in web standards and developments in new patterns of interaction between users and computers creates multiple challenges for digital archives and long-term provision of access to digital artefacts of significant cultural value [3]. Curation and presentation in archives of internet art, in particular, pose multiple sets of problems closely associated with the question of what constitutes the art object. Internet artworks are not single digital objects, but rather assemblages, dependent on specific software/hardware environments to be executed and rendered. They oftentimes change over time and require specific user input in order to be performed.

1 The term internet art is contentious. It is used here within the definitions given by Stallabrass (2003) [1] and Greene (2004) [2]. This is to distinguish it from the earlier narrower definition of the term net.art or the later expanded term internet-based art. Other more general terms like born-digital art, digital art and new media art are avoided in order to keep the focus on art which uses digital networked media and is made to be experienced online.
platform in 1996. The ArtBase was established in 1999 as an online space to present and archive internet art, as well as to build a community and a discourse around the works [6]. Initial strategies towards presentation of artworks in the ArtBase reflected contemporary developments in the fields of interaction design and digital preservation. More recently the archival system has struggled to accommodate the growing number and variety of artworks in the ArtBase. Increasingly the focus of preservation efforts at Rhizome has been on building tools to support the presentation of complex artworks with multiple dependencies.

The remote browsing framework, first introduced in Rhizome’s oldweb.today project to emulate old browser environments, has facilitated the online restoration and re-performance of historic internet art works in contemporaneous environments. Furthermore, the capacity to create high-fidelity archives of the dynamic web with Rhizome’s browser-based web archiving tool, Webrecorder, has enabled the preservation of artworks utilising third-party web services.

Remote browsers, in particular, could become a powerful tool allowing presentation of artworks either as a link out of the ArtBase page into a new page running the emulated browser, or as an embedded iframe within the ArtBase page of the artwork. In each of these cases, users would encounter a “browser-within-a-browser” presentation paradigm. Potential challenges here include users mistaking the remote browser environment for other secondary representations (a static screenshot, for instance), or users misinterpreting (or failing to interpret) outdated interaction patterns within the remote browser.

The poster will outline these challenges and showcase initial prototypes for user interactions which facilitate engagement with complex and ‘diffuse’ artworks within the case study scenario of the ArtBase.

3 METHODOLOGY

This project is multidisciplinary, combining qualitative research methods from the fields of the digital humanities, information sciences and HCI. Following common HCI qualitative research approaches, the research work uses ethnographic observation, contextual inquiry and grounded theory towards the design of a new interaction design framework. Additionally, drawing on relevant concepts from contemporary archival theory and curatorial practice, will allow the formation of a concept-driven interaction design approach [8] with an underlying design theory to inform the case study practice.

The initial phase of the research project, which will be visualised in the poster presentation, includes findings from the contextual inquiry for the case study. This phase of discovery and user research involves three key milestones: 1) understanding the organisational context and goals; 2) understanding the domain (of ongoing related R&D or pilot projects at other institutions); 3) understanding user needs and goals. The poster will include qualitative summaries of findings from ethnographic observation, quantitative data based on audits of the current ArtBase data structure, as well as initial interaction design prototypes.

4 CONTRIBUTION TO THE FIELD

Computers and networks have become transparent delivery systems for text-, image- or video-based media which they are able to fully represent and contain. This research project looks to the media-specific properties of computers and networks which can represent and contextualise digital artefacts that are inherently outside of these containable formats.

The research applies methods from HCI to study user behaviours and carry out concept-driven design exploration in order to develop a design framework supporting increasingly complex and ‘diffuse’ artefacts and their related user interaction models. While the design of the framework will answer the specific requirements of Rhizome’s organisational goals and policies, the thesis developed alongside the practice-based research will critically analyse the role of the institutional context in this R&D study. Through evidence gathered from user research and design practice, the thesis will aim to outline new concepts and interaction design principles in the HCI field relevant to the archiving and presentation of artefacts which fall outside easily containable formats.

ACKNOWLEDGMENTS

This PhD research project is supported by an AHRC Collaborative Doctoral Award 2016.

REFERENCES

ABSTRACT
This poster describes current processes and challenges for the preservation and dissemination of video games from the heritage collections of the National Library of France (BnF).

1 INTRODUCTION
The National Library of France holds over 15000 video games in its Audiovisual collections. The mandate to collect, preserve and give access to these games is part of its larger mission to manage the legal deposit of multimedia documents. As defined in the 1992 law that extended the scope of legal deposit in France, multimedia documents are those which include two or more types of document on the same physical media. This definition, however imprecise, has allowed the library latitude to gather all sorts of artifacts of the evolution of interactive publications. Several video games publishers voluntarily sent their products to the BnF from the early 1990s onwards. A more targeted effort has been made since the turn of the century to check the coverage of legal deposit, paralleling the global trend of reevaluating video games as works of art. The legal deposit collections have also been augmented with donations and acquisitions, in order to get coverage of the 1970s and 1980s, and to offer a choice of foreign games to researchers, who may have more difficulty accessing Japanese editions for instance.

It has always been the intention of the library to collect video game content distributed via app stores, console stores and online game stores, but the proportion of downloadable content in the BnF’s holdings is still low, due to the lack of preservation solutions for these digital objects, whose distribution channels are laden with DRMs.

The way the collection has been built at the National Library of France has several consequences for the solutions for preservation and access currently in place, and the opportunities to better its practices.

2 THE COPYING PROCESS
The first analysis concerns the systematic copying process from physical media, to preserve content and prepare for a wide range of access requests in the reading room.

The number of games in the multimedia collections is an estimation, based on the experience of curators and extant bibliographical records:
- 1,500 on floppy discs (Amiga, Atari...)
- 3,000 on cartridges (several generations of Nintendo and Sega gaming consoles, PlayStation Portable and HuCards for PC Engines) and figurines carrying playable content (Skylanders series, etc.)
- 10,500 on optical discs (PC games, Playstations 1 to 4, etc.)

The progress of the copying program depends on:
- availability of hardware and software specific to a platform;
- availability of specific skill sets from library personnel, interns or contractors.
So far, the success rates of the copying process, as reported by the copying software, reaches:

- around 98% of optical discs, out of a total of 31,000 discs copied, with a higher rate of failure on proprietary game content; however, discs from certain proprietary platforms are currently excluded from the campaign since it uses standard PC hardware and robotics (Wii, XBox, Playstation 3, PlayStation 4; around 7% of the collection total);
- around 90% of the games on cartridges, out of 1,301 documents (Sega Master System, Super Nintendo, Nintendo 64 and Nintendo DS);
- about 50% of the games on floppies, out of 4,956 floppies copied, representing 1,385 documents (Atari ST, Amiga, Amstrad CPC and Commodore 64).

There are no systematic tests to make sure the copies are actually readable. When tests have been done, the results are encouraging. Out of 373 Mega Drive games for instance, copied and tested in the summer of 2016, 326 were accessible in an emulator, 27 could not be copied, and 3 were copied but would not run in the test emulator.

### 3 THE EMULATION PROCESS

From the experience gained from making specific selections available for a thematic event, a qualitative analysis of the emulation and delivery process is possible. The BnF has, over the past year, made available a number of games from its heritage collections through emulation for events, both within the library and at the largest French video game convention, the Paris Games Week.

The first step is content selection. An ideal list based on knowledge of the merits and cultural impact of games is cut down by checking it against actual library holdings. Certain titles are unavailable due to lacks in legal deposit or acquisitions, document loss or theft, etc.

The second step is copying the games, when this hasn’t already been done through systematic campaigns or when the copy failed. Some games have been published on multiple platforms, and different editions may be tested, when copying and emulation software is available. When the physical media and the gaming consoles are in good working condition, original access set-ups may be used, but it is not the library’s policy: emulation is the preferred preservation strategy.

The third step is testing the copies using a selection of available emulators. There usually is one prominent emulator able to run most games on each platform; for some games however, it may be useful to test two or more. The emulators’ performance may be hampered by various computer and network safety requirements. Some glitches require extensive testing to be spotted and may only appear at a later phase of the process or when the game is played by the end user. Some games run well in the emulators but still make for a poor gaming experience and thus are removed from the selection: when the emulator runs slowly, or when a contemporary keyboard or gamepad cannot replicate the original interaction, for instance.

The fourth step is setting up the emulators to run with the available selection of PCs and controllers. This is a time consuming process as every action of every game needs to be assigned to a button or key and tested.

### Event 1: Open Days at the BnF, November 2015
- 4 PC games were selected,
- 3 PC games were satisfyingly copied
- 1 alternative edition on Nintendo 64 was selected.

### Event 2: a presentation of historic heroic fantasy games at the BnF, matching a promotional event by Blizzard Entertainment on one of the library’s towers, August 2016
- 14 games were selected, representing 19 different copies (3 were available on two or more platforms)
- 10 games were presented, 1 running on the Hatari emulator, 5 on the DOSBox emulator and 4 on WMWare

### Event 3: the BnF’s stand at the Paris Games Week, October 2017
- 20 games were selected, representing 25 different copies
- 13 games were presented: 1 Amiga, 4 1990s PC, 1 Sega Master System, 3 Sega Saturn, 1 Nintendo 64, 1 PlayStation and 1 Playstation 2

Reasons for excluding games are varied, for instance:
- 4 Atari ST games were selected but were not accessible: 3 were on degraded floppies and content could not be read;
- 1 required a change between two floppies, which the emulator can simulate between two disc images, but it was deemed too complicated outside of a reading room context;
- moreover, the Atari emulator was deemed not ergonomic enough for an event aimed at the general public.
- 1 Amiga game required a keycode input to run; it was not judicious to add an extra step to the access process in the context of a gaming show.

The BnF wishes to improve the information it provides to users and potential users about what can be accessed and in what timeframe. The library intends to continue gathering data about the preservation and dissemination processes to allow it to improve them over time.

---

A Demonstration of ePADD: Computational Analysis Software Facilitating Screening, Browsing, and Access for Historically and Culturally Valuable Email Collections

Extended Abstract

J. Schneider
Stanford University
557 Escondido Mall
Stanford, CA 94305
USA
josh.schneider@stanford.edu

ABSTRACT

ePADD is free and open-source computational analysis software facilitating screening, browsing, and access for historically and culturally significant email collections. The software incorporates techniques from computer science and computational linguistics, including natural language processing, named entity recognition, and other statistical machine learning-associated processes. This demonstration will highlight how these processes enable ePADD to support the appraisal, processing, discovery, and delivery of email archives held by archival repositories and other memory institutions, filling an important role in the preservation of these materials.

CCS CONCEPTS
• Computing Methodologies → Artificial Intelligence; Natural language processing • Computing Methodologies → Machine Learning • Information Systems → World Wide Web; Web applications; Internet communications tools; Email

KEYWORDS
Acquisition, Archival appraisal, Archival processing, Archives, Descriptive metadata, Email, Named entity recognition, Natural language processing, Privacy, Redaction, Screening, Web access

1 ePADD PHASE 2

ePADD Phase 2 began on November 1, 2015 and will end on October 31, 2018. Funded through an US Institute of Museum and Library Services (IMLS) National Leadership Grant for Libraries, Stanford University Libraries, with partners University of Illinois Urbana-Champaign, Harvard University, University of California, Irvine, and Metropolitan New York Library Council, are advancing the formation of a national digital platform by further developing ePADD, free and open-source computational analysis software that allows individuals and institutions to appraise, process, and provide access to email of potential historical or cultural value. During this grant period, Stanford University Libraries and grant partners will continue to improve the program’s scalability, usability, and feature set [1].

Figure 1: Browse options in the Harrison Studio papers - Email Series, Stanford University, ePADD Processing module, 2017 (Version 3.0).

2 DEMONSTRATED MODULES

2.1 Appraisal

Appraisal provides donors, curators, and archivists with a toolset to review and manage an email archive prior to accessioning it to a repository. ePADD can gather email from multiple sources. Upon ingest, ePADD de-duplicates messages, resolves correspondent names from the address book, and extracts fine-grained entities using a custom NER. These functionalities and others enable users to determine the relevance and importance of email messages, identify and flag confidential, restricted, or
legally-protected information, and impose access restrictions prior to transfer.

2.2 Processing

Processing is designed for an archivist to further perform all functions included in the Appraisal module, including scanning for confidential, restricted, or legally-protected information, as well as other tasks that prepare the archive for discovery by and delivery to end users, such as reconciliation of correspondents and extracted entities with established authorities (see Fig. 1).

2.3 Discovery

Discovery is designed to run under a standalone web server, and allows researchers to browse and search a redacted email collection prior to physically traveling to a repository’s reading room to access the full corpus. Only metadata from the processed email archive is published online.

2.4 Delivery

Delivery provides users with access to the full contents of the unrestricted portions of a processed email archive, including attachments, from a managed reading room workstation.

3 DEMONSTRATED FUNCTIONALITIES

3.1 Named Entity Resolution

ePADD uses a custom fine-grained named entity recognizer/classifier that recognizes categories of entities bootstrapped from DBpedia. These include persons, organizations, locations, government entities, political parties, companies, universities, diseases, and awards. ePADD learns from these categories and is also able to recognize likely entities it has not come across before.

3.2 Name Resolution / Correspondent Browsing

ePADD resolves names and email addresses associated with a single correspondent, improving browsing and visualization. All decisions can be manually overridden using a dedicated interface. Mailing lists can similarly be tagged and optionally consolidated using this functionality. Resolved correspondent names can be browsed and graphed alphabetically or by volume of messages exchanged with the email account holder.

3.4 Lexicon Search

ePADD includes tiered thematic keyword searches geared towards broad analysis of a variety of email collections, including lexicons to identify categories of sensitive correspondence. These lexicons can be edited and tuned, or the user can create all new lexicons to suit their research goals.

3.4 Advanced Search

ePADD includes an advanced search interface enabling sophisticated search queries. For instance, users can perform a search for messages containing entities from the disease entity category, or terms from the sensitive lexicon, and further limit this search by mandating that the search should exclude results from a mailing list. In this way a user can create a narrow search for potentially sensitive information to embargo for a specific period of time or to not transfer to a repository.

3.5 Query Generator

ePADD includes a query generator to aid in comparative entity analysis between the archive and any other textual corpus. Matching entities are highlighted and link to message results.

3.6 Bulk Actions and Annotation

ePADD allows the user to apply actions (including marking messages as reviewed, fit for transfer, or fit for embargo) and annotations to sets of messages meeting user-defined criteria, including all messages associated with a given correspondent, all messages from a given date range, all messages containing certain keywords or named entities in the subject or message fields, or some combination of the above.

3.7 Additional Functionalities

ePADD’s additional functionality includes features intended to further support screening for sensitive, confidential, or legally protected materials, as well as features intended to support user access to the intellectual content of the messages. These functionalities include: regular expression search, a redacted view of messages, account and folder-level browsing, built-in visualization tools, and image attachment browsing.

ACKNOWLEDGMENTS

ePADD development is managed by Stanford University’s Department of Special Collections & University Archives, part of Stanford University Libraries [2]. The ePADD development team is composed of Glynn Edwards, Peter Chan, Josh Schneider, and Sudheendra Hangal. Development work is performed with partners at Harvard University, the Metropolitan New York Library Council (METRO), University of Illinois at Urbana-Champaign, and University of California, Irvine. Funding for current ePADD development is provided through an Institute of Museum & Library Studies (IMLS) National Leadership Grant for Libraries. Development for the initial 2015 release of ePADD was primarily funded by the National Historical Publications and Records Commission (NHPRC).

REFERENCES

Digital preservation training needs assessment toolkit: a collaborative development for skills auditing

Sarah Mason  
The Bodleian Libraries  
University of Oxford  
Oxford OX2 0EW  
sarah.mason@bodleian.ox.ac.uk

Lee J. Pretlove  
Cambridge University Library  
University of Cambridge  
Cambridge CB3 9DR  
ljp62@cam.ac.uk

ABSTRACT
This poster is about one of the deliverables of the Polonsky Foundation-funded Digital Preservation at Oxford and Cambridge Project (DPOC), which is to assess the digital preservation skills and training needs within each institution’s workforce. During the review process of existing digital preservation skills and competencies it was found that there was no auditing template to use or adapt; as a result, the DigCurV competencies were reviewed and adapted into a toolkit to enable each institution to measure digital preservation knowledge and skill sets. For the DPOC Project, the results from the completed elements of the toolkit would enable the Fellows to identify potential skill and knowledge gaps in order to develop in-house curricula. For the wider community, the toolkit is intended as a resource for developing digital curation and preservation training programmes.

KEYWORDS
digital stewardship, digital preservation, digital curation, workforce development, skills, training, literature review

1 INTRODUCTION
The DPOC Project is a collaborative two-year project between Bodleian Libraries (University of Oxford) and Cambridge University Library (University of Cambridge), which commenced in August 2016 and is funded by the Polonsky Foundation. The project aims to enhance each institutions’ digital preservation programmes by building on existing expertise and research in the field of digital preservation and curation, forging closer links between key digital preservation and curation advocates to preserve each institution’s digital assets for the researchers of the future. Each institution has three digital preservation Fellows, focusing upon:

   (1) policy and planning development and implementation,
   (2) the development of expertise, training and community outreach, and
   (3) the specification and implementation of technical tools and workflows for digital preservation activities.

As local teams, the Fellows address and support their institutions’ specific digital preservation needs. Activities include auditing current provision, reviewing current good practice and recommending enhancements that might be made to improve the local digital preservation infrastructure (including policy framework, provision of people and skills, and technical infrastructure). The Fellows work collectively as a larger team, while each Fellow works closely with their partner at Oxford or Cambridge ensuring collaboration, enabling an exchange of knowledge between organisations and facilitating shared solutions if these emerge. Fellows are also required to share knowledge to the wider digital preservation community [4].

2 IDENTIFYING AND NARROWING A RESOURCE GAP
At the outset of the project the Outreach and Training Fellows worked collaboratively to develop materials to meet their specific project requirement of carrying out an audit of staff skills with regard to digital preservation. Whilst it was clear that there were existing methodologies from the risk and auditing perspectives within digital preservation (including, but not limited to, DRAMBORA, DSA, TRAC) there did not seem to be a methodology or resource available for auditing personal skills and training needs. With no widely available toolkits to use to measure the maturity levels of staff awareness, understanding or skills related to digital preservation, a deeper literature review was carried out.

2.1 Skills framework and competencies review
Whilst there was an apparent lack of skills audit methodologies, there are many high-level skills frameworks for digital preservation and curation. The DigCurV framework lists skills aimed specifically at executives, managers and practitioners whilst the DigCCurr matrix details skills and competencies for digital curation curriculum development [8, 9]. Several reports, such those produced by the CLIR and NRC have been recently released in the United States discussing skills required in the fields of digital curation and data curation [2, 10]. The skills and competency work of the NDSR has moved towards a greater understanding of the skills required by ‘digital stewards’ and the training methodologies to inculcate such expertise [5].

2.2 Adapting frameworks
While conducting the skills framework and competency review, a training needs assessment was developed using the digital preservation skills frameworks and also traditional ones provided by the ARA and CILIP [3, 6]. Skills from each framework were mapped to each other, shared skills were combined and those that lacked clarity were eliminated. The skills were also mapped to current training programmes at each institution and, where overlaps existed, were removed. The remaining skills were carefully assessed in a literature review before interview questions were drafted. The target interviewees for the initial test of the assessment criteria were practitioners and managers working with respective institutional research repositories. A simplified version of the questions were drafted for an organisation-wide online questionnaire.
3 TRAINING NEEDS ASSESSMENT TOOLKIT

3.1 Developing the toolkit

Individual interviews were conducted at each institution to gather the required information and the resulting data analysed using DigCurV skill descriptors through qualitative and mixed methods. The analysis from the training needs assessment informed the fellows of current skill gaps and training needs; those skills, when mapped to the OAIS Reference Model, will also highlight potential gaps in workflows, which can then be addressed [1]. The approach in using a unified template would also highlight any flaws or differences in the developed interview questions between the institutions. This would help develop a template that could be reused at other institutions.

3.2 Prototype toolkit

The poster describes a prototype digital preservation training needs assessment toolkit, which practitioners can use when assessing their own organisations’ skills, capabilities and training requirements. The toolkit provides questions that the practitioner can feed into an online survey program to send out to institutional staff. Bodleian Libraries used the SurveyMonkey tool and Cambridge University Library used the Qualtrics tool as both had existing institutional accounts. Both institutions were able to use quantitative and qualitative data for comparison despite using two different tools. The toolkit provides face-to-face interview questions and a methodology, which is lacking in other methods and frameworks. Bodleian Libraries used the interview questions in a printed format and developed separate question response form templates for recording answers. Cambridge University Library also used the interview questions in a printed format but recorded their interviews and then transcribed the results into Qualtrics for analysis, having preloaded the questions into the Qualtrics module. To summarise, the toolkit consists of:

1. an organisation-wide digital preservation skills survey questionnaire,
2. practitioner lens semi-structured interview template,
3. manager lens semi-structured interview template,
4. manager and practitioner combined semi-structured interview template,
5. developer semi-structured interview template, and
6. supporting documentation for the surveys and interviews.

4 DPC REVIEW

Whilst the toolkit has been developed and tested within Bodleian Libraries and Cambridge University Library with differing methodologies, the toolkit will also be reviewed by the Digital Preservation Coalition (DPC) to determine its usefulness for the wider digital preservation community. It is hoped that the toolkit will be tested at a variety of DPC member institutions and organisations, which cover the academic, business, charity, cultural, heritage and not-for-profit sectors [7]. As a result of this planned review and further consultation, the toolkit will have been rigorously tested in practice so that it can be universally applicable to as many organisations as possible that wish to develop staff capability and training resources in order to preserve their digital assets.

5 ACKNOWLEDGEMENTS

Our thanks are due to the present and past DPOC Project Managers: Michael Popham, Head of Digital Collections and Preservation Services, Bodleian Libraries, University of Oxford; Dr. Suzanne Paul, Keeper of the Manuscripts and Archives; Tuan Pham, Acting Director of Digital Strategy and Initiatives, Cambridge University Library; Grant Young, former Head of the Digital Content Unit, Cambridge University Library, University of Cambridge; and to all the colleagues who completed the surveys and patiently answered our interview questions. Many thanks to the Polonsky Foundation for funding the two-year collaborative project and to the DPC for their continued support.

REFERENCES

BitCurator NLP Demo: Applying Natural Language Processing to Processing and Accessing Digital Collections

Christopher A. Lee
University of North Carolina at Chapel Hill
callee@unc.edu

Kam Woods
University of North Carolina at Chapel Hill
kamwoods@unc.edu

ABSTRACT
We will demonstrate products of the BitCurator NLP project, which is developing software for LAMs to extract and expose features (e.g., people, places, organizations, events, relationships, topics) in text extracted from born-digital materials. The services and methods can be used by LAM professionals for appraisal and description, as well as facilitating a wider range of access and use scenarios.

CCS CONCEPTS
- Information systems--Digital libraries and archives
- Computing methodologies--Natural language processing
- Security and privacy--Data anonymization and sanitization

ADDITIONAL KEYWORDS AND PHRASES
BitCurator NLP; archival processing; named entity recognition; text processing; topic modeling

INTRODUCTION
Libraries, archives and museums (LAMs) are increasingly called upon to move born-digital materials from their original locations into more sustainable preservation environments. Information professionals must be prepared to extract digital materials from removable media in ways that reflect the rich metadata and ensure the integrity of the materials. They must also support and mediate appropriate access: allowing users to make sense of materials and understand their context, while also preventing inadvertent disclosure of sensitive data.

There has been a significant shift in recent years toward the adoption of digital forensics tools and methods by LAMs, in order to meet the above goals. This process has been facilitated by the BitCurator project (2011-2014), funded by the Andrew W. Mellon Foundation, which has packaged and disseminated an open-source software environment\(^1\) that allows users to create disk images; extract data and metadata from disks or directories; scan bitstreams for the presence of potentially sensitive data values; characterize the contents of disks; and perform other practical tasks, such as scanning for viruses, finding duplicate files, mounting forensically packaged disk images, generating cryptographic hashes, and viewing hexadecimal representations of bitstreams.

The BitCurator Access project (2014-2016), also funded by the Andrew W. Mellon Foundation, investigated mechanisms for providing access to forensically-acquired data. A major product of the project has been BitCurator Access Webtools, which allows users to dynamically navigate filesystems of disk images, as well as searching over the content of many common files types contained within the images.\(^2\) The project also created BitCurator Access Redaction Tools to redact strings and byte sequences identified in disk images.\(^3\)

BITCURATOR NLP PROJECT DEMONSTRATION
BitCurator NLP (2016-2018), funded by the Andrew W. Mellon Foundation and led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS), is developing and disseminating software for identifying, extracting and exposing contextual entities from the wide diversity of born-digital materials that LAMs already hold and continue to receive. This includes helping to identify and explore information based on specific entities (e.g. people, places, organizations, events) of interest to curators and researchers.

There are many existing mature open source natural language processing platforms, including platforms that provide web services and RESTful application programming interfaces (APIs) and integration with industry-standard testing and training corpora. Production-quality open source software toolkits for natural language processing include OpenNLP (Java-based) and NLTK, Pattern, and spaCy (Python-based).

Our target use cases differ from previous work in two fundamental ways. First, disk images are internally complex and require a significant software dependency stack that is already available through the BitCurator environment and BCA Webtools. These include the ability to read, mount and provide access to the contents of various filesystems, as well as extracting, presenting and reporting on their data and metadata.

A second distinguishing factor is that disks may contain a broad range of file types and data encodings, requiring substantial pre-

---

\(^1\) https://wiki.bitcurator.net/

\(^2\) https://github.com/bitcurator/bitcurator-access-webtools

\(^3\) https://github.com/BitCurator/bitcurator-access-redaction
processing to extract content so that it can be processed by NLP tools and organized into meaningful reports, access points and visualizations. BitCurator NLP is building from a variety of existing tools and initiatives to provide services that LAMs can be run independently or integrate into existing software environments and access portals via simple application programming interfaces (APIs). BitCurator NLP is exploring approaches that focus on improving the utility of reports produced about the contents of born-digital collections. Using data extracted from open text using NLP tools, along with techniques from digital forensics research to eliminate or deemphasize those that appear to be irrelevant or common to the system rather than the documents themselves (e.g., names and email addresses of developers or organizations that created the software used to produce a given document), the project team will also develop guidelines describing how to apply the tools in ways that support common access and research use cases. The BitCurator NLP team is ensuring close integration between the existing functionality of the BitCurator environment, BitCurator Access Webtools and the software developed by the BitCurator NLP project. For example, we are increasingly making the various elements of the BitCurator environment available as self-contained software installers (software packages that may be installed in Ubuntu and Debian Linux environments), so users can selectively install and update them as they find most useful. Institutions could load all of the access tools onto the same machine (or virtual machine) as the one they are using for the initial processing, or they could instead decide to run those tasks in different environments in order to better manage and allocate their resources.

ACKNOWLEDGEMENTS
This work has been funded by the Andrew W. Mellon Foundation. The BitCurator NLP team is composed of Jacob Hill, Christopher A. Lee, Sunitha Misra and Kam Woods. We have also benefited from contributions of the project Advisory Board.
Preservation as a Service for Trust
An InterPARES Trust Specification for Preserving Authentic Records in the Cloud

Adam Jansen
University of British Columbia
470-1961 East Mall
Vancouver, BC Canada V6T 1Z1
adamj@mail.ubc.ca

ABSTRACT
Questions regarding trust and authenticity of records stored in the Cloud, as well as the custodial obligations and storage management provided by the service provider of Internet-based records have yet to be resolved; but the adoption of Cloud-based technologies is not waiting for legislation or standards to resolve these issues. Therefore, there is an existing need clearly articulated requirements that provide effective, tested methods for documenting and maintaining the authenticity of records that are removed from creator’s systems and placed into the custody and control of Cloud Service Providers. To address this need, Preservation as a Service for Trust (PaaST), part of the InterPARES Trust research project, is modelling a suite of preservation services that detail specific actions and attributes that capture, create or document metadata and activities that provide supporting evidence of the authenticity of records entrusted to the Cloud Service Providers.

KEYWORDS
Trust; Cloud Computing; Privacy; Access; Digital Preservation; InterPARES; Records

ACM Reference format:
DOI:

1 Introduction
A recent survey conducted by O’Reilly Media revealed that 94% of respondents anticipated migrating to cloud technologies within the next five years [1]. The rapid increase in bandwidth availability, combined with the density increase in hard disk storage following Kryder’s Law [2], has presented new commercial opportunities to level economy-of-scale savings through co-tenancy leveraging of computing resources in centralized datacenter mega-warehouses. These Internet-based service models (collectively referred to as ‘the Cloud’), offer organizations both large and small the potential for lower upfront costs, decreased in-house technical staffing, and easy pay-as-you-go growth on demand. Given these numerous financial incentives, large numbers of both public and private organizations have been embracing the advantages that Cloud Services Providers CSPs offer in order to create, store and access vast amounts in highly centralized, and some would argue highly attractive to hackers, internet-based environment.

Among these organizations that are relying upon CSPs to store and maintain their records are in public institutions, such as banks, public utilities, health care providers, and government departments, that the public has vested with an immense amount of trust and responsibility to protect their personal and sensitive data (e.g. social security numbers, birth dates, etc.). As these public trust institutions adopt these Cloud base services and migrate their records from internally hosted and managed data centers to Cloud-base services, the implications that such a paradigm shift entails is not fully understood. Traditional Records Management questions -- where are the records being stored, how are they being managed, where are all the copies hard disk, tape or otherwise – often are not asked or do not have answers. The very definition of the Cloud allows for dynamic and elastic provisioning, allowing for the rapid relocation and allocation of resources from a datacenter in one location to another (potentially in another country [3].

This global system of interconnected presents the issue of records from one jurisdiction – and, therefore, a specific set of record-related laws and regulations – can rapidly, fluidly and without knowledge of the records owner, to move into a storage location that resides within another jurisdiction – and a potentially conflicting or less favorable set of records access/disclosure laws – such as [4]. The legal liability for any damages that occur as a result of any security breaches of the CSP is either unclear or, if the basic service level agreement of any major Cloud storage provide is any indication, the reasonability of the record provider. Additionally, the expected response from the CSP in the event of disclosure, subpoena and access rules, regulations, and law regarding these records stored in the Cloud are unknowns. From an evidential perspective, a major area of concern when utilizing CSPs to store records of important legal value is whether those records, once entrusted to a CSP, can be trusted after they leave control of the creating organization [5]. Should those records be needed again, will sufficient documentation exist to establish a detailed chain of custody of those records have been created and accessible to establish the authenticity of the record retrieved from the CSP – from the time they left the creator’s control through all
the movement within the CSPs mega-infrastructure and who accessed what and when? Whose responsibility is it to create and produce such documentation? To address these concerns of authenticity and legal admissibility, records creators and CSPs must work together to create the appropriate procedures and mechanisms to ensure that as records are transferred and/or moved from one location to another that they remain, and can continue to be proven to be, authentic.

2 Authenticity of Digital Records in Cyberspace

Records are a specific sub-set of data that are defined as any “document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference” [6]. A record is considered to be authentic when its identity can be established and its integrity can be demonstrated. The identity of the record is derived from those attributes that uniquely characterize that record and are used to help distinguish one record from others that participated in the same, or similar, activities. The integrity of a record refers to the completeness of the record, in that the record possesses all of the its necessary parts to convey the message for which it was created and its condition is unimpaired [7]. Authenticity of records, from an Archival Science perspective, encompasses the entire context in which the records were created, accessed, stored and managed, from the moment of their creation through their entire life and eventual disposition. This is important in that records, as byproducts of the activity they provide evidence of, are granted a special allowance under the hearsay rules [8] that allows them to be submitted as evidence of the activity that created them. Therefore, in order for a record to continue to serve as a faithful witness of the activity that created it, it must remain demonstrably authentic; that is, it must be what it purports to be, free from any manipulation, substitution or falsification.

The presumption of authenticity is afforded to records when they are created to serve administrative needs during the usual and ordinary course of business. This presumption is strongly influenced by the methods and means of its creation, handling and chain of custody. When a record is moved across space (e.g. sent from one storage location to another over a network) or through time (e.g. set aside for later retrieval), the message for which the record was created must not be substantially altered in the process. Retaining the authenticity of records past their creation necessitates that those records be created, managed and stored in accordance with regular, documented procedures that can be attested to through an unbroken chain of custody [9]. When the records have been removed from the original system of creation, or passed onto a third-party custodian, documenting and providing evidence of how these records were stored and transmitted across space and through time becomes increasingly important, as important evidential metadata that supports that records authenticity is often lost in such movement. The stronger and more documented the procedures used in the handling, transfer and storage of the records, the stronger the record’s presumption of authenticity that can be afforded to that record [10]. On the other hand, when a record is transferred into the care of a third-party custodian without documenting the procedures used or chain of custody, it becomes difficult to create post-facto to provide sufficient evidence of the identity and integrity of that record to support its presumption of authenticity. Once a record’s presumption of authenticity is lost, it is nearly impossible to reassert.

3 Trust and Digital Records in an Increasingly Networked Society

The InterPARES Trust (ITrust) research project, under Project Director Dr. Luciana Duranti of the University of British Columbia and funded by the Social Sciences and Humanities Research Council of Canada (SSHRC), is exploring the trustworthiness of digital records uploaded to the Cloud with the goal to:

...generate the theoretical and methodological frameworks that will support the development of integrated and consistent local, national and international networks of policies, procedures, regulations, standards and legislation concerning digital records entrusted to the Internet, to ensure public trust grounded on evidence of good governance, a strong digital economy, and a persistent digital memory [11].

The ITrust research team represents public and private institutions and universities from around the world with subject matter expertise in archival science, records management, diplomacy, law, information technology, communication and media, e-commerce, health informatics, cybersecurity, information governance and assurance, digital forensics, computer engineering, and information policy. The project has been organized into four regional teams (North America, Latin America, Europe, Asia) and a Multinational Organization team that are each focusing on a specific area of research that leverages their collective areas of expertise and geo-political environment. Represented amongst the larger institutions participating in the ITrust research project are: British Library, European Commission, International Federation of Red Cross and Red Crescent Societies, International Monetary Fund, International Records Management Trust, National Institute of Standards and Technology, NATO, UNESCO, University of British Columbia, University College London, and University of Washington. In order to ensure that the cross-disciplinary nature of the researchers utilize a common foundation for across all the teams, the ITrust project will build upon the findings of the first three phases of the InterPARES research project (1998-2012) by expanding upon those findings with additional case study research, current literature, legislation and regulatory review, and exploratory research.

The ITrust project has been organized into five primary domain areas considered to be of particular interest to the creation, handling, management and storage of digital records in Cloud-based environments, as well as five cross-domain areas.
The five research domains are:

- **Infrastructure:** This domain considers issues relating to system architecture and related infrastructure as they affect records held in online environments. Examples of areas to be investigated include such topics as: types of cloud and their reliability; types of contractual agreements (service level agreements or SLAs) and their negotiation, coverage, flexibility, etc.; costs, up front and hidden.

- **Security:** The security domain considers records issues relating to online data security, including: security methods (encryption, sharding, obfuscation, geographic location); data breaches; cybercrime; risks associated with shared servers; information assurance; governance; audits and auditability; forensic readiness; risk assessment; and backup.

- **Control:** The control domain differs from the security domain in its focus on the management of digital material in online environments. It addresses such issues as: authenticity; reliability, and accuracy of data; integrity metadata; chain of custody; retention and disposition; transfer and acquisition; intellectual control, and access controls.

- **Access:** The access domain researches open access/open data; the right to know/duty to remember/right to be forgotten; privacy; accountability; and transparency.

- **Legal:** The legal domain considers issues such as: the application of legal privilege (including the issue of extra-territoriality); legal hold; chain of evidence; authentication of evidence offered at trial; certification; and soft laws (in particular UN standard-setting instruments) - mapping, scope, potential impact, and constraints;

and the five research cross-domains are:

- **Terminology:** This cross-domain is concerned with the ongoing production of a multilingual glossary; a multilingual dictionary with sources; ontologies as needed; and essays explaining the use of terms and concepts within the project.

- **Resources:** This cross-domain is concerned with the ongoing production of annotated bibliographies, identifying relevant published articles, books, etc., case law, policies, statutes, standards, blogs and similar grey literature.

- **Policy:** The policy cross-domain considers policy-related issues emerging from the five research domains; for instance, it would cover policy issues pertaining to the development and implementation of the 'infrastructure' or 'security' standards, or as the facilitator for the implementation of laws. In general, it addresses recordkeeping issues associated with the development and implementation of policies having an impact on the management of records in an online environment; policies can be broad, such as a national policy on information management, or very specific, such as a policy on adopting certain standards within an organization.

- **Social/Societal Issues:** This cross domain is concerned with the analysis of social change consequent to the use of the Internet, including but not limited to use/misuse of social media of all types, trustworthiness of news, data leaks (intentional or accidental/force majeure) consequences, development issues (power balance in a global perspective), organizational culture issues, and individual behaviour issues.

- **Education:** This cross-domain is concerned with the development of different models of curricula for transmitting the new knowledge produced by the project [12].

### 4 Preservation Services for Online Environments

With the adoption rate of Cloud services outpacing legislation and case law, there exists a strong need for a clearly articulated set of functional requirements defining records-related services that support the presumption of authenticity within an online environment. Under the ITrust Control Domain, **Preservation as a Service for Trust** (PaasT) seeks to develop a preservation model that expresses actions and attributes capable of supporting the authenticity of records that are created, managed or stored within Internet-based environments. The purpose of PaasT [13] is to: ...provide insight and guidance to both those who entrust records to the Internet and those who provide Internet services for the records. The project will address relevant requirements, insights and concerns developed in other ITrust projects to enrich and strengthen its models. To provide a strong foundation for the proposed preservation services, the PaasT project team is leveraging the Chain of Preservation [14] model developed by InterPARES2. The CoP model stipulates that preservation activities begin with the creation of the record and must be continuously managed throughout the lifespan of that record. As a record moves from creation to active and then inactive stages of its lifecycle, the actions and attributes that are needed at a specific stage to support the record’s authenticity also change.

As the services are being written with Cloud Based Services in mind, rather than speaking in terms of preservation environments, the services use the concept of Preservation Environment. As such, PaasT introduces a new set of terminology and information concepts that borrow and adapt from existing standards (OAIS, for example). This concept refers to the highest level of set of Preservation Targets (those objects that are to be preserved by the Preserving Party) under the
Preservation Rules, together with the technological infrastructure and tools necessary to perform the functions specified in the services. The services are structure to be performed independently, be a single provider, or 'farmed' out to a series of providers depending on the needs of the organization. PaaST requirements address the preservation of digitally stored information and at the top level of the hierarchy is the Information Object. To provide a very brief overview of the hierarchy of information objects as viewed by PaaST (see Figure 1: Classes of Information Objects). Information Objects are comprised of Data Objects and Intellectual Objects. Data Objects are what are traditionally understood as the 'digital file', or an ordered set of numbers, characters, signs or other information encoded as binary bits. The Intellectual Object, on the other hand, is a human recognized object (the 'Record') comprised of one or more data objects along with related information and preservation targets. Related Information contains description information about the record in the form of submission information, preservation description information, preservation rules, or heuristic information. Finally, the Preservation Target is the focus of the preservation operation, and is comprised of the zero or more data objects and zero or more related information. Below this Preservation Target are additional Archival Aggregation Objects that are beyond the scope of this paper, but these concepts (as well as those briefly touched on above) are covered in much greater detail with the PaaST specification.

4.1 Preservation Services

To reflect this changing nature, PaaST has identified four distinct services to be modelled:

4.1.1 Receive Submission. The Receive Submission Service ensures that a set of records transferred to an Internet-based environment is complete and intact, and is in compliance with any agreements that are in force between the transferring party and the receiving preservation party (e.g. a Cloud Service Provider).

4.1.2 Preservation Storage. The Preservation Storage service captures, reports and makes available those attributes concerning the storage of the records, the movement within the storage system, and the replacement or upgrade of storage media and related technologies.

4.1.3 Preservation Change. The Preservation Change service captures, reports and makes available those attributes related to the migration, conversion or transformation of the digital objects that constitute a preserved record or the software used to translate the digital bits into a human readable form.

4.1.4 Access. The Access service provides retrieval and production of copies of records and assessment of the authenticity of the copies provided to the requestor.

4.2 Supporting Capabilities

In addition to the Preservation Services, PaaST will provide three categories of supporting capabilities to supplement the Preservation Services.

4.2.1 Management. The Management category is comprised of the Control and Problem Handling capabilities

4.2.1.1 Control. The Control capability is responsible for the determination of which Preservation Rules and Conditions apply to specific cases and enforces said conditions on objects, processes, parties and information.

4.2.1.2 Problem Handling. The Problem Handling capability is responsible for recognizing problems related to objects, processes, parties and information, for characterizing and rating the severity of the problem, assigning resolution of the problem to specific party and tracking problem resolution.

4.2.2 Information Processing. The Information Processing category is comprised of the Information Management, Reporting, Class Definition, Composition Definition, Characterization, Permanent Feature Designation, and Assignment capabilities.

4.2.2.1 Information Management. The Information Management capability is responsible for creating and maintaining Preservation Management Information about the controls, objects processes and parties. Includes categorizing information, extracting data from information sources, generating data from preservation actions and collecting data from inspecting and verification of preservation objects.

4.2.2.2 Reporting. The Reporting capability is responsible for producing, sending and managing reporting functions about objects, processes, parties, and problems.

4.2.2.3 Class Definition. The Class Definition capability is responsible for defining the composition of objects and their features, as well as establishing the conceptual framework for managing the Preservation Targets along with Related and Linked Objects.

4.2.2.4 Composition Definition. The Composition Definition capability is responsible for identifying the composition of Submission Sets and Preservation Targets and confirming that they meet the specified criteria.

4.2.2.5 Characterization. The Characterization capability is responsible for specifying those features that either individual or
sets of Preservation Targets possesses, or should possess, and, optionally, the values of those features. Information about the Preservation Targets may be derived from a variety of sources, including, but not limited to, Preservation Agreements, Preservation Service Contracts, Submission Information, Linked and Related Data Objects, information derived from Inspection, Verification and/or Authenticity Assessment.

4.2.2.6 Permanent Feature Designation. The Permanent Feature Designation capability is responsible for determining the essential requirements for preservation of a class set, individual Preservation Target, or Preservation Data Object by identifying those features that must remain unchanged throughout the preservation storage and retrieval process.

4.2.2.7 Assignment. The Assignment capability is responsible for assigning individual objects to categories. Categories are assigned via criteria as defined in Class Definition and Composition Definition capabilities based on information about the object and captured as Preservation Management Information.

4.2.3 Object Processing. The Object Processing category includes the capabilities of Inspection, Verification and Authenticity Assessment.

4.2.3.1 Inspection. The Inspection capability is responsible for specifying the methods to be used to examine Data Objects in order to identify the components of a composite object, such a Preservation Object, Preservation Aggregate, or Submission Set, or to determine whether the Preservation Target under examination has a particular feature or specific value for a given feature.

4.2.3.2 Verification. The Verification capability is responsible for providing confirmation of existence and values of features of Preservation Targets by comparing information from different sources or information obtained by inspection at different times; also responsible for verifying the success of Preservation Processes, such as Submission and Change.

4.2.3.3 Authenticity Assessment. The Authenticity Assessment capability is responsible for determining the authenticity of Preservation Targets at the time it enters the Preservation Environment, capturing data about the authenticity of Preservation Objects, comparing authenticity related data, and reporting discrepancies in authenticity data.

5 Specification to Standardization

The penultimate goal of the PaaST project is to release the specification to a standards body to have it reviewed, analyzed, and, ideally, approved as an internationally agreed upon standard. To realize this goal, InterPARES has joined the Object Management Group (OMG) and will be working as a member of this standards body to introduce and advance the PaaST specification. The OMG is an international, non-profit technology centric standards consortium whose mission to to: develop, with our worldwide membership, enterprise integration standards that provide real-world value. OMG is also dedicated to bringing together end-users, government agencies, universities, and research institutions in our communities of practice to share experiences in transitioning to new management and technology approaches like Cloud Computing [15].

In support of this mission, OMG hosts organizations such as the Cloud Standards Customer Council (CSCC) and the Consortium of IT Software Quality (CISQ) at its quarterly technical meetings in order to: increase industry exposure to technical specifications that are working their way through the OMG approval process, foster cross-sector collaboration, and encourage inter-domain knowledge sharing between organizations.

The standardization process developed by OMG differs from that used by most other standards bodies in that OMG employs a strict "No Shelf-ware" policy; that means that all specifications that are submitted to the OMG for review must have a working product that has been created in accordance with the specification, and therefore validates the clarity and comprehensiveness of the specification, before it will be approved. This requirement to test the implementability of the specification ensures that, upon approval, that the standard is immediately usable without further modification – or that it won’t just ‘sit on the shelf’. OMG support for specifications continue after approval as well, with OMG producing educational book, training workshops, certification mechanisms. Among the better known OMG approved specifications are several modelling languages widely used within the software and system development sector: Unified Modeling Language (UML), System Modeling Language (SysML), and Model Driven Architecture (MDA).

Working closely with both public and private sector organizations allows for small vertical industry-oriented standards bodies, consortia and other groups (such as research projects like InterPARES) to work alongside the OMG to create and test the metamodels, Applications Program Interfaces (APIs) and other types of specifications that are designed by, and meant by, specific sectors or industries. While the OMG has focused predominately on producing highly technical and widely used specifications that have cross industry applications (such as UML), as a whole OMG relies upon input and feedback from current and new consortia members to address emerging challenges that affect specific sectors that might be outside their normal purview – such as the challenges faced when storing and preserving digital records in Cloud-based environments that ITTrust is researching. In order to foster such cross-pollination of sector-specific knowledge and experience, the OMG maintains reciprocal membership agreements and exchanges with industry organizations in encourage industry specific organizations to bring their challenges and concerns to the OMG; among the groups with cross memberships exchange are: Association of Information and Image Managers (AIIM), Open GIS Consortium, Integrated Justice Information Systems (IJS) Institute, and World Wide Web Consortium (W3C).

Among the benefits to ITTrust of working with the OMG to develop PaaST into a publicly available specification is that, in addition to having access to and review by the many
professional and private organizations that participate in the OMG, the OMG also maintains close working relationships with other major global standards bodies in order to reduce duplicative efforts. By maintaining a formal liaison with other standards bodies that publish in similar areas, the OMG can work in concert with these other standards bodies to reduce the number of redundant and occasional conflicting standards issued by different standards bodies. Among the groups with which the OMG works are: International Organization for Standardization (ISO), European Computer Manufacturers Associations (ECMA), Institute for Electrical and Electronics Engineers (IEEE), and two Accredited Standards Committee (ASC) committees -- X12 (electronic data interchange) and TIM1 (network management). Of particular importance to the PaaST project is the special relationship that the OMG has with International Organization on Standards (ISO). Specifications that are approved by the OMG are recognized by ISO as Publicly Available Specifications; this special recognition allows them to be fast tracked by the ISO Committee on Information Technology Standards (ISO/IEC JTC1) directly onto a final ballot for approval for ratification as an ISO standard.

6 Next Steps for PaaST

The Preservation Services and supporting capabilities that are part of the specification comprising PaaST are still in the design phase of a two-year development cycle. An initial draft of the functional specification for the services and supporting capabilities has been creating, along with functional specifications detailing required and optional functionality. Throughout the development process, researchers from other domains within the ITrust project provide feedback on the PaaST work product as it impacts areas of their own research, as well as propose additional services and/or functional specifications as the needs are identified. Once the ITrust review has been completed, PaaST will be formatted as an OMG Request for Proposal and forwarded to the OMG’S Government Information Sharing and Services Domain Task Force for review. The OMG RFP will provide all the information necessary to software developers interested in creating a functional application that performs all the operations that are detailed within the suite of Preservation Services – i.e. create software based on the PaaST specifications with nothing more than the RFP as guidance. The RFP, as stated in The OMG Hitchhiker’s Guide [16], is:

... a statement of industry need and an invitation to the software supplier community to provide a solution, based upon requirements stated within. The process of identifying need is a culmination of experience within an OMG technical group...and solicitation of industry recommendation. While the RFP is not prescriptive in the sense of dictating how the solution is presented, it does provide guidelines – requirements – that again are derived from the sources noted above.

Contained with the PaaST RFP will be functional requirements, pre-conditions that must exist for a service to functional, and main and alternate workflows for each of the preservation services; along with appropriate UML Class diagrams of the methods and attributes corresponding to the functional requirements as well as any other supporting material deemed to be helpful to software developers. Should the RFP meet the high standards set by the OMG set by the OMG GovDTF and be approved for released, the RFP will be issued and any OMG member organization may develop and submit a package based on the RFP for evaluation by the group. Based on the quality of the submissions received, the RFP will then be forwarded for full approval as an OMG specification, or require further revision to address any shortcomings that were discovered by the developers who attempted to implement the specification as written.

7 Conclusion

The objective of InterPARES Trust project is to generate the methodological and theoretical frameworks necessary to support the development of an integrated network of policies, procedures, regulations standards and legislation that can be applied consistently across the broad spectrum of juridical boundaries that exist in the study. The goal is to increase public trust in records stored with Cloud based providers by creating such frameworks grounded on evidence of good governance, a strong digital economy, and a persistent digital memory. In support of that goal, the Preservation as a Service for Trust (PaaST) project is developing a series of preservation services that supports the presumption of authenticity of records entrusted to the Cloud-based Service Providers. These preservation services detail those actions and attributes that need to be documented as records are moved through space, such as transmitted from the creator to Cloud-based Service Provider, or across time, such as being stored in the Cloud for an extended period of time. By implementing the preservation services articulated by PaaST into a Cloud-based storage environment, the record keeping system will capture and document metadata that allows for the identity of that record to be established and its integrity demonstrated within a documented chain of custody. To ensure that the PaaSST preservation services can be fully integrated into existing Cloud-based environments, InterPARES is partnering with the Object Management Group to developed PaaSST into a working prototype for assessment by the GovDTF and, if found to be accurate, complete and implementable, approved by OMG as a Publicly Available Specification.

Acknowledgements

PaaSST has been a major undertaking, started in 2014, that has involved the efforts of many brilliant minds: researchers Dr. Luciana Duranti, Dr. Kenneth Thibodeau, Dr. Giovanni Michetti, Dr. Corrine Rogers, Dr. Joseph Tennis, Adam Jansen, Courtney Mumma, and Daryll Prescott, Katherine Timms, and Graduate Research Assistants Lois Evans, Mel Leverich and Sylia Seller. Primary funding for The InterPARES Trust Project has been provided by the Social Sciences and Humanities Research Council (SSHRC) of Canada through a Partnership Grant. Matching funds
for the project have been provided by The University of British Columbia’s Vice President of Research, the Dean of Arts, and the School of Library, Archival and Information Studies. Additional matching support in kind has provided by the over 70 partners participating in this research project. InterPARES Trust is also a partner on the SSHRC-funded Participedia Project led by Dr Warren of the UBC Department of Political Science.

References

How collaboration and sharing has made OAIS compliant archiving at (very) small archives possible

Kathrine Hougaard Edsen Johansen
Copenhagen City Archives/NEA
Københavns Rådhus,
1599 København V
Denmark
f27b@kff.kk.dk

Mads Neuhard
Copenhagen City Archives/NEA
Københavns Rådhus,
1599 København V
Denmark
madsne@kff.kk.dk

ABSTRACT
Can small archives with limited budgets and only a handful of employees do professional OAIS compliant digital archiving? In Denmark 26 small public archives have joined forces and found a way to do it. This poster describes how it is made possible through close collaboration and sharing within a network organization.

KEYWORDS
Digital Archives, infrastructure, small archives, collaboration, network organization, digital preservation, community, institutional challenges

1 INTRODUCTION
Not only does Denmark have a large public sector, but its archival legislation is also one of the most extensive of its kind. The legislation specifies in detail what, how and when to archive and it gives far-reaching powers to the archives. This is particularly true for born-digital material [1]. Danish Public archives carry out thorough appraisal of all born-digital material created by administrative bodies from the public sector. Sometimes appraisal is done as early as when an IT-system is commissioned. For materials that should be archived, producers are obliged by law to submit data to a public archive in a format decided by the archive, with a deadline set by the archive. All this is done at the cost of the producers. With regular submissions approximately every five years and a national archival strategy based on normalization before submission, it is evident that the archival responsibility enforced by law on producers is significant in terms of both extent and resources needed. This is in many ways an ideal situation for the archives as it gives them the power to enforce archival interests and ensure that material of value is transferred to archives for long-term preservation.

The empowerment of archives combined with the large public sector results in large quantities of data to archive and a functioning archiving process that ensures regular submissions of data. Submissions are made to public archives, which in a Danish Context covers all archives operating under the Danish Archival Act.[2] The largest archive is The Danish National Archives[3], but the landscape of public archives in Denmark also covers small public archives at municipality level. These are run independently by each municipality and often with a limited budget, and only a couple of employees.

2 CHALLENGE
Archival legislation empowers public archives and gives them authority to place demands on producers. At the same time this places demands on the archives themselves. It requires the archives to do their part of the archiving, which is no small task. Public archives are obliged to do appraisals, specify how each information package should be created and transferred, validate transferred information packages and ingest them into archival storage, do preservation planning, manage collections, provide access etc.. In short they have to run an OAIS compliant archive. This responsibility lies on all public archives that act under the Danish Archives Act. Being a public archive they per se commit to run an OAIS compliant archive. This is no insignificant task. It requires volume, expertise and resources and is challenging for all archives and even more so for small public archives with limited budget and size.

3 THE SOLUTION: COLLABORATION AND SHARING

3.1 Facing The Challenge Together
Small public archives all face the same challenge of being obligated to run a full scale OAIS, but lacking the volume, expertise and budget to do so. To this end, seven small public archives joined forces in 2007 and created a network organization, in which the small archives collaborate closely and take advantage of economies of scale. The network, named NEA (Network Electronic Archiving)[4] is located at Copenhagen City Archives, which is the largest member archive in NEA. The Network has grown steadily since 2007 and now comprises 26 small public archives.

3.2 The NEA Concept
NEA is based on close collaboration and sharing. The basic idea is that specialized knowledge, expertise and facilities are shared.
Instead of each archive employing specialized staff and building its own infrastructure for preservation, archives share this via NEA.

Member archives of NEA are run individually and each archive is responsible for its own business. However, the day-to-day operations are supported by NEA to the extent needed by each archive.

At the heart of NEA is a shared group of experts that have the experience, skills and expertise needed to do OAIS Compliant archiving. Members can draw on the expertise and knowledge of the shared staff as needed on a consultancy basis. NEA staff support member archives in all aspects of their day-to-day operations including appraisals, negotiations with producers, ingest and access. Each member has one primary consultant but can draw on all experts as needed.

NEA also provides a shared archival infrastructure, which means that each archive avoid having to make their own set-up. Management, archival storage and preservation planning are handled centrally in NEA, but data is clearly separated. Each archive only has access to its own data. Handling of data is contractually agreed on between NEA and each member archive in separate data processing agreements.

In addition to the services and support offered via NEA, member archives have a strong community where archives benefit and learn from each other. Because all member archives are using the same infrastructure collaboration is easy and often successful. Member archives can directly implement or build on experiences from other member archives.

3.3 The NEA Business Model

Because member archives are run individually it is essential that the solution is flexible and adaptable. NEA offers different “service packages” with varying levels of support. Member archives can tailor the support from NEA to their own individual situations based on their needs, their ambitions and their own level of expertise. By buying into service packages, member archives can get the expertise and facilities they need to create and complete their own OAIS compliant archive.

4 PROOF OF CONCEPT

In 2017 NEA celebrates its 10th anniversary. Looking back at the past 10 years NEA has proved its worth, having more than tripled in size. TBs of born-digital material have been archived and the small archives are doing professional digital archiving. This is only possible because of the network organization and the extensive collaboration and sharing.

In the 10 years NEA has existed, similar network organizations for small public archives have been formed in Denmark. This means NEA is no longer the only network of its kind in the country. However, it is the first of its kind and it is by far the largest.

The emergence of other networks only adds to the success of NEA. It shows that the basic concept of collaboration, sharing and economies of scale is attractive and viable. The different network organizations all have slightly different business models, but they all build on the same basic concept.

5 CONCLUSION

Through NEA, small public archives in Denmark are able to do full scale OAIS-compliant digital archiving and meet the legal requirements in spite of limited budgets, volumes and expertise. This is possible because of extensive collaboration and sharing. Instead of employing specialized staff and building infrastructure for preservation at each archive, this is shared via NEA. This exploits economies of scale and allows small public archives to run full scale OAISs using significantly fewer resources.

REFERENCES

Technical aspects from the Polonsky Digital Preservation Programme - The story so far at The Bodleian Libraries and Cambridge University Library

James Mooney
The Bodleian Libraries
Oxford OX2 0EW
james.mooney@bodleian.ox.ac.uk

Dr. David Gerrard
Cambridge University Library
Cambridge CB3 9DR
dg509@cam.ac.uk

ABSTRACT
As part of the two-year Polonsky digital preservation research project [21], the Bodleian Libraries (the University of Oxford) and Cambridge University Library (CUL) are researching and developing requirements for digital-preservation-specific services. Part of the project concerns gathering technical requirements for long-term digital content repository systems; this has included reviewing our own infrastructures, surveying our digitized collections and existing repositories, visiting a number of other institutions and assessing software from various vendors.

Our poster highlights the different challenges both institutions face with their current systems, the collaboration relating to auditing and reporting software used, the work which has already completed, and what is planned for rest of 2017-2018.

KEYWORDS
audit, dspace, droid, fedora, jhove, review, repository

1 THE BODLEIAN LIBRARIES
At Oxford, our initial review has focused upon four main areas within our institution: Imaging Services, Oxford’s Research Archives (ORA and ORA-Data [7]), Digital.Bodleian [2] and project-driven websites.

Our imaging services department have created over six million images and use Goobi workflow management software [6] to manage their day-to-day workflows, allowing for preservation actions to be incorporated such as the creation of checksums and validation checks. For 2D digitization the TIFF 6.0 format is currently used; previously using 5.0. These master TIFF files are stored to tape, along with a MD5 checksum. Due to legacy file naming conventions for digitized content additional software is used to track linkages between file names and associated metadata. Utilizing a repository environment to manage master image files in the future will help mitigate this preservation risk.

As part of our survey we reviewed this digitized content, one challenge was restoring 100+ Tb of content from tape to spinning disk. As part of ongoing infrastructure improvements, a 600 Tb RAID disk array was available to store the retrieved data, allowing us to then use parallel JHOVE [3] processes characterize, validate the large number of TIFF files. Transforming the JHOVE XML output with XSLT allowed us to import into Qlik Sense [10], a data visualization application with which reports such as cumulative file size over time could be created to help us predict future storage requirements. In addition SHA256 checksums were calculated outside of JHOVE for ongoing fixity checks along with image fingerprinting with Python image hash libraries to identify near duplicates.

Oxford University Research Archives (ORA) is our institutional repository for scholarly research output. ORA preserves an array of research publications, journal articles, conference papers, working papers, theses, reports and more. There is also an additional repository, ORA-Data, designed to help researchers archive, share and cite research data. Both these repositories based on the Fedora repository software, with ORA having previously migrated from Fedora 2 to 3, one major challenge is migrating to the latest version of Fedora [13].

By using regular automated DROID [12] scans of the repositories, we have also identified the current file formats in use and built a format risk registry. We have established that approximately 52% of content within ORA-Data is unknown, which is understandable as this is research data output. We will engage with our research community to help create new file signatures for these and submit to the PRONOM file registry [20], so we would like to see the number of unknown formats decrease. SHA-256 checksums generated with DROID have let us report on duplicates, and comparing checksums with previous scans will help to identify any fixity issues.

Digital.Bodleian is a repository bringing together our discrete digitized collections under a single user interface. This service currently hosts approximately 700,000 images online. Originally Digital.Bodleian utilized customized repository software based on Fedora 2, and like our Research Archives, Digital.Bodleian will migrate to latest Fedora repository system. Standardizing on the same repository software will enable more code reuse and many of the same preservation actions will be used across to both systems.

We also have approximately 40 project-driven websites containing digitized content from the last 20+ years, using a variety of different frameworks, databases, metadata formats and user interfaces. There is no common search interface, making content discovery and navigation across collections difficult. Only a few collections offer a machine-readable interface, or any way to link their data with similar data in other Bodleian collections, or with collections at other institutions. These project sites vary in size from 127 million compressed JPEG2000 files, which were created as part of a Google Books project [4], to earlier projects which contain just over thousand JPEG images. Approximately 2.5 million other digitized images have also been made available via these sites. This presents a major challenge, the sites need to continue to be supported until such time that the content can be migrated to Digital.Bodleian, however we face some problems with extracting various forms of metadata from the legacy sites.
One of the next areas of focus will be around the current storage infrastructure. Our services currently have a minimum of three copies of data, one on ultra- resilient spinning disk storage which is mirrored between two data centers, then an additional two copies stored on tape in two locations. However, at present these copies all lie within Oxford’s city boundaries. Therefore we will recommend that for our repository environments, two copies will be stored on spinning disk ideally on different technology stacks in more geographically diverse sites, along with further copies of these stored on tape [11]. Due to the extensive infrastructure already in place at Oxford, we have not yet considered 3rd party cloud storage; however this is one of a number of options we are still reviewing.

2 CAMBRIDGE UNIVERSITY LIBRARY

Like our colleagues at Oxford, the Polonsky Fellows at Cambridge University Library have also used DROID to scan the University of Cambridge’s DSpace [5] research data repository (Apollo [8]) to gather information about the files stored there. In October 2016, there were 415,900 files linked to 700 ‘dataset’ type submissions, taking up 308 Gb of storage space. These accompany nearly 20,000 research publications (almost exclusively in PDF format). Our analysis indicates that a similar 44% of the files stored within research datasets in Apollo also have unrecognized formats.

To contribute towards the business case for digital preservation, the CUL team also conducted two searches of archival and bibliographic databases. These searches used regular expressions to search for approximately 250 different terms related to digital and audio-visual media carriers. From these queries, we estimate that there are approximately 80,000 - 100,000 unique pieces of digital and AV media in our collections of published materials, and circa 2500 records in our union archival catalogue (Janus [14]) that refer to digital or AV materials. Analysis of the results indicated that the (comparatively) simple approach of using regular expressions achieved a precision measure of 0.8, which was (just) performant enough to base effective management statistics upon.

As with our colleagues at the Bodleian, however, a major area of focus was upon our digitized image collections. Thus far, we have scanned our central storage system with a custom-written Perl script (faster and lighter-weight than DROID, but designed to bring older content under the same degree of control as the material that for our repository environments, two copies will be stored on spinning disk ideally on different technology stacks in more geographically diverse sites, along with further copies of these stored on tape [11]. Due to the extensive infrastructure already in place at Oxford, we have not yet considered 3rd party cloud storage; however this is one of a number of options we are still reviewing.

We have also worked directly with our photography and digital library content and system development teams to analyse their digitization and content development workflows. We are now preparing some more in-depth tasks related to digitized materials, which are:

(1) To assist with an infrastructure upgrade to provide photographers with local redundant / robust, but fast ’scratch’ storage, and to work with them to develop storage management plans and policies. We may also trial Versioning Object Storage (e.g. Swift [16]).

(2) To follow the Bodleian’s approach, by conducting more in-depth scans, using characterisation tools such as JHOVE, to find out more about our legacy objects and prioritise cleaning them up and bringing them under control.

(3) To help the photographic, content creation and development teams improve their workflows. This is partially about tools - so we will experiment with Goobi and Apache ActiveMq / Camel [1] - but it is also, fundamentally, about working with staff to help them improve their processes.

Our survey work has also highlighted a similar need to consider the underlying storage architecture as that encountered by the Bodleian: all our data is currently stored and backed up within Cambridge, with no mirrored versions or backups held further afield.

Finally, all the above has been conducted alongside a review of preservation systems and vendor companies. Part of this has been to join the JISC Research Data Management Shared Services Pilot [15], where we are one of three UK universities (alongside St. Andrews and Lancaster) who are testing both Preservica [17] and Archivematica [9]. We also plan to test three other systems (Fedora 4, Rosetta [19] and Roda [18]) during a programme of case studies lasting until spring 2018.

3 ACKNOWLEDGMENTS

We would like to thank our colleagues Edith Halvarsson, Sarah Mason, Somaya Langley and Lee Pretlove for their help and advice, and the Polonsky Foundation for funding research of digital preservation at Oxford and Cambridge (DPOC) 2016-2018.

REFERENCES

Implementing Digital Preservation at the University of Melbourne

Jaye Weatherburn
University of Melbourne
Parkville, Melbourne, Australia
jaye.weatherburn@unimelb.edu.au

ABSTRACT
To implement the University of Melbourne’s digital preservation strategy, during 2016-2017 a small project team (1.6 staff) was tasked with achieving goals within four key areas: Culture, Infrastructure, Policy, and Organisation. These goals have involved planning for digital curation education improvements, assessing relevant university policies and procedures, and planning for a digital preservation ecosystem encompassing many different key stakeholders. The goals and achievements to date for the project are described, as are some next steps that are building on these activities.

KEYWORDS
Preservation strategy, Digital preservation project

1 THE STRATEGY AND ROADMAPS
The University of Melbourne’s ten-year digital preservation strategy [1] takes into account interrelated groups of digital materials, including research outputs (e.g. higher degree theses), research data, university records (born digital or digitised records forming organisational memory and legacy), and cultural collections (as defined by the university’s Collections Policy [2]). Four key principles underpin the strategy: Culture, Infrastructure, Policy, and Organisation. To implement the strategy, roadmaps were developed for research outputs and research data [3]. These areas were initially prioritised due to:

- Funding mandates for research data preservation
- The pressing need to implement digital preservation processes for the move to digital-only submissions of higher degree research theses
- The lack of project resources to tackle the research outputs, research data, university records, and cultural collections areas simultaneously.

2 GOALS FOR 2016-2017
Using the implementation roadmaps to guide the project work, in 2016-2017 the project team was tasked with high-level goals for each of the four key principles: Culture: Develop an engagement plan and education framework to increase digital preservation awareness; Infrastructure: Develop functional and non-functional requirements for implementing a preservation platform; Policy: Review and align university policies related to the management and preservation of research data and records; Organisation: Review digital processes and workflows for management of research product [4]. These efforts are ongoing, and some successful outcomes can already be noted.

2.1 Culture and Communication
A training framework was developed to guide improvements for four areas: Fundamentals of digital preservation, Archiving and data management, Technical requirements, and Business requirements [5]. The framework provides high level, overarching principles, and next steps are identifying different requirements of staff and researchers in order to implement these. A constant challenge throughout the project has been the confusion around terminology, particularly “digital preservation” and “archiving”. It has been useful to attempt to define these terms clearly, then to revisit these definitions with different stakeholders, noting the differences, and using these to generate conversation in order to begin building a shared understanding. This is a complex and ever-evolving task but a worthwhile investment, because it provides a sense of clarity for stakeholders where previously there was confusion, and has enabled a sense of connection and shared purpose. The digital preservation project has acted as a connector for some of the cultural dissonances that have arisen over time at the university, particularly around understanding and managing research data. Joining the Digital Preservation Coalition (DPC) in June 2017 has contributed to the university’s development of a research community of practice, through using the online resources offered by the DPC for various initiatives and events. Being a DPC member has allowed the university to learn about international developments and case studies happening in the field much more readily.

2.2 Infrastructure Blueprint Pilot Projects
To begin assessing functional and non-functional requirements for implementing a preservation platform, in 2017 the project team scoped and implemented four collaborative blueprint pilot projects. Selection of these pilots was based on readiness of the
digital materials to be investigated for long-term preservation, and the human resources available in each area to aid the project team. These pilots were led by different teams at the university to highlight the importance of collaborative workflows across university departments, necessary if long-term preservation is to succeed. By involving infrastructure service providers, archivists, and faculty support staff in the requirements gathering process, the people needed to play key roles in creating and maintaining a preservation ecosystem were engaged early. The pilots were designed primarily to identify and scope requirements for preservation storage, but also assessed the effectiveness of current workflows, staffing requirements, and costs associated with long-term preservation.

2.2.1 Digital Archives Pilot. This pilot, led by the University of Melbourne Archives, trialled Archivematica for digitised and born digital cultural collections, and a selection of university council minutes. It aimed to highlight infrastructure and skill gaps, and evaluate what is required to move from a pre-digital archiving paradigm to one that supports managing digital materials.

2.2.2 Architecture Building Archive Pilot. The Architecture faculty led this pilot to determine infrastructure requirements to support curation and preservation of research data for individual researchers and large research groups.

2.2.3 Digital Theses Preservation Pilot. In early 2017, digital-only deposit of theses became mandatory at the university. The institutional repository team assessed the suitability of Archivematica for long-term preservation processes for digital theses.

2.2.4 Active Research Data Preservation Pilot. The university’s research infrastructure team assessed current storage, tools, and systems available for managing research data, and began to investigate preservation storage requirements.

2.3 Policy Review

A review was commissioned to identify the effectiveness of current university policy to support long-term preservation [6]. Eight relevant university policies were identified and analysed. A key general recommendation was that adopting a sustainable funding stream to meet obligations for the long-term preservation of digital product is essential. This may seem obvious, but given the current lack of awareness of the need for and requirements of digital preservation at the university, it is important that this point is reiterated and promoted to relevant decision makers as a key policy issue. Currently, the onus lies with research units or departments to establish procedures for data retention, and to maintain a register of research data and their location and storage requirements. The policy review questions whether this siloed approach is suitable for long-term digital preservation, or whether developing a university-wide register of research data and records and storing them in a central repository would be preferable. The digital preservation project team remains committed to actioning the recommendations, but with 1.6 staff we have been stretched thinly across other project deliverables. For now, we maximise opportunities to lobby senior decision makers at the university, to promote clear understanding of the shared goals required for long-term digital preservation, and to promote the pressing need for policy direction.

2.4 Organisation Reports

One major activity for the Organisation principle was documenting the main barriers to implementing digital preservation at the university, specifically targeting the library, archives, and research IT support areas [7]. The resulting report summarises the findings from forty face-to-face interviews in these areas. Major findings included that holistic data curation processes are lacking, and that digital preservation is not widely understood as an important part of the research lifecycle. A siloed way of working was identified as a major organisational barrier to developing shared messaging and values. Also identified was the need for a community of practice, to enable key stakeholders to meet regularly to avoid duplication of effort and consolidate messaging to the research community. This report presented many issues and recommendations for action. Awareness has been raised that without organisational cohesion to improve ad hoc, unconnected services, confusion and lack of clearly defined responsibilities will negatively impact the goals of the digital preservation project. The second major activity was documenting fifteen case studies of current research data management practices in different disciplines [8]. These provided insights that the project team is using for building the ongoing business case for the digital preservation project, as they demonstrate that little to no long-term thinking is put in place at the start of projects, thus barring preservation from becoming an important part of the research lifecycle.

3 CONCLUSION

Implementing a digital preservation strategy has been a challenging undertaking with few dedicated project resources, yet our work to date has identified key findings that help build new opportunities. In-kind support provided by the wider research community at the university will continue to be essential, particularly for assessing gaps and gathering requirements for the implementation of a preservation ecosystem at the university.

REFERENCES

Promotion of InfoEver SIG (Special Interest Group), the Research Group Seeking to Establish a Grand Design for Long-Term Preservation of Digital Information in the Super Smart Society

Extended Abstract†

Hideo Nojima
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara
Japan
nojima@rsc.naist.jp

Shoji Kasahara
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara
Japan
kasahara@is.naist.jp

Kotaro Minato
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara
Japan
kotaro@is.naist.jp

ABSTRACT

Advancements in information science technology are generating anticipations of a “super smart society” that integrates the physical and cyber worlds. A research group that discusses the grand design for a long-term digital information preservation system that can respond to the super smart society from the perspectives of information science, materials science and social science is being promoted. The group started its activities on March 2016, and has completed the first phase on March 2017 by conducting a basic survey. For 2017, the research group will continue on to the second phase by specifically deciding on the targets for digital preservation and their methods, and then propose strategies on the long-term preservation of digital information in the super smart society in 2018.

KEYWORDS
digital information, long-term preservation, super smart society, blockchain, AI, IoT, Big Data

1 INTRODUCTION

Precious information on mankind is currently in danger of extinction in many fields of society, and various measures are being taken to develop a long-term preservation and recording system for digital information [1]. However, much of the actions focus on the research and development of media and devices [2] that preserve recorded data for extremely long times, and research on the system from the standpoints of information science and technology are seldom being conducted.

Meanwhile, a super smart society with the physical and cyber worlds integrated is being anticipated following the rapid advancements in the information science field such as AI, IoT and Big Data [3]. Moreover, blockchain that the virtual currency Bitcoin is based on is drawing attention for its innovations and significant potentials in application. A wide array of blockchain expansion is currently being proposed, and centralized databases to record information will consequently become unnecessary if transaction data are distributed and shared.

Considering such background, we have formed a research group that will develop the grand design for preserving digital information in the super smart society, with the belief that a long-term preservation system compatible with the society with physical and cyber worlds integrated will be essential in archiving digital information. Named the InfoEver SIG (Special Interest Group) Meeting, the research group started its activities on March 2016.

2 RESEARCH GROUP GOALS, PLANS FOR PROMOTION AND PROMOTION METHODS

2.1 Goals

To develop the grand design for long-term preservation of digital information that is compatible with the super smart society in the future, where physical and cyber worlds will be integrated.

2.2 Plans

- Phase 1 (March 2016 – March 2017) Basic survey
- Phase 2 (April 2017 – March 2018) Promote the activities of the research group by deciding on the data to archive and their preservation methods.
- Strategic proposal for long-term preservation of digital information that corresponds to the super smart society (April 2018)

2.3 Promotion Methods

The members elected from corporations, the National Diet Library and universities for the research group will gather bi-monthly for meetings. The meetings are open to the public and welcome nonmember participants, and with planned themes, invite external specialists involved in the latest research and business for lectures. Based on the lectures, the research group members hold discussions on the long-term preservation of digital information from various standpoints, and review options towards the super smart society.
3 RESULTS

The research group started activities from March 2016, and completed the first phase on March 2017 by conducting a basic survey. The results of Phase 1 activities (March 2016 - March 2017) are as follows:

3.1 Kickoff Workshop

March 7, 2016: Keihanna Open Innovation Center (“KICK”)
Forewords: Makoto NAGAO
(Director, International Institute for Advanced Studies)
Keynote speeches:
1) “Current Status and Future Challenges of Ultra Long-Term Storage Device/Systems R&D”
Seiichiro KAWAMURA (Fellow of Center for Research and Development Strategy, Japan Science and Technology Agency)
2) “Blockchain Technology of Bitcoin and its Applicability to Long-Term Preservation of Digital Information”
Shoji KASAHARA (Professor of Graduate School of Information Science, Nara Institute of Science and Technology)
InfoEver SIG (Special Interest Group) Meeting was established starting from this kickoff workshop.

3.2 InfoEver SIG (Special Interest Group) Meeting Activities

1st meeting: May 13, 2016 (International Institute for Advanced Studies)
Invited lectures:
- “The Millennium EHR Project toward Japanese National Medical Data Repository”
  Tomohiro KURODA (Professor, Kyoto University Hospital)
- “Preservation of Cultural Property Photography in the Digital Era: In-Between Bequeathing and Inheriting”
  Ichiro NAKAMURA (Specialist, Nara National Research Institute for Cultural Properties)
Special lecture:
- “The Importance of Information Recording and Discussions on the Latest Super-High Resolution Digital Records”
  Kenji SUMIYA (Senior URA, Kansai University (former CTO of Hitachi Maxell))
2nd meeting: June 20, 2016 (International Institute for Advanced Studies)
Invited lectures:
- “Long-Term Preservation of Digital Information”
  Sadayasu ONO (Vice President, O&G Corporation)
- “Issues in Digital Archiving: Continual Operation of Long-Term Recording Systems”
  Hiroyuki KAWANO (Professor at the Faculty of Science and Engineering, Nanzan University)
3rd meeting: October 7, 2016 (International Institute for Advanced Studies)
Invited lectures:
- “New Hardware Technologies for Cognitive Computing”

Based on these results, Phase 2 (April 2017 – March 2018) activities will decide on the data to archive and their preservation methods that are compatible to the super smart society.

4 CONCLUSIONS

A research group to discuss ways to develop the grand design of the long-term digital information preservation system compatible with the super smart society that integrates the physical and cyber worlds from the information, physical and social science perspectives was established. The research group started activities with members from corporations, the National Diet Library, and universities, and completed the basic survey in the agenda for Phase 1 on March 2017 after holding six meetings. For 2017, the group will continue on to Phase 2 to decide on the specific data to archive and their preservation methods, and then propose strategies on the long-term preservation of digital information in the super smart society in 2018.

ACKNOWLEDGMENTS

The research group has been funded by the Program for Promoting the Enhancement of Research Universities of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

Performing a Content Analysis of Biodiversity Literature

Alicia Esquivel
National Digital Stewardship Resident
Lenhardt Library
Chicago Botanic Garden
esquivelndsr@gmail.com

ABSTRACT
Biodiversity essentially describes all living organisms and their environments. As the world’s most comprehensive digital library collection of legacy biodiversity literature, the Biodiversity Heritage Library (BHL) is dedicated to building and maintaining their repository by determining the scope of biodiversity literature and analyzing the collection strengths, weaknesses and gaps in the BHL corpus [1]. Regular analyses have been performed since the creation of BHL in 2006 and as the corpus expands and methods for analyzing literature in the public domain develop, continuous analysis of BHL coverage and gaps remains critical. In the course of 2017, BHL has the opportunity to work with a team of National Digital Stewardship Residents who are committed to developing a plan for the next generation of BHL.

This poster will illustrate the current research and progress of the collection analysis project of the National Digital Stewardship Resident (NDSR) project led by the resident hosted at the Chicago Botanic Garden’s Lenhardt Library.

KEYWORDS
National Digital Stewardship Residency, NDSR, Biodiversity Heritage Library, BHL, biodiversity, collection analysis, content analysis, digital library

1 INTRODUCTION
1.1 Biodiversity Heritage Library
The Biodiversity Heritage Library (BHL) is an open access digital library for biodiversity literature. Member and affiliate partners of BHL form a consortium of natural history and botanical libraries that work together to digitize legacy literature on biodiversity held in their collections to be used as a global “biodiversity commons.” As of March 2017 over 51,000,000 pages have been digitized and made available since digitization began in 2006. In addition to public domain books and journals, BHL works to obtain permission from publishers to digitize biodiversity materials still under copyright. BHL expanded globally in 2009 and now has nodes in Europe, China, Australia, Brazil, Egypt, Africa, and Singapore [2]. In June 2016 BHL was awarded a grant funded by the Institute of Museum and Library Services (IMLS) to work with 5 residents hosted at BHL member and affiliate institutions across the United States. The “Foundations to Actions: Extending Innovation in Digital Libraries in Partnership with NDSR Learners” (Foundations to Actions) hosts residents for 12 months over the course of 2017.

1.2 National Digital Stewardship Residency
The National Digital Stewardship Residency (NDSR) program aims to develop a community of professionals in the fields of digital stewardship and informatics through collaborative field experience. Since the 2013 pilot program, NDSR has provided opportunities for recent graduates of library and information science to link theoretical knowledge to practice in a professional context through hands-on experiences [3]. The 5 residents working on the “Foundations to Actions” cohort are currently hosted at the Missouri Botanical Garden, Natural History Museum Los Angeles County, the Ernst Mayr Library of the Harvard University Museum of Comparative Zoology, Smithsonian Libraries and Chicago Botanic Garden.

1.3 Lenhardt Library
The Lenhardt Library is one of the great treasures of the Chicago Botanic Garden. Open to the public 7-days a week, its 125,000 volume collection encompasses resources on gardening, botany, plant conservation and landscape design, in formats from rare books to e-books. The Library supports all aspects of the Garden’s mission: “We cultivate the power of plants to sustain and enrich life” [4]. The resident hosted at Lenhardt Library (Alicia Esquivel) is dedicated to performing content and collection analyses for BHL.
2 OBJECTIVES

In order to manage the BHL’s collection scope, the Collections Committee at BHL created a Collection Development Policy that defines specific areas of interest to digitize based on BHL user needs. The committee defines BHL users as an interdisciplinary audience composed of “zoologists, botanists, evolutionary biologists, taxonomists, systematists, ecologists, natural history collections managers, scientific illustrators, biological science historiographers, and amateur scientists & hobbyists” [5]. Given these user groups, BHL focuses digitization efforts on the core literature of zoological and botanical literature especially those with high concentration of taxonomic names. While BHL can easily calculate the amount of pages and items digitized, it becomes more challenging to calculate the entire scope of biodiversity literature in order to assess which areas are missing in the BHL corpus.

Collections Committee members estimated the core biodiversity literature in 2010 to consist of 495,000,000 pages. This measurement was calculated by estimating the amount of botanical literature in the public domain based on two extensive bibliographies (Taxonomic Literature: A selective guide to botanical publications and collections with dates, commentaries and types and B-P-H: Botanico-Periodicum-Huntianum). This calculation was then compared to the total amount of botanical species. This ratio of pages to species was then compared to zoological and mycological species counts [6].

This current project explores methods to improve the accuracy of these estimates in order to determine the scope of biodiversity literature both in and out of copyright. Additionally, an aim of this work is to develop visualization methods for examining the coverage within the corpus by comparing BHL materials to known ontologies. This work-in-progress of digital collection analysis explores geographic, taxonomic, and topical tools for analysis. It is performed using BHL metadata such as subject headings and taxonomic name data, as well as exploring full text indexing for potential data mining using natural language processing and machine learning.

To visualize geographic representation in BHL, the full text is mined for geographic names, or toponyms, then assigned latitude and longitude coordinates and placed onto a map.

The taxonomic representation in BHL is done using scientific names mined from collection text using the open source Global Names and Recognition and Discovery (GNRD) architecture, a suite of machine learning and named entity recognition algorithms, to extract scientific names to index and attach page records. The list of scientific names can be exported through BHL’s public data tables. In order to determine collection coverage of different biological kingdoms (Plantae, Animalia, Fungi, etc.), the list of scientific names can be sorted by using the Global Names Resolver tool, from the GNRD suite. This tool matches names based on different data sources (taxonomic ontologies such as NCBI, ITIS, and Catalogue of Life). By filtering scientific names into kingdom level taxonomic ranks, we can get a more granular look at species coverage in BHL.

Topical analysis of the collection is being done in collaboration with external researchers. The text is analyzed using a topic model and a controlled vocabulary to extract topics from items based on the full text of the items.

The lessons learned from this research will inform the next iteration of BHL in how to provide meaningful data to its users. By exploring collection analysis methodologies of the BHL corpus, biodiversity knowledge can become better connected across open access platforms and databases.

3 ACKNOWLEDGEMENTS

Funding and support for this National Digital Stewardship Residency comes from the Institute of Museum and Library Services.

REFERENCES

Reviving past research data by utilizing institutional repositories

Eriko Amano
Kyoto University
Yoshida-honmachi
Sakyo-ku
Kyoto 606-8501
amanoo@kura.kyoto-u.ac.jp

Yasuyuki Minamiyama
National Institute of Polar Research
10-3 Midori-cho
Tachikawa-shi
Tokyo 190-8518
minamiyama@nipr.ac.jp

SUMMARY
A considerable number of databases created through Japanese public research funds are now under unstable management or use an outdated publication methodology; therefore, future access to the research results these databases contain cannot be guaranteed. Here, we examined the current status of databases listed in two major sources, Grant-in-Aid-funded databases and the National Diet Library Database Navigation Service (Dnavi), and reaffirmed this problematic situation. To ensure the sustainability of these databases, the Research Data Task Force of the Japan Consortium for Open Access Repository (JPCAR) is currently testing a workflow for the transfer of contents from unstable databases to institutional repositories.

1. INTRODUCTION
Numerous databases in various fields have been published as the result of publicly funded research in Japan. From FY2009 to FY2011, the Grant-in-Aid for Publication of Scientific Research Results: Databases program, administered by the Japan Society for the Promotion of Science, supported the creation of 130 databases, 74 of which contained research results in the natural sciences and 56 of which contained research results in the humanities and social sciences (HSS).

2. PRELIMINARY SURVEY OF DATABASE SUSTAINABILITY

2.1 GRANT-IN-AID-FUNDED HSS DATABASES
To examine database sustainability, we conducted a preliminary survey in which we evaluated the sustainability of the 56 HSS Grant-in-Aid-funded databases with respect to their maintenance status and publication methodology. As of April 2015, 47 (84%) of the HSS databases were live, and their maintenance status was as follows: stable, 24 (41%); unstable, 20 (42%); inaccessible (including CD-ROM-based, fee-based, and registration-based databases), 3 (6%). Among the databases with unstable management, we found some that appeared to have not been updated in recent years, and for one of the databases the lead researcher had moved to another institute and the current owner of the database was not specified. Regarding publication methodology, some databases were provided in the legacy FileMaker format or as a single PDF file. Such methodologies do not ensure future accessibility to the database so newer technology should be applied to these databases to prevent loss of the research results they contain.

2.2 OTHER DATABASES
The National Diet Library Database Navigation Service (Dnavi), formerly managed by the National Diet Library of Japan and currently archived as a static dataset, is a searchable list of databases that includes not only databases published by academic institutions and individual researchers but also public use databases. Approximately 25% (5333) of the databases in Dnavi are managed by universities, laboratories, and technical colleges. A preliminary
survey conducted in 2016 revealed that the majority of databases in Dnavi created after 2000 were in most cases of discontinuation, and about half of these idle databases contained HSS research results. These databases will likely become inaccessible in the near future resulting in loss of the valuable data they contain. Therefore, to secure access to these databases and prevent data loss, the data should be migrated to a stable infrastructure such as an institutional repository.

3. CHALLENGES

3.1 DEVELOPING A WORKFLOW

To migrate unstable databases to institutional repositories, a workflow that is applicable to most databases and streamlines the process of converting and transferring the database metadata to the institutional repository is essential. In addition, a new metadata schema, the JPCOAR metadata schema, built specifically for the Japanese situation was developed recently by the Metadata Task Force of JPCOAR to replace the current standard schema for institutional repositories, juini2. The JPCOAR metadata schema assumes preservation of non-literature materials that were created in accordance to international standards.

The JPCOAR metadata schema was developed to ensure that the metadata contained in Japanese institutional repositories is compatible with the Open Science framework and is interoperable internationally. In the development of the JPCOAR metadata schema, the Metadata Task Force

- focused on interoperability with six established international metadata standards (i.e., RIOXX, BIBO, COAR vocabulary, OpenAIRE, DataCite, and Dublin Core),
- adopted major identifiers (Crossref Funder ID, ISNI, ORCID, etc.),
- monitored compliance with open access policies,
- hierarchized priority elements (creator, fundingReference, etc.), and
- enhanced elements related to research data.

In FY2016, the Research Data Task Force and Metadata Task Force carried out a trial migration of a Grant-in-Aid-funded database to an institutional repository; after subsequent review of the trial, they proposed the following workflow:

1. Select a database in Dnavi that is under unstable management (live or discontinued databases are considered).
2. Acquire permission from the database administrator(s) to migrate the contained data to a new database platform.
3. Examine the database metadata and select and implement the most appropriate conversion method for registration of the metadata in an institutional repository.
4. Map the database’s unique metadata schema to the JPCOAR metadata schema.

3.2 FUTURE PLAN

In cooperation with several universities we are now testing our proposed workflow. By following this workflow, we intend to map the metadata from unstable research databases to the JPCOAR metadata schema and register the databases in real-world institutional repositories.

4. REFERENCES

Grant-in-Aid for Scientific Research (KAKENHI)

National Diet Library Database Navigation Service (Dnavi)
http://doi.org/10.11501/8427554

Japan Consortium for Open Access Repository (JPCOAR)
https://jpcaro.repo.nii.ac.jp/
Abstracts

Workshops
Digital Preservation Storage Workshop: Exploring Preservation Storage Criteria and Distributed Digital Preservation

A. Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 01970 USA
andrea_goethals@harvard.edu

N. McGovern
MIT Libraries
77 Massachusetts Ave (14E-210)
Cambridge, MA 02139 USA
nancymcg@mit.edu

J. Mandelbaum
Library of Congress
101 Independence Ave
Washington, DC 20540 USA
jman@loc.gov

S. Schaefer
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093 USA
sschaefer@ucsd.edu

G. Truman
Truman Technologies
4096 Piedmont Avenue, Suite 217
Oakland, CA 94611 USA
gail@trumantechnologies.com

E. Zierau
The Royal Danish Library
P.O. Box 2149
Denmark
elzi@kb.dk

ABSTRACT
Preservation storage is an essential component of a sustainable digital preservation program. This workshop uses recent community developments to explore good practice for preservation storage as well as implementation use cases.

Preservation Storage “provides the services and functions for the storage, maintenance and retrieval of [Archival Information Packages, or] AIPs. Archival Storage functions include receiving AIPs from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the media on which Archive holdings are stored, performing routine and special error checking, providing disaster recovery capabilities, and providing AIPs to Access to fulfill orders” (see OAIS, 4-2, [2]). For this workshop, the term “Preservation Storage” is used, instead of Archival Storage or Digital Preservation Storage, but is understood to mean storage for digital material under preservation management.

This workshop references two documents that have been informed by community discussions and derive from digital preservation standards and two previous workshops: “Preservation Storage Criteria” initially presented at “What is Preservation Storage?” workshop at iPRES 2016 [5] and “Outer OAIS - Inner OAIS (OO-IO) Model” the focus of the Distributed Digital Preservation Workshop presented as part of the Digital Preservation Management Workshop series prior to iPRES 2015 [4].

KEYWORDS
Preservation Storage, Archival Storage, OAIS Reference Model, Digital Preservation

1 INTRODUCTION
Storage is a fundamental component any organizational program designed to preserve digital content. Digital preservation programs must make use of a combination of storage systems or services that leverage the requirements, opportunities, and challenges of distributed digital preservation approaches. Good practice for preservation storage must respond to the evolving capabilities and capacity of emerging technologies. Digital preservation managers must continually identify, evaluate, and incorporate available options to provide preservation storage that is suited to the requirements, scale, content, and cost models of individual programs. This workshop will demonstrate the use of the proposed storage criteria as a framework for making decisions and the use cases of the distributed digital preservation model for considering possible approaches to implement preservation storage.

2 WORKSHOP STRUCTURE
Prior to the workshop, attendees will receive current versions of the preservation storage criteria and the OO-IO Model documents along with brief supporting materials as background for the discussion and exercises. During the workshop, presenters will introduce the two main topics, storage criteria and distributed digital preservation, then facilitate interactive discussions and breakout groups explore and apply core preservation storage concepts. After the workshop, presenters will incorporate relevant workshop outcomes into revised versions of the documents discussed as a continuation the digital preservation community’s discussion of preservation storage standards-based practice.

Exercises and discussions will explore the categories of the preservation storage criteria (Content integrity, Cost considerations, Flexibility and resilience, Information security, Scalability and performance, Storage location, and Transparency) and distributed digital preservation use cases.

3 WORKSHOP OBJECTIVES
Attendees of the workshop will develop a deeper understanding of preservation storage, be exposed to current options and challenges, consider the role standards play in achieving good practice, and apply the concepts to realistic use cases. Instructors will include:
Andrea Goethals leads the development and operation of Harvard's digital preservation program and the management and oversight of the Digital Repository Service (DRS), Harvard's large-scale digital preservation repository.

Nancy McGovern is the Director for Digital Preservation at MIT Libraries. She leads the Digital Preservation Management (DPM) workshop series, offered fifty times since 2003. She has held senior positions at ICPSR; Cornell University Library; and the U.S. National Archives. She is the President of the Society of American Archivists, 2016-2017. She completed her PhD on digital preservation at UCL in 2009.

Jane Mandelbaum currently works on IT strategic planning and IT architecture development in the Office of the Chief Information Officer, Library of Congress. She leads and guides enterprise-wide projects and architecture initiatives focusing on digital content management and digital content delivery. She previously served as implementation and operations manager for the Library’s largest business system (the Library’s Integrated Library System), led the team to establish and operate the Library’s first LAN environment, and served as the automation officer for the National Library Service for the Blind and Physically Handicapped. She has also worked as a project manager and business analyst in a management consulting firm.

Sibyl Schaefer is the Chronopolis Program Manager and Digital Preservation Analyst for Research Data Curation at the University of California, San Diego. In addition to working with national digital preservation efforts like the Digital Preservation Network and the National Digital Stewardship Alliance, she helps define long-term digital preservation solutions for the UCSD campus. Schaefer holds an MLIS with a specialization in Archival Studies from the University of California, Los Angeles.

Gail Truman is the lead for Truman Technologies, providing consulting and development services relating to digital archives. Truman Technologies help clients assess and deploy digital preservation systems, from web archives and digital repositories to large-scale on-premise and cloud-based digital storage, on-line archive and business continuity infrastructures.

Eld Zierau is a digital preservation specialist at the Royal Danish Library, responsible for research, design and implementation of various aspects of digital preservation. She has been heavily involved in preservation storage implementation there as a chief architect, auditor of implementation, and creator of the initial Outer OAIS-Inner OAIS model, enhancing it through international collaboration. She completed her PhD in 2011 on digital preservation with a specific focus on preservation storage.

ACKNOWLEDGMENTS
This digital preservation storage criteria document referred to in this workshop proposal was informed by discussions at the 2016 Designing Storage Architectures for Digital Collections Conference [1] and at the PASIG 2016 conference [3]. In addition to the authors listed, Steve Knight of the National Library of New Zealand and Kate Zwaard of the Library of Congress contributed to the development of this workshop.

REFERENCES
A Dutch approach in constructing a network of nationwide facilities for digital preservation together

Joost VAN DER NAT  National Coalition Digital Preservation  
P.O. Box 17 097, 1001 JB Amsterdam  
The Netherlands  
+316 5204 8788  
jvdnat@ncdd.nl

Marcel RAS  National Coalition Digital Preservation  
P.O. Box 90 407, 2509 LK Den Haag  
The Netherlands  
+316 1477 7671  
marcel.ras@kb.nl

ABSTRACT

In November 2016 this research was rewarded the Digital Preservation Coalition Award 2016 in the category Research and Innovation [3]. According to the jury the research report presents an outstanding model to help archives and memory institutes to share facilities and create a distributed, nationwide infrastructure network for Digital Preservation. In this workshop participants will be able to gain a deep understanding of this collaboration model and will be provided exercises to work with and apply it to their own archiving facilities in order to explore the possibilities of a similar shared network for digital preservation in their own environment.

The goals of the workshop are: (1) to introduce the participants to the model, gain understanding of the way to define supply and demand for DP-services, Then (2) to apply these elements to their own situation by rating and ranking the grid with essential DP Building Blocks; and (3) to explore the appropriate means to acquire the Services needed within a nationwide collaborative network. Throughout these 3 steps, practical examples are given of distributed use of infrastructure in the Netherlands.

KEYWORDS


1 INTRODUCTION

1.1 Organisational setting

As part of the national strategy for Cultural Heritage in the Netherlands, under the umbrella of the Network Digital Cultural Heritage [1] (Dutch acronym: NDE), a model [2] for a network of distributed facilities to ensure long-term accessibility of digital cultural heritage was developed. This model is part of a nationwide strategy of development of collaborative Services based on three pillars: to make the digital heritage wider visible, better usable and more sustainable preserved. The network was established on the initiative of the Ministry of Education, Culture and Science and consists of a number of large organizations occupying key positions in the field of digital heritage [4].

Work packages are established for each of the pillars, outlining the projects necessary to achieve its central goals. The development of the network of facilities is part of the Sustainability work package, which is carried out by the Dutch National Coalition for Digital Preservation (NCDD) [5] between 2015 and 2017. The first concepts of the collaborative model were actually developed by the NCDD as part of its strategy to set up distributed national infrastructure based on a collaborative approach. The NCDD acts as the national platform for exchange of knowledge and expertise and has a role in coordinating and facilitating the establishment of a national network in which long term access to digital information =which is of crucial importance for science, culture and society= is guaranteed.

Within the framework of the Digital Heritage Network a catalogue of existing preservation Services to be used in a collaborative infrastructure, has been developed. Alongside, a range of case studies will bring the model to a practical level. Having a theoretical model is very helpful, but bringing it to an operational level is even better.

1.2 Very brief description of results of the research

In this workshop we will focus on practical implementations of the model for a network of distributed facilities. How can we move from theory into practice?

The workshop will help the participants to gain a deep understanding of the underlying concepts of the model. What are the steps we took in the development of the model?

Two basic concepts of the model are depicted in the two following diagrams.
Based on this high-level model and other sources [6] we developed a “Business – IT stack” for digital preservation, with Building Blocks representing all elements we think are the parts of the digital preservation environment.

In figure 2 all these elements are depicted as Building Blocks. From IT-elements as storage facilities, to standards, training, R&D, and semantics. Any organisation, or collaborative effort, obligated to ensure long-term access to digital information should have (most of) these elements organised. The model then focuses on those elements which could be shared within a network or a collaborative effort. The dark blue Building Blocks are the ones that are potentially shareable. The grey ones seem by definition organisation specific. Having the shareable elements defined, scenarios were developed to set out the lines for collaboration. In this, collaboration is always based on two sides of the coin: the Supply and the Demand for Services.

2 TOPICS

In the workshop we will connect Supply and Demand for Services for digital preservation:

1. The participants will be introduced to the underlying concepts of “distributed facilities for digital preservation”;
2. In a first exercise the participants will try to articulate the demand of sample organisations. A hand out is given;
3. In the second session the participants will identify suitable Building Blocks on the Supply side for the Demand they have articulated in their first session;
4. Two [7] or three practical examples will be given of cooperation in the Dutch context;
5. In the last session the same work groups are now asked to select the proper Means to achieve their goals for cooperation. A hand out is given with a provisional list of possible Means.

If and when email addresses of participants are available, we intend to send the participants a concise reader with background material as a preparation to the workshop [8].

Total time of the workshop, including breaks, is 3:00 hours.

3 INTENDED AUDIENCE

Anyone interested in learning more about creating a network of distributed facilities for digital preservation. This tutorial has a strong practitioner focus and will be especially interesting for those working with Digital Archives and Digital Collections. Custodians of these collections who are faced with a multitude of questions and developments will gain a better understanding of the subject.

4 EXPECTED LEARNING OUTCOMES

After the workshop, participants will have knowledge / understanding of:

- the notions of “Infrastructure” and “Services” for Digital Preservation;
- existing facilities in the Netherlands;
- how to articulate Demand for a Facility;
- Building Blocks, determine the most important ones, identify “quick wins”;
- what Means are available; what type of Means are suitable for various Services.

5 PRESENTERS

Marcel Ras is the Program Manager of the Dutch National Coalition for Digital Preservation (NCDD).

Joost van der Nat is the principal researcher.

6 REFERENCES


[4] The participants in this network are national organizations with large digital collections and a mandate to preserve them, like the National Library (Koninklijke Bibliotheek or KB), the Institute of Sound and Vision (BenG), the Royal Netherlands Academy of Arts and Sciences (KNAW), the National Archives (NA), the Cultural Heritage Agency of the Netherlands (RCE) together with other partners like for example the knowledge centre Digitaal Erfgoed Nederland (DEN) and the National Coalition for Digital Preservation (NCDD).


[7] The case of video-recordings of the Dutch Parliament is one, the need for “storage” of smaller Frisian museums is another.

[8] The reader will contain (1) a Summary of the Research, (2) a graphic of the model of the Building Blocks with a short description of each Block, (3) the program of the workshop, and (4) the intended hand outs
Preserving and Providing Access to Privacy-sensitive Collections

Unmil P. Karadkar, Pat Galloway, King Davis
School of Information, The University of Texas at Austin
1616 Guadalupe St. Ste 5.202, Austin, TX 78701-1213
USA
{unmil, Galloway}@ischool.utexas.edu, king.davis@austin.utexas.edu

KEYWORDS
Digital preservation, privacy-sensitive records, historical records

1. OVERVIEW
Recent developments in digitization and dissemination technologies present the possibility of making archival collections broadly available to a global audience. In addition, new collections of born-digital documents will be readily available to support a multitude of objectives of their diverse, worldwide stakeholders. Demographics such as family members, journalists, social services providers, and policy makers can all benefit from access to these historical collections.

While these advances are broadly welcomed in most circumstances, some archival collections include restricted or privacy-sensitive collections. Examples of such privacy-sensitive records include mental health institutional records, prison records, records of the Truth and reconciliation commissions, Nazi archives, and the Guatemalan national police archives. While access to paper documents is protected by distance, physical barriers, and varying state and national policies and laws, digital access may exacerbate threats to the privacy of individuals named in these records. The online availability of such records has a potential to stigmatize or embarrass the families or descendents of those named in the records when they bear no responsibility for the acts or health conditions of the named individuals, raising ethical issues in providing open, direct access to these records. In some cases, the legal frameworks for digital records are substantially less clear than those for physical records.

2. TOPICS
This workshop will invite broad participation from scholars and practitioners who work with or are interested in issues surrounding the preservation of and providing access to digital, privacy-sensitive collections. A non-exhaustive list of topics of interest include:

- Digitization, curation, and preservation of privacy-sensitive collections
- Theoretical and metadata models
- Policies, workflows, and protections for accessing materials
- Issues in using cloud services for privacy-sensitive materials storage and scholarship
- Scholarly information behavior and needs
- Models that recognize diverse user needs (for example, aggregate data, individual information)
- Institutional and political negotiations surrounding access to privacy-sensitive collections
- Mechanisms and models for data retrieval from handwritten documents
- Privacy-aware digital repository architectures
- Privacy-aware crowdsourcing and transcription methods
- Privacy issues in designing user interfaces and data visualizations
- Privacy mitigation in data analytics and presentation
- Evaluation of existing software, infrastructure, and techniques
- Social justice issues and non-scholarly outcomes of work with restricted collections

3. TARGET AUDIENCE
While our first workshop has been accepted for hosting at the Digital Humanities 2017 conference, we have not conducted it yet. We have heard anecdotal expressions of interest from several communities but are unsure how broad and deep such interest goes. In this light, we have tried to make the topic as broad as possible, while retaining the core characteristics and constraints that we have found in our work. We anticipate receiving 10 to 20 submissions and accepting 4 to 6 for presentation at the workshop. We are unable to gauge the level of interest in the iPres community in attending the workshop without presenting. With connections in the Asian iSchools and the digital libraries communities.

4. SPECIAL REQUIREMENTS
The workshop will not require special technical support. A projector for presentations and an internet connection for participants to showcase web-based collections will suffice.
5. DURATION AND FORMAT
This workshop will be held over half a day. This time frame will allow for adequate exploration of the various aspects of the workshop topics as well as domains via presentations, as well as in-depth discussion. The workshop will be held in a seminar style, with several short and long presentations. Individuals may participate in the workshop without presenting. The organizers will include an open discussion time to engage the audience and, especially, to tease out aspects of scholarship with privacy-sensitive digital collections that the presentations do not cover.

6. ORGANIZERS

Unmil P. Karadkar is an assistant professor in the School of Information at The University of Texas at Austin. He situates his work at the intersection of digital libraries, human-computer interaction, and visualization. He studies data practices of researchers with an eye toward identifying unmet information needs. Based on an understanding of these needs, he designs software to support their evolving practices and evaluates the impact of this software on their work. His research contributes to areas such as the design of digital collection interfaces and digital humanities. His research has been funded by the National Science Foundation, Texas General Land Office, USA, and most recently, the Andrew W. Mellon Foundation.

Pat Galloway is a professor in the School of Information at The University of Texas at Austin. Her research interests include institutionalization of digital repositories and appropriate appraisal practices for digital records. In keeping with her interests as a historian, she is also interested in understanding how archiving and cultural preservation in general fit into their historical and cultural contexts. Pat has developed a suite of courses designed to prepare students to become what has recently been referred to as “Archival Engineers,” capable of managing and maintaining digital cultural objects indefinitely; she also teaches archival appraisal and a course on historical museums. She worked as a medieval archaeologist in Europe in the 1970s and then became involved with humanities-oriented computing, which she supported in the Computer Unit of Westfield College of the University of London, where her primary interest was text analysis. From 1979 to 2000 she worked at the Mississippi Department of Archives and History (MDAH), where she was a documentary editor, archaeological editor, historian, museum exhibit developer, and electronic records program director, while at the same time creating the MDAH’s automation program from scratch as manager of information systems. She is the author of Choctaw Genesis 1500-1700 (1995) and Practicing Ethnohistory(2006). From 1997 to 2000 she directed the NHPRC grant-funded project at MDAH to create an electronic records program for the state of Mississippi.

King Davis is a research professor in the School of Information at The University of Texas at Austin and has made outstanding contributions in the field of Health and Mental Health over the last three decades. Dr. Davis held the Robert Lee Sutherland Chair in Mental Health and Social Policy at the University of Texas at Austin, School of Social Work. From 2003-

2008, Dr. King also served as the Executive Director of the Hogg Foundation, which awards grants and manages programs to improve mental health research and services in Texas. Prior to his work in Texas, Dr. Davis served as the Commissioner of the Department of Mental Health, Mental Retardation and Substance Abuse Services for the Commonwealth of Virginia by Virginia Governor L. Douglas Wilder. He also has served as the John Galt Chair in Public Mental Health at the University of Virginia’s Department of Psychiatry. Dr. Davis has held academic appointments at Washington University in St. Louis, Virginia Commonwealth University, Eastern Virginia Medical School and Norfolk State University. Dr. Davis received his PhD from Brandeis University, Florence Heller School for Social Policy and Management, and his MSW from California State University Fresno, School of Social Work. He has written and published numerous articles and reports on mental health, fund raising, managed health care and social justice. His book, The Color of Social Policy, was published in 2004.

ACKNOWLEDGMENTS
This workshop is supported in part by the Andrew W. Mellon Foundation via grant number: 11500653 under the Scholarly Communications program.
Innovative approach for project viability: from a diversity of business models to harmonized and scalable national services

Pierre-Yves Burgi, Eliane Blumer, & André Jelicic
Université de Genève
DISTIC/DIS
Rue du Général Dufour 24,
CH-1204 Genève
+41 22 379 71 45
eliane.blumer@unige.ch
andre.Jelicic@unige.ch
pierre-yves.burgi@unige.ch

Aude Dieudé
Ecole Polytechnique Fédérale de Lausanne (EPFL)
VPAA SISB
RLC D1 220
Station 20, CH-1016 Lausanne
+41 21 69 34784
aude.dieude@epfl.ch

Ana Sesartić
ETH Zürich
ETH-Bibliothek, Digital Curation
Rämistrasse 101, CH-8092 Zürich
+41 44 632 73 76
ana.sesartic@library.ethz.ch

ABSTRACT

Within the context of a national-level project on Data Life-Cycle Management (DLCM), we aim to present in this workshop the methodology our team applied for developing service business models for serving the researcher community. Based on this experience, working groups will consider concrete cases of long-term data management, which then will be shared among participants to discuss the challenges and opportunities brought by the proposed methodology.

Keywords

Data life-cycle management; research data management; service business modeling; Swiss Higher Education Institutions

1. DLCM PROJECT

The DLCM project [1] is a collaborative initiative regrouping eight Swiss Higher Education Institutions (HEI)¹. The primary objective is to provide sustainable services to implement research data lifecycle management in Switzerland. Those services target the efficient management of active research data, and aim at ensuring publication, long-term reference and preservation of data subsets selected by researchers. The project is financially supported by each participating institution and by the Swissuniversities’ program P2/P5 (until July 2018, and 2020 if an extension is afforded). At the term of the project, services should however become self-financed. That’s why there are strong efforts put on viability methodology based on strategic management and lean startup method for improving the chance of having sustainable services on the long term, which is essential for services concerned with long-term preservation of research data.

2. VIABILITY METHODOLOGY

To assess the viability of the future services, we applied a customer-centric approach based on Lean Startup principles by Eric Ries [4], in combination with industry reference tools, namely the Business Model Canvas (BMC) and the Value Proposition Canvases (VPC) created by Osterwalder and Pigneur [2,3]. The VPC aims mainly to explain how to create value for the customer, while the BMC explains how to create value for the legal entity delivering the service. The Value Proposition Canvas consists of two parts: (i) the Customer Segment, that lists jobs (what the customer tries to achieve) as well as pains (risk/barriers preventing a proper job execution) and gains (benefits from achieving/exceeding customer’s objectives). (ii) the Value Map, that defines a suitable product/service relieving the customer from some of the pains that have been identified and that creates gains. Once defined, the Customer Segment and the Value Map are integrated into the Business Model Canvas and forms 2 out of its 9 building blocks. Here are the 7 remaining ones:

- Customer Relationships (i.e. self-service or personal assistance and the strategy to get, keep and grow a customer base)
- Distribution channels (i.e. downloads, software integration, etc.)
- Revenue Streams (i.e. subscription fees, licensing, etc.)
- Key Resources (i.e. human, physical and financial)
- Key Activities (i.e. to operate the services)
- Key Partnerships (e.g., joint ventures)
- Cost Structure (i.e., cost models to run the services)

In the context of the DLCM project, starting with a customer-centered approach based on the Value Proposition Canvases, several services have been identified in relation to the perceived needs and their match with the different DLCM steps, namely: Training and Consultancy, Electronic Laboratory Notebook (ELN) as well as a Long-Term Preservation Services. For each of these services, a corresponding business model has been elaborated and tested through a validation process. As a final consolidation phase, the resulting business models are assembled to yield a global service covering the whole data lifecycle.

3. CHALLENGES & OPPORTUNITIES

We initiated the business-modeling methodology in the fall of 2016 and intend to get the overall service consolidated by Autumn 2017. To reach this goal, several challenges of different natures must be faced: the geographical distribution of the service providers complicates the consolidation of the business models; having to work at a national level across different regions, each with their own languages and culture

¹ There are two Federal Institutes of Technology, four Universities, one University of Applied Science, and SWITCH, a foundation providing ICT services to HEIs
(federal context); adopting the same tool, templates and vocabulary to share the progress of individual business models and to validate important milestones; finding the right balance between economic and research practice interests.

However, once all partners adopt a common methodology, a more coordinated view of what is happening regarding data lifecycle management in different institutions can emerge, with a direct positive impact on the target audiences' needs.

4. AIM OF THE WORKSHOP

With this background experience gained at a national level from the DLCM project, the workshop aims at: (1) presenting the application and outcomes of this viability methodology from 2015-2020 with a focus on the outcomes of the long-term preservation service; (2) inviting participants to work in small groups to apply the methodology (with the set of tools) on concrete cases of their choices regarding long-term preservation; (3) opening a general discussion so that participants can express their insights and opinions on the methodology based on their use cases.

5. CONTENT OF WORKSHOP

The workshop is organized in two parts of 90 minutes. The first one and a half hour will start after an introduction round with a more detailed overview of the viability methodology as well as an in-depth introduction into the DLCM project. Then, participants will have time to take a closer look into the already existing Business Models elaborated within the DLCM project, with a focus on consultancy, ELN/LIMS as well as archiving services. Having taken notice of those existing services, participants will have the opportunity to share their opinion and discuss about the service description.

The second one and a half hour will be dedicated to the concrete application of the methodology on a given or free use cases. In this context, participants are asked to work in small groups. In the end, they will have the opportunity to present their work to the other groups.

6. TARGET AUDIENCE

This workshop is especially designed for administrators and practitioners of digital preservation as well as information professionals who deal with service design as well as everyone interested in service business models practices.

7. ACKNOWLEDGMENTS

Our thanks go to the entire DLCM team across institutions led by Pierre-Yves Burgi.

8. REFERENCES


Jefferson Bailey
Internet Archive
300 Funston Avenue
San Francisco, CA, CA 94118, USA
jefferson@archive.org

Maria Praetzellis
Internet Archive
300 Funston Avenue
San Francisco, CA, CA 94118, USA
maria@archive.org

Vinay Goel
Internet Archive
300 Funston Avenue
San Francisco, CA, CA 94118, USA
vinay@archive.org

ABSTRACT

The process of collecting, managing, and serving archives of web published materials poses many challenges. Web archives tend to be quite voluminous, data-wise and number-of-items-wise, with even highly curated collections being many terabytes in size and tens, if not hundreds, of millions of URLs. Web archives also have longitudinal complexity, bearing a temporal aspect similar, but different, to serial publications, with frequent changes in content (and often state of existence) even at the same URL, gobs of metadata both content-based and transactional that has research or administrative value, and have overall characteristics that make them highly suitable for data mining and computational analysis. However, web archiving as a practice is still mostly fractionally-staffed, with few institutions able to contribute engineering resources to innovation and research and development.

The Internet Archive has been pioneering web archiving technologies for 20 years. A number of recent initiatives have sought to reinvigorate innovation around new tools designed to address improving content capture, reconceptualizing access and discovery, and facilitating the increasingly popular scholarly approaches of computational analysis of large born-digital collections. This workshop will provide hands-on demonstrations of, and, when possible, training in, new tools for capturing, analyzing, facilitating use of web archive collections. This includes training in browser based capture methods, curatorial tools, data derivation and extraction methods for research use, demonstration of APIs and their uses for studying archived web data, new search and aggregation tools, and frameworks for web archive processing.

Workshop Scope & Intended Content

Attendees will use a range a tools as part of the workshop, including the ability to run scripts and tools on web archive data that they collector prior to, or bring with them to, the workshop. In general the workshop will cover these areas of practice through a mix of hands-on work, demonstrations, and discussion:

- Participants will use and discuss new API based tools for enhanced access to, and research use of, web archive data. In-production and test APIs will be used for data transport, content analysis, metadata aggregation, and other collection management and discovery activities.
- Participants will be trained in how to generate datasets from web archive collections, including gaining familiarity with the suite of derivative formats, indexing tools, and how to facilitate and support researchers interested in using archived web data.
- Participants will be trained in the latest open-source crawling technologies including browser based capture and API harvesting. This includes both Brozzler, a distributed web crawler, and new other tools designed to harvest social media data. Participants will be given the opportunity to compare crawling mechanisms by using a variety crawling methods, and discuss how they can incorporate these tools in their own institutions.
- Demonstration of recent innovations in providing access to large web archive corpora, including user testing and technical review of new search technologies for querying large-scale web archives, including a review of methods, technologies, interfaces, and other cutting-edge approach to dealing with web archive content discovery.
- Review new dataset formats and tools designed to enable easy data processing, extraction and derivation of web archive data for computational analysis. This includes the development of a community space to share modules and recipes for analyzing data sets that can be reused among scholars as well as the exploration of derivative data formats and interfaces like Python-base Jupyter Notebooks for pre-scripted data analysis.

Requirements & Target Audience

The workshop requires a meeting room with wireless internet access and a projector with screen. Participants will bring their own laptop computers and there should be sufficient power outlets. The workshop hosts will coordinate preliminary activities over email and provide some basic technical support beforehand for workshop attendees.

The target audience for this workshop includes professionals working in digital library services that are collection, managing, or providing access to web archives, scholars using archived web data in their work, or digital library and preservation professionals working to provide computational access to digital collections. Some knowledge of or experience with web archives is helpful but not required and some familiarity or comfort with working at the command line interface is helpful.

Keywords

Web archiving; digital collections; data mining; research
Abstracts

Tutorials
ABSTRACT
Fedora is a flexible, extensible, open source repository platform for managing, preserving, and providing access to digital content. Fedora is used in a wide variety of institutions including libraries, museums, archives, and government organizations. Fedora 4 introduces native linked data capabilities and a modular architecture based on well-documented APIs and ease of integration with existing applications. Recent community initiatives have added more robust functionality for exporting resources from Fedora in standard formats to support complete digital preservation workflows. Both new and existing Fedora users will be interested in learning about and experiencing Fedora features and functionality first-hand.

Attendees will be given pre-configured virtual machines that include Fedora bundled with the Solr search application and a triplestore that they can install on their laptops and continue using after the workshop. These virtual machines will be used to participate in hands-on exercises that will give attendees a chance to experience Fedora by following step-by-step instructions. Participants will learn how to create and manage content in Fedora in accordance with linked data best practices and the Portland Common Data Model. Attendees will also learn how to import content into Fedora and export content to external systems as part of a digital curation workflow.

Keywords
Fedora, repository, linked data, open source.

1. OUTLINE
The tutorial will include three modules, each of which can be delivered in 1 hour.

1.1 Introduction and Feature Tour
This module will feature an introduction to Fedora generally, and Fedora 4 in particular, followed by an overview of the core and extended Fedora 4 features. It will also include a primer on data modeling in Fedora 4, which will set the audience up for the next section.

1.2 Linked Data and PCDM
The Fedora community is deeply invested in linked data best practices; this is exemplified by our alignment with the W3C Linked Data Platform recommendation and the Portland Common Data Model (PCDM). This section will feature an introduction to linked data and PCDM, with a particular focus on the way Fedora implements linked data. Attendees will have an opportunity to create and manage content in alignment with PCDM using the Fedora 4 virtual machine.

1.3 Data Curation Workflows
Fedora 4 is fundamentally a middleware application – it is meant to be used in conjunction with other applications as part of a data curation workflow. This section will provide an overview of the recently developed import/export utility and demonstrate how content can be packaged for both ingest into Fedora and export from Fedora to external preservation systems and services.

2. DURATION
Half-day (3 hours)

3. AUDIENCE
This tutorial is intended to be an introduction to Fedora 4 - no prior experience with the platform is required. Repository managers and librarians will get the most out of this tutorial, though developers new to Fedora would likely also be interested.

4. OUTCOMES
Tutorial attendees will:
- Learn about the latest and greatest Fedora 4 features and functionality
- Learn how to create and manage content in Fedora in accordance with the Portland Common Data Model
- Understand how to import content into Fedora and export content to external systems

5. PRESENTER
David is the Product Manager for the Fedora project at DuraSpace. He sets the vision for Fedora and serves as strategic liaison to the steering committee, leadership group, members, service providers, and other stakeholders. David works together with the Fedora Technical Lead to oversee key project processes, and performs international outreach to institutions, government organizations, funding agencies, and others.
Persistent Identifiers for Digital Cultural Heritage

Jonathan Clark
International DOI Foundation
Rembrandtlaan 12
1231 AC Loosdrecht
The Netherlands
+31654733572
jonathanmtclark@gmail.com

Remco van Veenendaal
Marcel Ras
National Archives
P.O. box 90520
2509 LM The Hague, Netherlands
+31629451951 / +31703140180
remco.van.veenendaal@nationaalarchief.nl
marcel.ras@kb.nl

Juha Hakala
National Library of Finland
Fabianinkatu 35
00170 Helsinki
Finland
+358 50 382 7678
juha.hakala@helsinki.fi

ABSTRACT
This is an introductory level tutorial. The goals of the tutorial are: to explain what persistent identifiers are and why they are so important and what the criteria are for trustworthy identifier systems; to present the latest version of a decision tree tool that helps determine which system is the most appropriate for a particular set of needs; and to discuss the current features of PIDs and the development of future properties.

Keywords
Persistent Identifiers; Handle System; DOI; ARK; URN; URN:NBN; PURL.

1. INTRODUCTION
Over the past years, a growing number of collections belonging to archives, libraries, media, museums, and research institutes have been digitized and made available online. These are exciting times for ALM institutions. On the one hand, they realize that, in the information society, their collections are goldmines, as “data is the new gold” [1].

On the other hand, unfortunately most heritage institutions do not yet meet the basic preconditions for long-term availability of their collections. Apart from the problems of digital preservation the digital objects often have no long lasting fixed reference. URL’s and web addresses change. For instance, some digital objects that were referenced in Europeana and other portals can no longer be found. References in scientific articles can have a very short life span, which is damaging for scholarly research, as is shown in the well-known article of Herbert van de Sompel and others [2].

Thus, in this digital world there is a need to unambiguously determine what a resource is and where it can be found in a way that is persistent over time. However, the identifiers themselves are simply strings of numbers and not inherently persistent. What make them persistent are the policies, organization and agreements that sit behind these numbers.

Many different systems have emerged over the years, some more robust than others. What is the difference between them all and how to choose between the various options available? What are the criteria to judge whether a PI system can be trusted or not? [3]

Finally, of course a choice must be made for a PI system that meets the long-term needs of the ALM institution. In 2015, the Dutch Digital Heritage Network (NDE) [4] started a two-year work program to co-ordinate existing initiatives in order to improve the (long-term) accessibility of the Dutch digital heritage for a wide range of users, anytime, anyplace. [5]

The NDE is a partnership established as an initiative of the Ministry of Education, Culture and Science. The members of the NDE are large, national institutions that strive to professionally preserve and manage digital data, e.g. the National Library, The Netherlands Institute for Sound and Vision, the Netherlands Cultural Heritage Agency, the Royal Netherlands Academy of Arts and Sciences, the National Archive of the Netherlands and the DEN Foundation, and a growing number of associations and individuals both within and outside the heritage sector.

Meanwhile, other institutions such as the British Film Institute, the National Diet Library of Japan and the European Union are moving forward with interesting projects to assign persistent identifiers to their archives.

2. TOPICS
The tutorial on persistent identifiers for digital cultural heritage will consist of three parts

1. Introduction: what are persistent identifiers (PI) and why are they important?
   a. Knowing what’s what and who’s who
   b. The importance of Social Infrastructure
   c. Recommendations of the Research Data Alliance (RDA) for reliable & trustworthy PIs [6]
   d. Review of current identifier systems

2. The NDE decision tree for PIs
   a. A tool in form of an online questionnaire that guides cultural heritage organizations through the process of selecting a particular type of Persistent Identifier (Handle, DOI or NBN:URN) [7]
   b. Discuss the applicability of the decision tree outside the Netherlands

3. Persistent Identifiers: current features and future properties
   a. Resolution
   b. Future developments of PIDs

SCOPE
The intention is that this tutorial be “PI-agnostic”; that is, any persistent identifier (Handle [8], DOI [9], ARK [10], URN:NBN [11], URN [12] and PURL [13]) will be discussed.

3. INTENDED AUDIENCE
Anyone interested in learning more about persistent identifiers for digital information. This tutorial has a strong practitioner focus and will be especially interesting for those working with Digital Archives and Digital Collections. This tutorial also acts
as an introduction and level set for the Workshop on Smart Persistent Identifiers. Participants may be interested to read some background articles [14] [15] [16] [17]

4. EXPECTED LEARNING OUTCOMES
Participants will leave this tutorial with a clear understanding of what persistent identifiers are and why they are important. They will have an overview of the different identifier systems in use today and have ample opportunity to answer their questions.

5. SPEAKERS
The tutorial speakers will be Jonathan Clark, Managing Agent for the DOI Foundation, Remco van Veenendaal of the Dutch National Archives, Marcel Ras of the National Coalition for Digital Preservation and and Juha Hakala of the National Library of Finland.

6. REFERENCES
[10] ARK (Archival Resource Key) Identifiers https://confluence.ucop.edu/display/Curation/ARK
[17] Introduction to persistent identifiers, The THOR Project Knowledge Hub https://project-thor.readme.io/docs/introduction-to-persistent-identifiers
Understanding and Implementing PREMIS

K. Bredenberg  
National Archives of Sweden  
P.O. Box 12541  
Stockholm  
Sweden  
karin.bredenberg@riksarkivet.se

A. Dappert  
British Library  
96 Euston Road  
London, NW1 2DB  
United Kingdom  
Angela.Dappert@gmail.com

E. Ziera  
The Royal Danish Library  
Søren Kierkegaards Plads 1  
1016 Copenhagen K  
Denmark  
elzi@kb.dk

ABSTRACT
This tutorial will provide participants with an introduction to the PREMIS Data Dictionary [1]. It will give a basic overview of the standard and explore different models of implementation.

KEYWORDS
Preservation strategies and workflows; Infrastructure, systems, and tools; Case studies, best practices and novel challenges; Training and education

1 INTRODUCTION
The PREMIS Data Dictionary for Preservation Metadata is a specification that provides a key piece of infrastructure for digital preservation activities, playing a vital role in enabling the effective management, discovery, and re-usability of digital information. Preservation metadata provides provenance information, documents preservation activity, identifies technical features, and aids in verifying the authenticity of digital objects. PREMIS is a core set of metadata elements (called “semantic units”) recommended for use in all preservation repositories regardless of the type of materials archived, the type of institution, and the preservation strategies employed.

2 SUMMARY OF TUTORIAL
The PREMIS Data Dictionary was originally developed by the Preservation Metadata: Implementation Strategies (PREMIS) Working Group in 2005 and revised in 2008 and 2015. It is maintained by the PREMIS Editorial Committee and the PREMIS Maintenance Activity is managed by the Library of Congress [2].

We have seen a constant call for PREMIS to undertake tutorials, such as this, as more and more organisations come to grips with digital preservation. This tutorial provides an introduction to PREMIS and its data model and an examination of the semantic units in the Data Dictionary organized by the entities in the PREMIS data model, objects, events, agents and rights.

In addition it presents examples of PREMIS metadata and a discussion of implementation considerations, particularly using PREMIS in XML and with the Metadata Encoding and Transmission Standard (METS) [3]. It will include examples of implementation experiences through the institutional experience of the tutors.

The tutorial aims at developing and spreading awareness and knowledge about metadata to support the long term preservation of digital objects.

3 CONTENT OUTLINE
The draft outline for the tutorial is outlined below.

Introduction to PREMIS
- Background (brief history and rationale of PREMIS)
- Benefits of implementing PREMIS
PREMIS in detail
- Core entities
- Simple examples to build familiarity
Implementation
- PREMIS in METS
- Case studies
- Support and the PREMIS community
- Conformance
Next Steps
- Round table discussion for institutional plans
Wrap up

4 INTENDED AUDIENCE
The tutorial will benefit individuals and institutions interested in implementing PREMIS metadata for the long-term management and preservation of their digital information but who have limited experience in implementation. Potential audience includes cultural heritage operators, researchers and technology developers, professional educators, and others involved in management and preservation of digital resources.

The presentation is structured into two parts. The first part is intended for beginners with little or no knowledge of PREMIS, while the second half will cover more advanced use of PREMIS for practicing users and present the opportunity for questions and discussions.

5 EXPECTED LEARNING OUTCOMES
Participants will understand:
- What PREMIS is and why it exists;
- How PREMIS has changed across versions;
- The benefits of implementing PREMIS;
- The nature of the existing PREMIS community;
- The critical role PREMIS plays in the digital preservation community.
In addition, participants will get insight into:
- How PREMIS may be used in conjunction with METS;
- How different organisations implement PREMIS within their own repositories;
- The nature of conformance with PREMIS.

5 SHORT BIOGRAPHIES OF ORGANIZERS

Karin Bredenberg is a technical advisor on metadata at the Swedish National Archives. She graduated in Computer Engineering (programming C#), with a Bachelor of Science at the Royal Institute of Technology (KTH) in Stockholm 2006. Bredenberg works with Swedish adaptations of international metadata standards and is responsible for the common specifications regarding e-archiving and e-records management maintained by the Swedish National Archives. She currently serves as co-chair of the Society of American Archivists Technical Subcommittee on Encoded Archival Standards (TS EAS), chair of the newly created DLM Forum Archival Standards Board (DAS Board) as well as a member of the PREMIS Editorial Committee and as a member of the METS Board. During 2011-2014 she served as the work package lead regarding metadata exchange formats specifications in the Swedish project eARD. The project part on exchange format specifications received the Swedish e-Diamond Award 2014. 2009-2014 Bredenberg was part of the European APEX (http://www.archivesportaleurope.net/) as a standards specialist and 2014-2017 part of the European project E-ARK (http://www.eark-project.com/) working on Information Packages and use of metadata standards.

Angela Dappert Dr Angela Dappert is Project Manager for the EU-funded THOR project (project-thor.eu). She has widely researched and published on digital repositories and preservation; consulted for archives and libraries on digital life cycle management and policies; led and conducted research in the EU-co-funded Planets, Scape, TIMBUS, and E-ARK projects; and applied digital preservation practice at the British Library through work on digital repository implementation, digital metadata standards, digital asset registration, digital asset ingest, preservation risk assessment, planning and characterization, and data carrier stabilization. She has applied work towards preservation of research data and processes, software environments and ejournals, with an emphasis on interoperability and standardisation. Angela holds a Ph.D. in Digital Preservation, an M.Sc. in Medical Informatics and an M.Sc. in Computer Sciences. She serves on the PREMIS Editorial Committee and a number of advisory bodies.

Eld Zierau is member of the PREMIS Editorial Committee, since 2013. She is a digital preservation researcher and specialist, with a PhD from 2011 within digital preservation. Originally, she is a computer scientist, and has worked with almost all aspects of IT in private industries for 18 years, before starting in digital preservation in 2007. She has been working with many aspects of digital preservation, and she is involved as an architect or a consultant on major initiatives such a new digital repository including data modelling of metadata for preservation.

REFERENCES
Diverse Digital Collections Meet Diverse Uses: Applying Natural Language Processing to Born-Digital Primary Sources

Christopher A. Lee
University of North Carolina at Chapel Hill
calle@unc.edu

ABSTRACT
In this tutorial, participants will learn about and gain hands-on experience with products of the BitCurator NLP project, which is developing software for libraries, archives and museums (LAMs) to extract and expose features (e.g. people, places, organizations, events, relationships, topics) in text extracted from born-digital materials. The services and methods can be used by LAM professionals for appraisal and description, as well as facilitating a wider range of access and use scenarios.

CCS CONCEPTS
Information systems—Digital libraries and archives
Computing methodologies—Natural language processing
Security and privacy—Data anonymization and sanitization

ADDITIONAL KEYWORDS AND PHRASES
BitCurator NLP; archival processing; named entity recognition; text processing; topic modeling

INTRODUCTION
Libraries, archives and museums (LAMs) are increasingly called upon to move born-digital materials from their original locations into more sustainable preservation environments. Information professionals must be prepared to extract digital materials from removable media in ways that reflect the rich metadata and ensure the integrity of the materials. They must also support and mediate appropriate access: allowing users to make sense of materials and understand their context, while also preventing inadvertent disclosure of sensitive data.

There has been a significant shift in recent years toward the adoption of digital forensics tools and methods by LAMs, in order to meet the above goals. This process has been facilitated by the BitCurator project (2011-2014), funded by the Andrew W. Mellon Foundation, which has packaged and disseminated an open-source software environment¹ that allows users to create disk images; extract data and metadata from disks or directories; scan bitstreams for the presence of potentially sensitive data values; characterize the contents of disks; and perform other practical tasks, such as scanning for viruses, finding duplicate files, mounting forensically packaged disk images, generating cryptographic hashes, and viewing hexadecimal representations of bitstreams.

The BitCurator Access project (2014-2016), also funded by the Andrew W. Mellon Foundation, investigated mechanisms for providing access to forensically-acquired data. A major product of the project has been BitCurator Access Webtools, which allows users to dynamically navigate filesystems of disk images, as well as searching over the content of many common files types contained within the images.² The project also created BitCurator Access Redaction Tools to redact strings and byte sequences identified in disk images.³

BITCURATOR NLP PROJECT
BitCurator NLP (2016-2018), funded by the Andrew W. Mellon Foundation and led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS), is developing and disseminating software for identifying, extracting and exposing contextual entities from the wide diversity of born-digital materials that LAMs already hold and continue to receive. This includes helping to identify and explore information based on specific entities (e.g. people, places, organizations, events) of interest to curators and researchers.

There are many existing mature open source natural language processing platforms, including platforms that provide web services and RESTful application programming interfaces (APIs) and integration with industry-standard testing and training corpora. Production-quality open source software toolkits for natural language processing include OpenNLP (Java-based) and NLTK, Pattern, and spaCy (Python-based).

Our target use cases differ from previous work in two fundamental ways. First, disk images are internally complex and require a significant software dependency stack that is already available through the BitCurator environment and BCA Webtools. These include the ability to read, mount and provide access to the contents of various filesystems, as well as extracting, presenting and reporting on their data and metadata. A second distinguishing factor is that disks may contain a broad range of file types and data encodings, requiring substantial

¹ https://wiki.bitcurator.net/
² https://github.com/bitcurator/bitcurator-access-webtools
³ https://github.com/BitCurator/bitcurator-access-redaction
pre-processing to extract content so that it can be processed by NLP tools and organized into meaningful reports, access points and visualizations. BitCurator NLP is building from a variety of existing tools and initiatives to provide services that LAMs can be run independently or integrate into existing software environments and access portals via simple application programming interfaces (APIs).

BitCurator NLP is exploring approaches that focus on improving the utility of reports produced about the contents of born-digital collections. Using data extracted from open text using NLP tools, along with techniques from digital forensics research to eliminate or deemphasize those that appear to be irrelevant or common to the system rather than the documents themselves (e.g., names and email addresses of developers or organizations that created the software used to produce a given document), the project team will also develop guidelines describing how to apply the tools in ways that support common access and research use cases.

The BitCurator NLP team is ensuring close integration between the existing functionality of the BitCurator environment, BitCurator Access Webtools and the software developed by the BitCurator NLP project. For example, we are increasingly making the various elements of the BitCurator environment available as self-contained software installers (software packages that may be installed in Ubuntu and Debian Linux environments), so users can selectively install and update them as they find most useful. Institutions could load all of the access tools onto the same machine (or virtual machine) as the one they are using for the initial processing, or they could instead decide to run those tasks in different environments in order to better manage and allocate their resources.

TUTORIAL OUTLINE

This half-day tutorial will have the following structure:

- Brief lecture and discussion that focuses on the motivation for using the tools and several foundational technical concepts
- Demonstration and hands-on exercises that demonstrate specific tools and methods, including text extraction, entity and entity relationship extraction and topic segmentation and recognition
- Brief summary of targeted visualization scenarios using e.g. the scattertext python library
- Concluding discussion about implications for participant’s institutional practices

EXPECTED LEARNING OUTCOMES

This tutorial will prepare participants to use the open-source BitCurator NLP tools to process, investigate and visualize born-digital collections. This will include text extraction from heterogeneous collections of file and executing NLP tasks such as part-of-speech tagging, entity and entity relationship extraction, and topic segmentation and recognition.

Upon completion of this tutorial, participants should understand several of the major motivations and uses cases for applying the BitCurator NLP tools to meet digital preservation and access objectives. They should also recognize the limitations and how they might be used in combination with other tools and methods. Participants will also become aware of the resources that are available for learning more about the software and engage with other users after completion of the tutorial.

ACKNOWLEDGEMENTS

This work has been funded by the Andrew W. Mellon Foundation. The BitCurator NLP team is composed of Jacob Hill, Christopher A. Lee, Sunita Misra and Kam Woods. We have also benefited from contributions of the project Advisory Board.

INTENDED AUDIENCE

This tutorial should be of interest to information professionals who are responsible for acquiring or transferring collections of digital materials. Another intended audience is individuals involved in digital preservation research, development and IT management, who will learn how data generated by the BitCurator

NLP tools can complement and potentially be integrated with data generated by other tools and systems.

\(^4\) https://github.com/ImaginativeLabs/scattertext
Abstracts

Organized Sessions
Digitization and Preservation of Historical Images Based on the Investigation of Historical Materials of Pre-modern Japanese History by Historiographical Institute the University of Tokyo

Extended Abstract

Taizo Yamada
Historiographical Institute / Collaborative Research Organization for Historical Materials on Earthquakes and Volcanoes
The University of Tokyo
Japan
t_yamada@hi.u-tokyo.ac.jp

Akiyoshi Tani
Historiographical Institute
The University of Tokyo
Japan
tani@hi.u-tokyo.ac.jp

Toru Hoya
Historiographical Institute
The University of Tokyo
Japan
hoya@hi.u-tokyo.ac.jp

ABSTRACT

In the session, we report on the investigation on historical materials of pre-modern Japanese history by Historiographical Institute the University of Tokyo, the digitalization of the historical materials based on the work, and the efforts aimed at long-term preservation and utilization of them.

KEYWORDS

historical material, Japan, history, digitalization, pre-modern

1 INTRODUCTION

Historiographical Institute the University of Tokyo, ‘HI’ for short, has been investigating historical materials concerning pre-modern Japanese history. In the process, we have been making and managing metadata as the results and shooting the materials. Recently, we have been digitizing of the historical materials based on the results so that the data can be used at long-term preservation and utilization of them. In the paper, we report on the approach and the efforts.

2 COLLECTING AND REPRODUCING OF HISTORICAL MATERIALS

2.1 PURPOSE

HI has the three missions which are to conduct research on primary source materials concerning Japanese history, compile based on and used by the materials, and publish the results. In order to the research, HI has been investigating and collecting historical materials concerning pre-modern Japanese history in earnest from 1885 and the operation has been ongoing for about 120 years. HI has two type compilations: chronological organized materials collections such as the Dai Nihon Shiryo and Dai Nihon Ishin Shiryo, and complication organized around specific archives such as the Dai Nihon Komonjo and the Dai Nihon Kokiroku. In order to make the compilations widely available, HI has been publishing the results since 1901.

2.2 REPRODUCTION AND DIGITALIZATION OF THE MATERIALS

The results have been accumulated as reproduced by several different means, such as microfilm, calligraphic reproduction, and photography and by making them public, we have enabled sharing them among researchers. The efforts have been contributing greatly to the development and deepening of Japanese historical study. From the beginning of HI, we had been creating a calligraphic reproduction as a result of the operations. In parallel with this, the replication by a photographic plate based on glass had also been carried out. We had been beginning micro camera shooting from about 1950 and genuinely engaged in the operation from about 1970. Shooting with a digital camera has been starting from around 2010. We scanned films by the shooting with the micro camera and released digitized images from the web site or databases of HI. From the late 1990s, we have taken photographs of historical materials from collections hold by HI which also includes the calligraphic reproductions, have been digitizing them. In HI, the method of reproducing and collecting of historical materials has been changed with the time-changing. In order to make long-time use of the accumulated historical materials, we have formulated a method for uniformly performing photography, digitization, management and other processes, and constructed and operated a system to realize them. In the session, we report about the approaches and the efforts.

2.3 MATERIALS IN OTHER COUNTRIES

Since around 1930 we had been beginning to collect Japanese historical materials in other countries. After that, we have been collecting historical materials by microfilm with the support of the U.A.I (Union Academique Internationale). We are also systematically collecting historical materials which are concerning pre-modern Japan in Russia and China, because there were few researches on the materials. Recently, we are promoting...
digitization and metadata addition to make these available. We also report on the approaches.

ACKNOWLEDGMENTS
This work was partially supported by JSPS KAKENHI Grant Number 26220402, and International Center for the Digitization of Premodern Japanese Sources in HI.
Constructing Integrated Studies of Cultural and Research Resources, and Renovating Sharing Infrastructures of Research Resources in Japanese History and Culture

Introduction

The Center for Integrated Studies of Cultural and Research Resources
National Museum of Japanese History
Japan
meta-shiryo@rekihaku.ac.jp

The National Museum of Japanese History (NMJH, REKIHAKU called in Japanese) opened the Center for Integrated Studies of Cultural and Research Resources to establish its core research project in 2016, “Constructing Integrated Studies of Cultural and Research Resources”. This project offers unique insight into a variety of studies and Japanese historical resources, through multidisciplinary collaboration with universities, museums, and other institutions. Here, we share information on the following topics.

Researchers in humanities and sciences focusing on “history” need to collaborate with their peers in other relevant fields to produce more diverse and substantial academic data. The plan for the new academic field is to construct a “research circulation model” (Figure 1) between academic studies and resource information through advanced cooperative studies. In addition, resources in institutes encompass a wide variety of materials. However, the poor quality of current information infrastructure has hampered researchers’ access to original data and convenience in producing the primary evidence of research results.

The number of researchers specializing in the study of China, Korea, and Taiwan is increasing owing to the rapid progress of the networks of historical resource information in these countries. In contrast, their peers in Japan are decreasing owing to non-digitized and partially closed systems. This trend is particularly apparent in the younger generation that is hard-wired for the use of the internet. In our poster, we show an example of our collaborative studies with relevant local and international institutes. The name authorities and concept label authorities in the translation are related to the entities mentioned in the RDF (Resource Description Framework). In addition to the issue of preserving data in case of disaster, the poster explains our development of a support system in the affected areas and an advanced sharing infrastructure.

Our project, “Integrated Studies of Cultural and Research Resources” approaches Japanese historical resources through interdisciplinary study in the humanities and sciences using information infrastructure. Classifying various cultural and research materials into time periods, regions, and research fields, and analysis with an interdisciplinary scope can lead to more advanced sharing infrastructure and additional cooperative studies. “Research Circulation Access Model” is formulated in line with the aim of the NMJH’s core project: to enable access to resource information based on a research result supports another new research. Moreover, creating an environment that enables access to resource information in Japan can ensure ease in referring to “the original evidence” of various research results. NMJH implements a “museum-based research integration” using the research results as permanent or special exhibits and promotes interdisciplinary studies and educational activities in the humanities and sciences. Such accumulated expertise supports the foundation of the research circulation access model.

Unit for Digital Humanities (Figure 2) aims to create an environment of information infrastructure that enables access to resource information in Japan. In this unit, knowledge is analysed and accumulated to support access to resource information based on research results. Focusing on Japanese historical resources as “materials”, Unit for Cooperation in Different Fields (Figure 3) aims to conduct interdisciplinary study in the humanities and sciences. As such, this unit further applies richer information taken from each resource to various research fields. In cooperation with institutes, Unit for Regional Cooperation and Education (Figure 4) aims to create a model applying research results based on “Integrated Studies of Cultural and Research Resources” to regional societies. Using the results of Unit for Digital Humanities and Unit for Cooperation in Different Fields, this unit explores local histories and cultures, with the collaboration of institutes. It also conducts outreach activities, such as education programs and exhibitions focused on NMJH’s research results.

To achieve the current results based on resources in Japan, we also constructed an initial application of Linked Data and ILIF (International Image Interoperability Framework). It has five features. First, each resource has its own one address, because it enables access to resources from the research results using Google and other search services. Second, it can access resources that follow another resource, such as “a resource- an institute- another
resource.” Third, to apply smart devices, spatial information is added to the resources. Fourth, the system and the data are divided in this application. If a system is replaced or updated in the future, the data should remain usable. An institute provides infrastructure data and other institutes or researchers can use the data to create any applications. In so doing, an access path in the institute needs to be confirmed. Finally, RDF data are applied to the database.

We are currently implementing all the functions and tests concerning actual data: involving 150 thousand triples of our museum collection, 20 thousand triples of related facto data, and approximately 500 IIIF images. We plan to apply the data sets of other Japanese universities by August 2017, and this application will be fully launched at the beginning of 2018.

We are now constructing a system to connect the painter to other Ukiyo-e information of the same person. Some of these links use metadata including FOAF (Friend of a Friend) as the information data of people, and other specific links use vocabularies defined by our project.

With this project, NMJH emphasizes collaboration with relevant local and international institutes. Cooperative studies with universities and the renovation of a digital network through collaboration with institutes will also enable a backup of resource information in case of disaster, further supporting infrastructure. Moreover, we are developing an English translation of the available resource information and creating an environment that enables access to resources in Japan. We plan to conduct cooperative research, exhibitions, and international symposiums.

NMJH is constructing and developing a digital network and English translation of the available resource information through collaboration with relevant institutes in Japan. The goal is to create an environment that enables access to resources in Japan. A backup of resource information, including non-disclosure for various reasons, will further support and restore infrastructure in case of disaster. We welcome any relevant information and appreciate cooperation in the construction of integrated studies of cultural and research resources and renovation of this digital network.

---

**Figure 1:** Research Circulation Access Model.

**Figure 2:** Unit for Digital Humanities renovating access data.

**Figure 3:** Unit for Cooperation in Different Fields applying constructed data to studies.

**Figure 4:** Unit for Regional Cooperation and Education applying results to societies.
Issues in archiving manga, anime, games and related cultures

Abstract†

Kaichiro Morikawa
School of Global Japanese Studies
Meiji University
4-21-1 Nakano, Nakano-ku, Tokyo
Japan
GFH04434@nifty.ne.jp

KEYWORDS
manga, anime, video game, fanzine

1 INHERENT ISSUES

When viewed as subjects of preservation, manga (comics), anime (animated cartoons), and video games are totally different animals. Mangas were -- and to a large extent still are -- printed and bound media. Animes are mostly shot and distributed in digital nowadays, but piles of films and videotapes are left from the last century. Video games require obsolete (or to-be-obsolete) devices such as consoles and arcade cabinets if they are to be playable in their original state, and the emergence of online games further complicates the challenges of preservation. The one thing in common is that they are mostly commercially produced and are copyrighted, making archival digitization by public or academic sectors nearly impossible. Plus there are multitudes of to-be-archived materials other than the end product, such as hand-drawn original drawings, preliminary manuscripts and documents, advertising materials, all sorts of related merchandise, and fanzines, just to name a few. Institutions and communities such as manga libraries and preservation circles of video games do exist, but most of them dedicate themselves to a single and not multiple medium.

Yet, in Japan, in the past few years, new situations have emerged at a political level, which may encourage the interconnection and integration of archival efforts concerning these mediums, with a potential solution to overcome the copyright issue.

2 RECENT TOPICS

In late 2014, an all-party parliamentary group on manga, anime, and video games was established, dubbing itself as the MANGA Giren (MANGA: acronym of Manga, ANime, GAmes; Giren: APPG in Japanese). Core members include former prime minister Taro Aso, and former cabinet members Keiji Furuya and Hiroshi Hase, the three of whom were involved back in 2009 with the governmental plan to erect a national center to house manga, anime, and video games together with media art. After becoming a subject of political attacks against the then prime minister Aso, the plan was shortly terminated due to the change of government. Such is the context of one of the founding purposes of the APPG; to reboot the plan, revamped under the name of MANGA National Center.

With political efforts by the APPG, the following lines were included in the Basic Policy on Economic and Fiscal Management and Reform of 2017, a major governmental policy which guidelines Japan's fiscal budget:

"... plans will be made to preserve, utilize, and transmit cultural properties by putting in place central functions pertaining to opening up and utilizing cultural properties, and measures for a digital archive will be promoted. Moreover, information hubs will be put in place for the media arts, such as manga, anime, and computer games, of which our nation is justly proud [1]."

Furthermore, the Basic Act for the Promotion of Culture and the Arts was amended in 2017 to address the preservation of production materials of media arts (a term used by the Agency for Cultural Affairs of Japan to address manga, anime, video games together with media art) [2].

One of the core concepts of the MANGA National Center, which the APPG aims to secure its execution, is to establish the Center as a branch of the National Diet Library of Japan. The NDL is the legal deposit library of Japan, as well as the institute which holds exceptional legal rights to digitize copyrighted materials for the purpose of preservation.

The presentation shall elaborate on the issues in archiving and preserving manga, anime, and video games, as well as on the recent situation which is brought about by the APPG.

REFERENCES


Archiving and Utilization of Manga Materials: in the case of Kyoto International Manga Museum

Abstract†

Tadahiro Saika
International Manga Research Center,
Kyoto Seika University
Kyoto International Manga Museum
604-0846 Karasuma-Oike,
Nakagyo-ku, Kyoto
Japan
saika@kyotomm.jp

KEYWORDS
manga, Kyoto International Manga Museum, popular culture

1 INTRODUCTION
Kyoto International Manga Museum is a communal project of Kyoto Seika University and the city of Kyoto; it opened in November 2006 and aims to collect, preserve, and exhibit manga (comics) and related materials, and its project is to research manga culture. From the perspective of regular visitors, it is a cultural facility that combines the function of a library specialized in manga and a museum that regularly holds exhibitions. The aim of this presentation is to consider the meaning of this facility's organization as a library and its efforts to archiving manga materials.

2 MANGA MATERIALS AT THE KYOTO INTERNATIONAL MANGA MUSEUM
As a manga archive, the Kyoto International Manga Museum currently holds over 300,000 manga materials. Most of them are primary sources like single manga volumes, manga magazines, but there are also books on manga research. This is where it is different from Western comic museums who archive original comic art as a primary resource. Our museum doesn’t put much effort into collecting original art, and we don’t have much in our holdings. There is a public library space where visitors to the museum can peruse 50,000 of our 300,000 manga. These volumes were mainly published between the 1970s and 2005, and fill the books shelves you can find against almost all walls in the museum, what we call the “Wall of Manga”.

These collection of the Kyoto International Manga Museum is made up from donations for the most part. Therefore the most of the manga related materials Kyoto International Manga Museum collects are not original manga art, but manga books. That we are able to build a collection out of donations from the general public shows how many people enjoy manga as popular culture in Japan. Our mission to collect manga magazines is also very important given that most of the time. Manga magazines do not just contain manga, but also background information of manga as popular culture. It is very hard to gain this sort of background from one page of original art.

One of our most important policies on collecting manga materials is that when collecting, we don’t discriminate based on content. We do not limit our collection to manga and the work of manga artists that has historical value, we also collect the work of manga artists forgotten by history, works that have arguable, or little to no artistic and literary value, equally. Of course, we also collect works of adult nature not collected by general libraries. We even collect illegal material such as pirated books, because of their cultural historical value.

This presentation shall introduce on detail of our effort to collect and utilize manga materials as mentioned above, and examine current issues on archiving manga culture comprehensively.
Abstracts

Posters/Demos Lightning Talk
Abstract

Over the decade the Open Access movement has evolved with the support by academic communities. In this context, we recently promote to make rare materials (e.g., manuscripts) public on the web, in order to support researchers in humanities. We have digitized the rare materials and investigated how to make them public. In this paper, we introduce our novel digital archive system Kyoto University Rare Materials Digital Archive. The digital archive system provides free access to the digitized rare materials to academic communities. It supports the IIIF (International Image Interoperable Framework), which promotes interoperable image delivery on the web. The digital archive system is composed of two main components, content management system based on Drupal and IIIF image server. In the future, we investigate how to assess the usage of the rare material images and how to present transcriptions and translations for them.

CCS CONCEPTS
• Information systems → Digital libraries and archives; • Applied computing → Arts and humanities; Publishing;

KEYWORDS
digital archive, open access, IIIF

1 INTRODUCTION

Open access to scientific articles refers to the removal of barriers (including price barriers and permission barriers) from accessing them [2]. Over the decade the Open Access movement has evolved with the support by academic communities. In this context, Kyoto University adopted the Kyoto University Open Access Policy in 2015. The policy mandates faculty members of Kyoto University to make their academic articles public on the web using the institutional repository Kyoto University Research Information Repository (KURENAI). In addition to academic articles, we recently promote to make rare materials (e.g., manuscripts) public on the web, in order to support researchers in humanities. We have digitized the rare materials and investigated how to release them onto the web. To this end, we launched the Kyoto University Rare Materials Digital Archive (http://rmda.kulib.kyoto-u.ac.jp/), a digital archive system, which provides free access to the rare material images.

Section 2 provides the details of our digital archive system. Thereafter, we describe what we investigate for the digital archive system in the future in Section 3.

2 DIGITAL ARCHIVE SYSTEM

This section describes the details of our digital archive system. Section 2.1 provides the system architecture of the digital archive system. Subsequently, we show the user interface in Section 2.2.

2.1 System Architecture

Figure 1 illustrates the system architecture of the digital archive system. The system comprises two main components: content management system (CMS) based on Drupal and IIIF image server. The CMS keeps metadata of the rare materials and their bibliographical introductions. The administrator manages contents of the digital archive system through the CMS. Users can search images by executing queries on the CMS. The IIIF image server stores images as well as IIIF manifests, which describe metadata. The images are delivered to the CMS via the IIIF Image APIs. The IIIF manifests are generated based on metadata stored in the CMS. They are brought to the CMS via the IIIF Presentation APIs. The images are stored in the Tagged Image File Format (TIFF). The backup server keeps copies of the images.

KULINE is an online catalog, which stores metadata of materials held by the Kyoto University Library Network. It includes metadata of the rare materials in addition to those of books and journals. Thus, users can search the rare materials not only on the digital archive system but also on the online catalog. The cover images of the rare materials are stored for providing thumbnails. In the future, we have to investigate how to synchronize metadata in the digital archive system with those in the online catalog.

2https://repository.kulib.kyoto-u.ac.jp/dspace/?locale=en

This paper elaborates on the digital archive system. It supports the IIIF (International Image Interoperable Framework), which promotes interoperable image delivery on the web. The IIIF is a community of libraries, research institutions, museums, archives, nonprofits and commercial organizations that are committed to interoperable image delivery on the web [1]. The community has defined a set of application programming interfaces (APIs) for delivery of images as well as their metadata. Thus, the rare material images available in the digital archive system can be accessed and reused by using these APIs.
2.2 User Interface

Figure 2 is a screenshot of the top page of the digital archive system. "Pick Up" in the middle provides several categories, in which important or/and popular rare materials are classified. Collections are given in "Collection" in the bottom. Users can execute queries for searching rare materials. In terms of viewers for the rare material images, we support both Universal Viewer4 and Mirador5 that are most popular in the IIIF.

3 FUTURE WORK

Basing on the digital archive system introduced in Section 2, we will further tackle the following issues to improve it.

Metric of image usage It is important to evaluate the detailed usage of the rare material images. In the case of traditional digital archive systems, we can assess it using page views (PVs). In contrast, it is inappropriate for our digital archive system, since the images are stored in the IIIF image server. Although we may use the number of IIIF Image API calls as metric, it does not reflect the actual usage. Because IIIF Image APIs are called every time users zoom-in and zoom-out images. Therefore, we will develop an appropriate metric by investigating the correlation between actual usage (e.g., dwell time) of images and the number of IIIF Image API calls.

Understanding user activity patterns Furthermore, we will analyze IIIF Image APIs calls, in order to understand user activity patterns on the digital archive system. Since IIIF image APIs are called every time users zoom-in and zoom-out images, we can see more precise patterns. These patterns may reveal, for instance, when users get difficulties and stop browsing images. This insight helps us to improve the digital archive system.

Generating transcriptions and translations The IIIF allows to present annotations over images. Thus, it allows to show transcriptions and translations, which assist users in understanding contents better. However, only a few rare materials have transcriptions or/and translations. In order to populate metadata of rare materials, it is necessary that these rare materials are studied by researchers. By making the rare materials public, we hope that studies in the rare materials make progress and researchers generate metadata for them.

Presentation of transcriptions and translations The IIIF provides functionalities of presenting transcriptions and translations as annotations over images. We have to investigate how to align images and annotations, since annotations do not have positions over images.

REFERENCES


Activities of "Research Resource Archive, Kyoto University"

Haruyoshi Gotoh  
Kyoto University Museum  
Japan  
h.gotoh@inet.museum.kyoto-u.ac.jp

Kentaro Toda  
Kyoto University Museum  
Japan  
toda.kentarou.4r@kyoto-u.ac.jp

Hidetoshi Nagamasu  
Kyoto University Museum  
Japan  
nagamasu@inet.museum.kyoto-u.ac.jp

"Research Resource Archive, Kyoto University" (KURRA) is a project to preserve materials that show educational and research activities in Kyoto University, and to provide them for the further activities. These materials consist of photographs, films, recordings, field books, records of research meetings, lecture notes, and manuscripts, as primary sources.

For the project, we developed a digital archive system. It is a fundamental software to manage archival information, digitized data, and related movies. It helps various archival operations, such as management of metadata, storing digitized data of materials, and opening them to the public. We also called for suggestions for a nickname of the system from the public, and "Peek" was chosen as the nickname.

Figure 1: A screen of “Peek.” Peek has over 63,000 digitized objects, and various finding aids including over 48,000 metadata records are available.

Digitized archival materials and their information are not only searched and browsed, but also provided for various services. People can use them for teaching, exhibitions, workshops, and symposia. It is also possible that these activities reveal the existence of unknown archival materials.

Furthermore, based on the digitized materials, we create movies that present educational and research activities in Kyoto University, and show them to the public.

The important characteristic of KURRA is that management of materials in KURRA complies with fundamental principles of archives and standards of metadata. However, it accepts also materials that normal archives do not.

Bringing materials into KURRA is a process in three steps: investigation, appraisal, and registration. First, we investigate a place where materials are kept and record current condition. Then, we decide what to keep by discussing with owners and relevant researchers. Finally, we register the materials from same origin as a “collection.”

Preservation of archives has two aspects: preservation of information that materials hold, and physical preservation of materials. We create metadata for a collection, namely indices, descriptions, and documents of operations. Also, we preserve archival materials and data themselves.

Note that physical preservation of the materials may be performed in other departments or organization due to technical reason.

Figure 2: We also run the “Audio-Visual Station” that is a public facility for using “Peek” and viewing related movies.

KEYWORDS
Digital archive, university museum, digital preservation

ACKNOWLEDGMENTS
KURRA is supported by Academic Center for Computing and Media Studies, and by Institution for Information Management and Communication. Especially following people help practically: Ayumu Saito, Yukiko Nishimura, Mihoko Hirazawa, Masaaki Iiyama, Masaji Iwakura, Naomi Nagata, Takaaki Aoki, Tamaki Motoki.
Research Lifecycle-based Research Data Management Requirements and Its Alignment with Institutional, Domestic and International Contexts

Shoji Kajita  
Academic Center for Computing and Media Studies, Kyoto University  
Yoshida Nihonmatsu, Sakyo-ku, Kyoto, Kyoto 606-8501 Japan  
kajiita@media.kyoto-u.ac.jp

Kazutsuna Yamaji  
Open Science Infrastructure Research Center, National Institute of Informatics  
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan  
yamaji@.nii.ac.jp

Heidi J. Imker  
University Library, University of Illinois at Urbana-Champaign  
1406 W Gregory Dr. Urbana, IL 61821 U.S.A.  
imker@illinois.edu

ABSTRACT

This session discusses Research Data Management (RDM) along research lifecycle by having inputs from University of Illinois at Urbana-Champaign in U.S.A. and Kyoto University in Japan as the representative from research intensive university and National Institute of Informatics in Japan as the representative from nationwide academic service provider.

In terms of the development of RDM infrastructure at each institution, we need to consider a lot of requirements and constraints to support diverse research communities at campus and to promote the use of RDM infrastructure. For example, the RDM plan requirement in Call for Proposal have been changing dramatically the necessity of the campus-wide RDM in campuses in U.S.A. and Europe countries. In special, it is very crucial for the research institutions in U.S.A. to support researchers by providing RDM infrastructure to get research funding in very competitive situation based on slightly different RDM plan required by funding agencies. In Japan, on the other hand, the academic scandal in the field of stem cell research has initiated requiring research data preservation for more than ten years after final publication for all of researchers, in the sense of research integrity. In addition to such domestic contexts, the international trends on Open Science and Open Data have been pushed by governmental and international organization to advance science and innovations. As the result, each nation must consider the nation-wide RDM infrastructure to support campuses and researchers because researchers work beyond institution and nation within the domestic and international contexts.

We will talk about the current status and challenges along research lifecycle in each context shown in Figure 1, and discuss the future of RDM to envision the RDM infrastructure in 2030 required in "R1: Research Universities in Carnegie Classification."

Figure 1: The big picture for aligning research data management requirements in the intersection of campus, domestic and international contexts
How well do we describe historical resources for digital archives?
Japan, Asia, and the World

Extended Abstract†

John Ertl
Institute of Liberal Arts and Science,
Kanazawa University
Japan
jjertl@staff.kanazawa-u.ac.jp

Yoshiko Shimadzu
Museum Science Division,
National Museum of Japanese History
Japan
yshimadzu@rekihaku.ac.jp

Shigeki Moro
Faculty of Letters,
Hanazono University
Japan
s-moro@hanazono.ac.jp

Makoto Goto (Moderator)
Research Department,
National Museum of Japanese History
Japan
m-goto@rekihaku.ac.jp

ABSTRACT

There are a variety of methodologies for preserving historical resources in Japan. Ultimately, however, whether the resources represent original or digital data, the main point to be addressed here is: “What and how can we preserve these resources for the future.” In other words, this session discusses the following: “What do we record?” and “What are the underlying issues in recording information from historical resources?” from the perspective of archaeology, scientific studies on cultural properties, and digitization.

KEYWORDS

Cultural resources, digital description, museum, archives

An issue of long-term preservation of cultural resources has a long history of reviews. Preservation of digital data in Japan also has similar issues, but they have not been resolved yet. This session discusses the standpoint for organizing these perspectives.

For instance, in terms of an archaeological site, the following preservation methods are generally adopted in Japan.

(1) Preserving an archaeological site in situ.
(2) Reconstructing an archaeological site to the condition when built.
(3) Recording information in an archaeological site, even though it is compelling to disrupt an archaeological site in land use.

The first (1) method of resource preservation is certainly preferable. However, a visit can cause damage or disruption to the site; or the site itself may be inaccessible due to inhospitable surroundings. As a result, the site and its resources may be lost without any access. The second (2) method may result in better access or increased frequency of visits to the site, but the accumulated history may be lost. Method (3) may not contain all of the site resources; while authentication of historical values is also important. The long-term preservation of ordinary resources requires confronting a dilemma between an ideal situation and the reality.

Digital data provides for a near-successful resolution of the above issues. Data can be maintained by different methods from ordinary resources of an archaeological site. A dilemma in applying methods (1) or (2) is least likely in digital data. The long-term preservation of ordinary resources and digital data can help showcase our richness of cultural resources in the future.

In contrast, we are faced with the problem of how to describe digital data for method (1). Indeed, the issue of what to describe is not only faced for digital data alone, but for all information recorded on paper. However, digital data is different from paper records in terms of the following features: (1) capacity of inputting data (2) flexibility of types of media. If only recorded on paper, we need to fit the information within an appropriate number of words, which cannot be exact. Due to this limitation, it was necessary to examine only the available important information for preserving the highly well-selected cultural resources. While digital data can resolve this limitation to some extent, it is still impossible to store all relevant cultural resource information digitally. In addition, digital data may suffer from the handicap of not having precise written descriptions of resources.

Flexibility of media can resolve the limitation of written records to some extent. Digital data enables recording of sound, motion picture, and 3D, which was not possible earlier. In addition, it can record enormous imaging data. However, representing content in 3D has a limitation, in that lost information cannot be identified when converting an ordinary resource to digital data. While a written record can alert people with the text “Some parts are out of preservation,” 3D data may misguide or misinform people who do
not have enough knowledge about digitization features, if alerted with the message “More complete copies are preserved,” more so, when such data is not available. In addition, there is another issue, “What can we preserve as metadata, along with 3D data?” Digitization of moving images brings forth a similar problem, “Which part of an image should be shot,” not faced in written records.

This session focuses on these issues. From the scope of problems of digitization, utilization, and reconstruction of archaeological sites, and the current conservation and preservation of cultural properties, three speakers will examine methods for effective long-term preservation of ordinary and digital records of cultural resources in the digital era.

**ACKNOWLEDGEMENTS**

This session is supported by the National Museum of Japanese History Institute-based Project, “Constructing Integrated Studies of Cultural and Research Resources, and Renovating Sharing Infrastructures of Research Resources in Japanese History and Culture.”
Japanese History Research by Historiographical Institute the University of Tokyo and its Contribution

Extended Abstract†

Taizo YAMADA
Historiographical Institute
The University of Tokyo
Japan
cert@hi.u-tokyo.ac.jp

ABSTRACT
Historical Institute the University of Tokyo (東京大学史料編纂所; HI for short) has as its primary objective, rather than historiography in general, analysis, compilation, and publication of historical source materials concerning Japan. HI has become a major center of Japanese historical research, and makes historical sources available through its library, publications, and recently, databases. In the poster, we introduce overview of our research and our contributions.

KEYWORDS
Japanese history, historical material, digitalization, pre-modern

1 INTRODUCTION
The research by HI consists of conducting research on, compiling, and publishing primary source materials dealing with Japanese history. The foundations of all historical understanding lie in the preservation and precise reading of primary source materials. Because it is a basic principle of HI that source materials should be preserved where they are found, we spend several weeks each year investigating and examining historical records wherever they are kept, both in Japan and abroad. We have over 100 years’ experience conducting these examinations, and our journeys have taken them all over Japan as well as to many different parts of the globe. When possible, documents are reproduced by several different means, such as microfilm, calligraphic reproduction, and photography. Over the decades, we have thus accumulated an unparalleled collection of duplicate documents as well as tens of thousands of originals. In some instances, documents are lent to us for further analysis, or for repair when damaged by water, insects, or the passage of time. We carefully catalog and index all source materials and list their findings in a variety of reference works. Specialized searches can be conducted based on individual research themes to facilitate advanced source criticism.

Along with assisting scholars by providing document catalog and index information, we are also eager to share the findings of their own research. We teach, both in the Graduate School of Humanities and Sociology at the University of Tokyo, and at other universities, and are active research scholars in their own right, participating in scholarly symposia, and publishing our research both in traditional print formats, and in digital formats via computers and the Internet.

2 COMPILATION AND PUBLICATION
We strive to make historical materials readily accessible to researchers in ways that they can easily use. Consequently, we divide its compilation and publication work into two broad sections. First, we compile chronologically organized document collections which incorporate materials from many different sources in the order in which events occurred. Principally, these include the Dai Nihon Shiryô series and the chronological sub-series of Dai Nihon Ishin Shiryô. Second, we publish compilations organized around specific archives. These include three series: Dai Nihon Komonjo, Dai Nihon Kokiroku, and Dai Nihon Kinsei Shiryô. To make these compilations widely available, HI has been publishing since 1901. Currently it produces over ten volumes each year, and the total number of Institute publications is over one thousand. Some principle publications are described follows.

Dai Nihon Shiryô ("大日本史料"; Chronological Source Books of Japanese History) The series is an ongoing and comprehensive attempt to cover all major events mentioned in the existing records between the years 887 and 1867. The starting date corresponds with the date when the ancient Japanese state abandoned compilation of official histories, and 1867 corresponds with the last year of Tokugawa rule. Publications of these historical sources are divided into sixteen parts or periods with multiple volumes in each part. HI has also published a 17-volume guide to Dai Nihon Shiryô entitled Shiryô Sôran. It lists all materials from 887 to 1639, including unpublished volumes of Dai Nihon Shiryô.

Dai Nihon Komonjo ("大日本古文書"; Old Documents of Japan) The publications are divided into three parts. The chronological part focuses on the Nara-period and the Shôsôin-Documents. The second part focuses on individual warrior family and temple archives. Bakumatsu Gaikoku Kankei Monjo focuses on key events of foreign relations in late Edo-period (1853-1867). Fig 1 shows a sample of the publication ("The Documents of Shimazu Family").
Dai Nihon Kokiroku ("大日本古記録"; Old Diaries of Japan)
Diaries remaining from the ninth to the nineteenth centuries are
invaluable sources for historical studies. The series are
publications of these diaries. The volumes are based not only on
the original text but on the references to later versions.
Explanatory notes are based on extensive comparative research.
Indexes are added for easy use.

Dai Nihon Kinsei Shiryô ("大日本近世史料"; Historical
Materials of the Edo Period) Basic Materials to understand the
Edo-period are selected for publication in this series. The
selections include documents from the Tokugawa shogunate, the
various clans (藩; han), and the Imperial court.

Dai Nihon Ishin Shiryô ("大日本維新史料"; Historical
Materials of the Meiji Restoration) This series originated in
another research institute located within the Ministry of Education
that focused exclusively on the Meiji Restoration. Copies of
materials were arranged chronologically in over 4,200 volumes
covering the years 1846 to 1871. Publication ceased with the
dissolution of HI in 1949. The Historiographical Institute was
asked to resume work on documents related to the Meiji
Restoration, and currently chronological materials are being
prepared directly onto microfilm. The special part of the
collection focuses on the Ii family document. Fig 2 shows a letter
written by Ii Naosuke (井伊直弼) as a sample of the publication.

Figure 1: The Documents of the Shimazu Family and Dai
nihon Komonjo

Figure 2: A letter written by Ii Naosuke (1815-1860), Great
Councillor of the Tokugawa bakufu

3 INNOVATION AND NEW DEVELOPMENT

To refine the research process and promote the availability of
our Institute materials and findings, our staff are constantly trying
new approaches and cooperative projects. It is essential that HI
continue to accumulate copies of historical source materials from
all over Japan and around the world in order to realize its mission.
As we in the discipline of history expand our vision of what
constitutes a "historical source," HI has kept pace by continually
adapting and refining its notion of the types of documents to be
collected, and the ways in which they are analyzed. Every year,
we undertake new research projects to focus on previously
unexplored themes and to employ new methodologies and
techniques.

Recently many of these research projects have become
international in scope. For example, the investigation of
documents concerning late Edo-period international relations has
grown from earlier studies, which focused exclusively on
diplomatic history, and now include military and commercial
documents. In addition, researchers have been looking at maps
from other parts of the world in order to conduct comparative
analysis of Japanese maps from similar periods. In these and other
ways, current research projects are moving beyond the goal of
document compilation and publication to a broader and more
complex understanding of our mission.

An area where the HI has been a leader in recent decades is the
development of computer applications for the management of
historical records and data. Research began in 1984, and quickly
progressed to include multiple Internet databases, computerized
records and search devices, and worldwide accessibility to certain
scanned images of original documents. Computers have become
an essential part of Institute research. To ensure that HI remains at
the forefront of technological applications, staff are active in
symposia and meetings to exchange findings with other scholars
and computer professionals. The Online Glossary of Japanese
Historical Terms and the Online Dictionary of Sources of
Classical Japan were established particularly for foreign users by
the Japan Memory Project (2000-2004). And, in 2006, on the
basis of such achievements, the International Center for the
Digitization of Premodern Japanese Sources was established,
which aims at new way of scholarly and international approaching
to historical materials in cooperation with information science.

4 INFORMATION SYSTEM

HI has been accumulating and applying research information,
including historical research information and historical document
information, through the continuous editing and compilation of
books. Many of these have now been computerized, and are made
available under SHIPS (Shiryohensan-jo Historical Information
Processing System) which includes our web site (URL www.hi-u-
tokyo.ac.jp) and our database system. SHIPS can accommodate
the demand for large volumes of information of edited documents.
SHIPS will be structured to provide a multi-channel connection
between the history researcher and historical documents, by
formatting various types of documents.
Digitisation and Digital Preservation in the National Archives of Japan

Extended Abstract

Yoshiyuki Kazama
National Archives of Japan
Tokyo, Japan
ykazama@archives.go.jp

ABSTRACT
The National Archives of Japan serves to properly preserve historical public records and archives, and promote public access. These holdings include not only paper-based records but born-digital records.

NAJ has proceeded with digitisation of paper-based records, and after the digitisation, it keeps both paper format as the original and the digitised imagery.

Based on the investigations done by various institutions such as the NAJ and the Cabinet Office, the NAJ ensures that preservation and access of records are properly handled, and also prepares for the long-term preservation and public access for the future generations.

CCS CONCEPTS
• Applied Computing → Computers in other Domains → Computing in government

KEYWORDS
Digital Preservation, Administrative records, digital records

1 INTRODUCTION
The National Archives of Japan (hereinafter, referred to as NAJ) is an organisation which serves to ensure proper preservation and the promotion for public access of the “official documents and other records that are important as historical materials” (hereinafter referred to as “historical public records and archives”) transferred from various state organs, and to conduct wide range of business from preservation practices to activities for public use. The Public Records and Archives Management Act (Act No. 66 of 2009) prescribes that historical public records and archives above mentioned are required to be permanently preserved as well as to be made available for public use. Also, the act requires NAJ and the facilities designated as possessing functions similar to NAJ (“the National Archives of Japan, etc.”) to accept not only paper-based records but digital records generated by electronic means, and to make sure they are preserved.

NAJ keeps paper-based records in the way in which they were acquired at that time. For some of those records it keeps their digitised images as well. Also, it accepts and preserves digital records. The information of our holdings is available to access via the National Archives Digital Archive, operated by NAJ. The Digital Archives is an online service which allows anybody to use from anywhere for free. It provides users catalogue information (metadata) of historical public records and archives and their digital imagery (limited to a part of the holdings) as well as a part of digital records.

2 Holdings and their digitised images

2.1 Holdings in NAJ and the digitisation
As of the end of fiscal year of 2016, NAJ keeps and manages over 1.4 million historical records and archives. Those include records transferred from not only current state organs, but those of the former government in the Japanese modern era, and pre-modern antique books and documents created and collected by the Edo Shogunate. These holdings differ in format and material, not just in content.

Up to this day, NAJ has digitised its holdings. As of the end of fiscal year of 2016, the number of digitised materials exceeded 210,000 volumes, and it reached 15% of entire holdings. NAJ plans to continue digitising the holdings in the fiscal year of 2017.

In terms of digitisation methods, NAJ adopts the JPEG 2000 (ISO1544-1) format which is free from any particular commercial software application, and is available to handle high compression and high resolution image data.

NAJ provides information of holdings via the online service called ‘the National Archives of Japan Digital Archive’ which users can access the catalogue information (metadata) and digital imagery of the holdings. The catalogue information in the Digital
Archives is handled in accordance with EAD (Encoded Archival Description). The format of digital imagery delivered by the Digital Archive includes JPEG 2000 as the downloadable format, and PDF and JPEG format as viewable in browsers.

2.2 Preservation of the digitised images
As mentioned above, NAJ adopts JPEG 2000 format when digitising, however digital imagery provided via the Digital Archive are lossy compression images in order to reduce the image size.

NAJ preserves the digital images of the materials separately from the images provided on the Digital Archive. The digital image are kept as the original and lossless compression image data of the holdings, and are saved to optical discs (blue ray discs) in accordance with the Japanese industrial standard for document management long-term preservation for electronic imaging documents in optical discs (JIS Z6017:2013). The digital imagery is kept in another information system at the Tsukuba Annex, a remote location. Tsukuba Annex is located in an area where the impact of powerful inland earthquakes is supposed to be mild compared to the location of the NAJ main office in Tokyo metropolitan area. The original data is stored there in case of a large scale disaster.

3 The digital records of the holdings
3.1 Acquiring the digital records
According to the survey conducted in 2016, the number of administrative documents possessed by state organs are 18,046,295 files as of 2015, and 996,157 files (5.5%) are digital records. The ratio of digital records versus entire administrative documents seems to be a little low. For reference, the Administrative Document File refers to a term in the Japanese record management, describing a unit of groups of documents closely related one another.

NAJ has accepted digital records transferred from the state organs and incorporated administrative agencies since 2011, and the total number of the Administrative Document Files which NAJ accepted has reached 1,015 as of the end of fiscal year of 2016. The 1,015 document files include approximately 26,000 digital files, and the file formats of Microsoft Word and PDF are often seen.

Amongst these circumstances, Japanese government currently promotes paperless business process as setting the goal for reformation of administrative information systems across the national and local level government. For instance, “Declaration on the Creation of the World’s Most Advanced IT Nation (Cabinet Decision on May 20, 2016)” encourages introducing remote work office environment such as telework at home. It is considered that further development of office environment using IT will require more management of storing and circulating records generated digitally. In addition, ‘Policy for the Promotion of Digital Government (Decision by the Strategic Council for the Promotion of the Use of Public Data by the Private Sector, the Strategic Headquarters for the Promotion of an Advanced Information and Telecommunications Network Society) was issued in May 2017, and the Japanese government will further engage with the promotion of e-government. While the majority of records transferred from state organs are currently paper-based, in the future it is expected that digital records will replace them.

3.2 Preservation/Service of the electronic records and documents
As for the preservation of digital records, the range of research and examination has been conducted, especially by the Cabinet Office and the NAJ. During the research and examination, based on the principle of preservation for digital records, various investigations surrounding empirical verification of the preservation systems were carried out. NAJ established the system for transferring, preserving and using the digital records from the result of the examination as well as the policies for acquiring and preserving these records issued by the Cabinet Office, and is currently operating a system to preserve electronic records and archives. The features of the system are as follows.

- Thorough quarantine
- File format conversion for assuring long-term readability
- Attachment of metadata considering the preservation and the future access
- Remote back-ups considering disaster recovery

In addition to that, NAJ is striving to ensure future access by implementing system migration every few years and preserving optical disc for backups compliant with the Japanese industrial standard for document management long-term preservation for electronic imaging documents in optical discs (JIS Z6017:2013).

4 CONCLUSIONS
The relevant laws prescribes that NAJ is required to permanently preserve historical public records and archives, and to make them available for public access.

While the majority of NAJ’s holdings created in the past are paper-based, the forms and the production methods vary such as bound and scrolls, even though they are the same medium.

Since then, more diverse media types have been used for historical archives and documents, and new methods of creating records have been invented, which in turn means we should consider preservation methods.

Furthermore, due to the advent of digital records, new perspectives and responses are required when it comes to assuring preservation and access.

For digital records (including digitised records) in particular, the progress and development of information technology not only encourages the environmental diversification of preservation and use, but the rapid technical obsolescence.

Paper-based records are humanly readable without any explicit instructions for reading. However digital records are readable once processed by computers. While users may be required to
properly process information, digital records are not integrally connected with the information of the system, and an environment for processing such information needs to be prepared.

In such situation, so as to support users, 2 methods as below are considered according to the migration of information systems for digital records.

- conversion of the digital records to the best format at that point in time
- emulation functions to regenerate the digital records created in the past

Whichever way is used, ensuring long-term preservation and future access requires considerable amount of expenses, and it cannot be said that it would be the slight burden for the public archives.

In Japan, there is a trend to newly establish the public archives among not only national and prefecture level, but municipalities. Under such circumstances, I consider that the issues of how to respond to the preservation and access of digital records should be the common challenges for all public archives in the future.

Regardless of the organisational scale I believe various technologies and technical/operational framework which enable digital records management will be pursued more in the future.

REFERENCES
Preserving Digital Motion Picture: the Present Situation and Issues for the Future

Extended Abstract

Katsuhisa Ozeki
National Film Center
The National Museum of Modern Art, Tokyo
Japan
oozeki-k@momat.go.jp

Naosuke Okamoto
National Film Center
The National Museum of Modern Art, Tokyo
Japan
okamoto-n@momat.go.jp

Kazuki Miura
National Film Center
The National Museum of Modern Art, Tokyo
Japan
miura-k@momat.go.jp

Tomonori Nakanishi
National Film Center
The National Museum of Modern Art, Tokyo
Japan
nakanishi-t@momat.go.jp

ABSTRACT

A big digital wave has come to the cinema industry and a huge amount of digital data is being produced every day. However, the technology to preserve such a huge amount of data for a long time does not yet exist. NFC started the BDC project to study this issue. The present situation and issues for the future concerning the digital preservation of motion picture are overviewed and discussed based on the research of the BDC project.

KEYWORDS
digital preservation, digital cinema, film, motion picture, community, production workflow, format, reproductivity, OAIS

1 INTRODUCTION

The National Film Center, The National Museum of Modern Art, Tokyo (NFC; see Fig. 1) is the only national institute for film preservation and research in Japan. NFC started a three-year governmental-level research project concerning various archival problems around digital cinema in 2014 (BDC project) under the big digital wave where the long-term preservation of “born-digital” or “made-digital” motion picture is an urgent issue [1,2]. In this report, needs, requirements and problems for the digital preservation of motion picture are overviewed and discussed based on the research performed in the BDC project.

2 THE DIGITAL SHIFT IN CINEMA

As a result of the big wave of digital technology, motion picture production has shifted from photographic film to digital. Furthermore, projection facilities have also become digital (the ratio of digital screen in Japan was 98% at the end of 2016 [3]). Eventually, photographic films will not be used from shooting to projection and it becomes harder and harder to see films by film projection. Accordingly, the demand to digitize motion pictures recorded on photographic films is increasing.

The technology of digital motion picture depends on information and communication technology (ICT) and new solutions related to the management and preservation of digital data are required. The workflow of motion picture production is continuously changing from the beginning of this digital shift because of the rapid and continuous evolution of imaging technology and/or ICT. As a result, motion pictures are produced and recorded by various formats and preserved in various sorts of media. Our research finds that management systems are individually established and different from each other among major studios in Hollywood. The situation is the same among Japanese production companies.

![Figure 1: National Film Center, The National Museum of Modern Art, Tokyo.](image-url)
3 CHALLENGES OF THE DIGITAL SHIFT

3.1 STANDARDIZATION
Digital cinema is presently in a transition period from the viewpoint of standardization while an unified standard has been established for a long time in the case of conventional motion picture. Standardization is one of the key factors for longterm preservation.

3.2 PRESERVATION OF FILM AND DIGITAL DATA
The difference between film preservation and digital preservation is shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1: Preservation of Film and Digital Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Media</td>
</tr>
<tr>
<td>Life Expectancy</td>
</tr>
<tr>
<td>Storage</td>
</tr>
<tr>
<td>Reproductivity</td>
</tr>
</tbody>
</table>

3.3 PRESERVATION PROBLEMS BASED ON THE MOTION PICTURE PRODUCTION WORKFLOW
Motion pictures are basically produced by the similar workflow both in film production and digital production. However, in the case of digital production, the work flow is complicated and there are so many options and various products could be produced easily. Further, digital technology makes it possible to digitize images recorded on film or restores them digitally, resulting in the increase of a variation of products. It is very important to decide what kind of data should be preserved, while taking cost into consideration.

As shown in Table 1, Digital data require software and metadata such as LUT (Look Up Table) and color space to reproduce images the same as what the creator intended. The increase of the variation of products and many kinds of metadata attached to the images makes it difficult to preserve all the digital products with their software and hardware which is changing rapidly. Hence, it would be required to select a specific product for preservation.

3.4 WHAT WE PRESERVE
NFC aims to preserve not only theatrical films but also private films produced on an individual basis including feature films, cultural and documentary films, news reels, animated films and so on. It is a distinct feature of cinema that many people including producers, directors, camera operators, imaging technology engineers apart from the casts are engaged in film production. Hence, the objective of “cinema preservation” is not easy to define because it depends on each person or institution related to the cinema. Cinema should be preserved not only as the physical media or image on the screen but also as the cultural and/or aesthetic contents with the accompanying background information.

4 FUTURE WORK
Some requirements for digital preservation mentioned above could be solved by the use of present ICT. On the other hand, serious problems such as cost, copyrights, responsibility for preservation and legal systems come to the surface. These problems are found to be hard to solve by individuals or by single institution. It is desirable to deal with such difficult problems by cooperation among the people who are related to the preservation of digital cinema. Meetings on the possibility of sustainable preservation systems including the utilization of digital content were held among the imaging companies, archives and imaging engineers in the BDC project. It provided a good opportunity to share the information and problems among the participants. At the meeting, the availability of the OAIS or cataloging model based on FRBR model was discussed. Further, preservation systems based on the above discussions were implemented and studied. Especially in Japan, the discussion concerning digital preservation is not sufficient. Establishing a cross-sectional community and encouraging continuous collaboration is needed on this issue.

REFERENCES
Management of Collection Information at Tokyo National Museum

Extended Abstract

Satoshi Tarashima

Tokyo National Museum
13-9, Ueno-Koen, Taito
Tokyo
Japan
stara@tnm.jp

ABSTRACT
Tokyo National Museum (TNM) is one of the biggest museums in Japan and has a history for over 140 years. This report shows present situation of management of digital information at TNM, creating database, digitizing images, constructing websites and offering digital resources for the public.

KEYWORDS
museum informatics, image database, open data

1 Introduction
Tokyo National Museum (TNM) was founded in 1872, as the first museum of Japan, after only five years since the new Emperor’s government had taken over political power from Shogun’s government. At first, the museum was expected to become an intellectual resource center for modernization and industrialization of the newly developed nation. Therefore, various artifacts concerning to natural history, industry, fine arts, and history not only from all parts of Japan, but also from many foreign countries by purchase or exchange. Through the history for more than 140 years, the main concept of the museum has gradually changed and now, TNM has a great collection, mainly focused on Japanese and Asian art history, including more than 130,000 titles on its inventory. This report introduces present works of information management by TNM and shows some expected future view.

2 Types of Digital Information in TNM

2.1 Collection Database
Building up the collection database of TNM began from early 1990’s, but it encountered various difficulties for a long time. Some problems were caused by complicated orthography of Japanese, especially related to Kanji, others were results of lack of standard metadata that describe properties of artifacts. It was around 2005 that the TNM collection database got proper usability for the inner staff of the museum who were responsible for curatorial work.

At present (2017), this database includes all titles of the collection and is used for everyday curatorial jobs, managing the process of acquisition of artifacts, planning the permanent exhibition, controlling loan, recording conservation works, and so on.

Figure 2: Screenshot of TNM collection database working on the internal network.

2.2 Digital Images
Digitizing images of the collection items at TNM began in 1994. As TNM had already stocked more than 50,000 color transparent and 150,000 black/white films, the first digitizing work was high resolution scanning of these films. This work continued until 2009, when the supply of sheet films was stopped and all photographic jobs were replaced with digital. On 2017, TNM has 100,000 scanned color images and 210,000 b/w images

Figure 1: Ukiyoe image of the art museum at 1st Domestic Industry Exposition held on 1877.
in stock, adding around 10,000 new color digital images every year.

Figure 3: scanned image of a transparent sheet film with 1000 dpi resolution

2.3 3-D Data

Taking 3-D digital data from museum assets was introduced in 2009. For 3-D scanning can get the whole shape of the object without touching it, it is very useful not only for an academic survey but also making replicas for museum exhibitions or editing VR contents.

In 2014, high-quality 3-D X-ray CT system was installed in TNM. With these new facilities, 3-D scanning has become much easier and can be done with higher resolution. TNM has already scanned many archaeological assets like an Egyptian mummy and ancient Buddhist statues.

3 Utilizing Digital Information

3.1 Web-Based Database

TNM has already released several web-based databases related to cultural properties since the Internet came into wide use in the end of 20th century. Sorry to say that almost of DB are still written in Japanese only, they are useful resources for academic studies of Japanese and Asian cultural properties.

“TNM Image Search” is one of the earliest trials among various DBs on the Web, and now it includes 110,000 images of around 20,000 artifacts among the TNM collection. Some other databases of research resource are found on “TNM Digital Research Archives”(http://webarchives.tnm.jp), All the contents within this site can be used freely for non-commercial purpose.

Figure 4: TNM Image Search http://webarchives.tnm.jp/ imgsearch/

3.2 Web Gallery

From 2010, National Institute for Cultural Heritage (NICH), the umbrella organization for TNM, has been maintaining digital image gallery of masterpieces from 4 National Museums’ collections of Japan, Tokyo, Kyoto, Nara, and Kyushu. Viewing high-resolution images, that can be easily magnified, of more than 1000 titles of designated National Treasures and Important Cultural Properties is available on the website (http://www.emuseum.jp). Free smartphone applications for iOS and Android are provided on each store and both are highly graded by many customers. Virtual reality moving image gallery based on 3-D data from Buddhist sculptures, prehistoric clay figures, and delicately designed lacquer wares also can be seen on a YouTube channel. This enables to see invisible parts of an item or to adjust conditions and colors of the surface.

Figure 3: 3-D VR Gallery of National Museum collection channel on YouTube

3.3 Making Data More Open

On March 2017, NICH has released “Integrated Collections Database of National Museums, Japan” (Colbase http://colbase.nich.go.jp). Almost all the titles in 4 Japanese national museums’ collection can be found from this site. Some of them are with English data and images. All the contents are offered under the Japanese government’s standard open-data license compatible with CC-BY 4.0.

In the newly planned national program for promotion of intellectual properties in 2017, Japanese government declares to establish a gateway to all kinds of digitized Japanese cultural resources in a few years. NICH and TNM are closely cooperating with this program and are going to offer more contents related to the museum collection to be widely used for academic, educational, artistic and also business use.
The National Digital Preservation Program for Scientific Literature in China

Extended Abstract†

Xiaolin Zhang, Zhenxin Wu, Honghu Fu
National Science Library, Chinese Academy of Sciences
33 Beishihuan Xilu, Zhongguancun
Beijing P.R.China ,10019
+86-(10)-82628382
zhangxl,wuxz,fuhhh@mla1as.ac.cn

ABSTRACT
With the support of the Ministry of Science and Technology of China, A cooperative national digital preservation program has been set up in China to preserve major databases of e-journals and e-books published inside and outside of China. With carefully formulated selection principles, rights management framework, open and trusted tech platforms, and standard-based auditing & certification arrangements, the system succeeded in preserving more than 13000 thousand e-journals from 10 publishers.

KEYWORDS
Digital Preservation, e-Journal, Scientific literature, National Program, China

1 INTRODUCTION
Test starting in 2007 by the National Science Library, Chinese Academy of Sciences (NSLC), and then formally transformed in 2013 into a National Program funded by the National Science and Technology Library Consortium (NSTL) under the Ministry of Science and Technology of China (MoST), the National Digital Preservation Program (NDPP) has achieved a large scale operational digital preservation system for a substantial portion of scientific literature subscribed by Chinese academic libraries.

NDPP is a cooperative system participated by more than 200 academic libraries and with multiple preserving nodes at major institutions, and is funded from a regular budget for nation-wide scientific information services by MoST as a national government commitment. The target resources are those important to research and innovation, with wide national usage, mainly in digital formats or relying on networked access, and especially those published outside China.

NDPP believes that the purpose to subscribe is to secure a long term access to knowledge manifested in e-literature, and immerse threats to accessibility of networked resources entail a natural right to preserve for those resources its member libraries subscribe. Thus NDPP requires that publishers recognize NDPP’s right to archive, right to process achieved content for the sake of preserving, right to provide access services under trigger events, and right to cooperate in preserving. Meanwhile, NDPP protects the copyrights and legitimate interests of authors and publishers, including rights of attribution, content integrity, content security, mutually defined trigger events, and auditing.

2 ROAD MAP
The predecessor of NDPP was a pilot digital preservation system by NSLC when the original selection principles, rights management framework, and model long term preservation agreements were tested, and ingest, archive, and dissemination workflows were built. SpringerLink journals and Chinese VIP Scientific Journal Database were preserved under these arrangement. Since NDPP formally came into being in 2013, it has gone through 3 main stages and now moving into the 4th phase under its plan for 2013 to 2020.

An improved demonstration system has been developed and 3 preservation centers were constructed. In the 4th phase, in addition to more resources being achieved, a national preservation registry and a coordinated operation center will be set up, and standard procedures for triggered access, auditing, certification, and financing will be established to make NDPP a trusted and sustainable repository.

Figure 1: Road map of National Digital Preservation Program in China.

3 RESULTS
There are currently three preservation nodes in NDPP, respectively, National Science Library (NSLC), Institute of
Scientific and Technical Information of China (ISTIC), and Beijing University. They mainly focus on e-journals and e-books from international publishers and society publishers in STM fields, with a few in social science.

Currently, NSLC has preserved 15 databases from 10 publishers, including e-journals, e-books and protocols. ISTIC has preserved more than 150 journals and signed a preservation agreement with Wanfangdata company (one of the three major Chinese aggregative providers for journals, proceedings, and theses and dissertations). Peking University signed agreement with ProQuest and Emerald Insight.

### Figure 2: A Co-operation Network Model.

To build the preservation network, we designed a concept model with 5 types of nodes. An A-type node would be a node only providing access-service. A P-type node would do the preservation works only. A B-type node is for backup. An F-type node would provide a full set of preservation services, including preserving objects and providing public services. It also can act as a backup node. A M-type node is a central management node, while other nodes joining the network through registering at M.

In practice, we followed a principle of node autonomy which means each preservation node can be a stand-alone F-type node to run independently. When registering to the M-node, it will become a part of the network which facilitates the mutually supporting remote back-ups and worst-case succession operation.

The tech system was based on Open Archival Information Systems (OAIS) model with function modules mapping to the OAIS requirements.

The preservation rights are secured by a special preservation agreement between a NDPP archiving institute and the publisher, specifying the target resource, scope of the content, ingest & archival processing requirements, and trigger events.

General auditing procedures are established so NSTL, participating libraries, and even the publishers, can periodically conduct audits to make sure what is meant to be preserved is actually preserved. And guidelines for certification of trusted archiving institutes are compiled following international standards and tests runs for auditing and certification were conducted.

### Figure 4: Archived Resources Statistics

### 4 CONCLUSIONS

NDPP, as a nation-wide effort and a long term governmental commitment, is determined to move forward to cover resources from other publishers. It will mobile the full force of the huge market it represents and increasingly important sharing of China in STM publishing to make sure that strategic knowledge resources will be secured for sustainable access under any circumstance. At the same time, it will improve its balanced and smart framework of fair rights, trusted processes, and transparent and standard-based management, so it will be relied upon by libraries, publishers, and scientific communities, to be their long term knowledge infrastructure in the time of many threats and uncertainties.

### ACKNOWLEDGMENTS

This work was partially supported by National Science and Technology Library.
A Review of Current State of Digital Preservation in the Academia Sinica Center for Digital Cultures

Abstract

Shu-Jiun Chen
Institute of History and Philology,
Academia Sinica
Taiwan
so phy@sinica.edu.tw

Hsiang-An Wang
Academia Sinica Center for Digital Cultures
Taiwan
sawang@gate.sinica.edu.tw

Wan-Jung Shu
Academia Sinica Center for Digital Cultures
Taiwan
sofia@gate.sinica.edu.tw

ABSTRACT

This report reviews the activities of digital preservation in Academia Sinica Center for Digital Cultures which maintains the Taiwan Digital Archives Union Catalog and more than 300 websites and databases across libraries, archives and museums. Three types of digital preservation strategies have been identified and a future outlook in the near-term is introduced.

KEYWORDS

digital preservation, redundant replication, emulation, migration, refreshment

1 INTRODUCTION

Academia Sinica Center for Digital Cultures (ASCDC) maintains a large of Taiwan’s Digital Cultural Heritage which includes Taiwan Digital Archives Union Catalogue, similar to the Europeana, a platform from the TELDAP (Taiwan e-Learning & Digital Archives Program) with over 5.6 million digitized materials collected mainly in Chinese languages over the past 15 years from more than 100 libraries, archives, museums, academic institutions and government agencies in the whole Taiwan [1]. In addition, the Center sustains over 300 websites and databases. Digital stewardship is one of the fundamental tasks of the ASCDC, to ensure the intellectual heritage can be used by future scholarly research, educational activities, and industrial applications [2]. The diversity of digitized materials and the frequent changes in information technology are the most challenge for the Center to plan the long term preservation strategy.

2 THE STRATEGIES OF THE DIGITAL PRESERVATION

2.1 Redundant Replication

Data redundancy has been the dominant approach to help preserve digital objects during the TEDAP period. The data types include static web pages, programs, databases, (multimedia) files, Virtual Machines (VM), and digitized materials. Among of the static web pages most come from the webpage catch of over 300 websites of the TEDAP. There are two backup strategies, one is distributed backup and the other is remote backup which has three copies are at least 30 kilometers from each other. The backup system maintains the updates daily or weekly depending different data types.

2.2 A Combination of Emulation and Migration

Emulation refers to an environment that imitates the behavior of a computer system. The key concept behind the approach is to be able to access or run original applications and data on a new platform by running software on the new platform that emulates the original platform [3]. The ASCDC adopts the emulation approach using a Virtual Machine that can be run in an interactive manner using current technology. In the meanwhile, the migration approach for the software and hardware is taken to avoid security vulnerability and ensure the technological obsolescence including older hardware and software can be transferred to a current or new platform.

2.3 A Combination of Refreshment and Migration

According to New World Encyclopedia [4], refreshment is “the transfer of data between two types of the same storage medium so there are no bitrate changes or alteration of data”. The approach has been adopted in ASCDC on demand to help transfer into a up-to-date or more appropriate storage media, for instance, response to the increasing use of mobile devices, the refreshing approach transfer the traditional webpage from HTML to HTML 5 in order to support the presentation for mobile devices. Another example includes converting file formats form Flash Video (FLV) to Mpeg-4 (MP4).

3 CONCLUSIONS

In summary, we have preliminary adopted a number of digital preservation strategies including redundant replication, emulation, migration and refreshment to ensure the digital collections can be reliably preserved. The ASCDC recognizes that a more comprehensive strategy including the vision and key activities should be fully developed to response challenges of technological obsolescence of the cultural heritage materials maintained in the Center. In the near future, an evaluation of the digital preservation strategies will be made, and the international standard for metadata to support the preservation of digital objects (i.e.
Preservation Metadata: Implementation Strategies, PREMIS) will be introduced and implemented in the Center.

REFERENCES


Indigenous and Rare Material Collection and Management at the National University of Laos

Sisavanh SINGVONGSA

Abstract—Paper describes collection management of indigenous and rare material resources in support of Academic Programme based on The National University of Laos (NUOL) experience as a case study. This paper includes descriptions of selected items in the indigenous materials collection, as well as information about institutional policies, obstacles faced in the effective management of the resources, and future plans. This paper discusses some of the steps the University Central Library (UCL) at NUOL has taken towards promoting the use of these materials and preserving indigenous knowledge for future generations. UCL recognizes the importance of utilizing modern technology to support the management of indigenous material resources, and disseminate Lao information globally.

I. INTRODUCTION

Since 2006, after participating in the previous meeting of the Local Information Network at Mahasarakham University, the Central Library of the National University of Laos has begun to prioritize the sharing of local information among South East Asian nations. A variety of local information resources at the Central Library have been identified and organized to respond to the needs of the academic staff, students, and other community members. Bibliographic records of these materials have been prepared and added to databases in the Central Library of NUOL.

With regards to the goals of the network, we; the librarians of the National University of Laos, are very motivated to promote mutual understanding and to find solutions to some of the more difficult problems encountered within the Southeast Asian Network members in sharing local information resources. Managing knowledge in general and indigenous knowledge in particular has become an important and valuable input in the management of sustainable development programmes (Ngulube 2002). We, at the National University of Laos, recognize the value and importance of preserving local information and local resources, as these materials are unique to the Lao PDR and are of great national and cultural value.

II. NATIONAL AND INSTITUTIONAL POLICIES

Regarding policy matters, preserving, sharing knowledge of culture including traditional and indigenous information has been accepted by the Ministry of Industry and Culture for decades and the National University of Laos, Ministry of Education for over ten years. Throughout its evolution NUOL has played a leading role in responding to the needs of the nation, stressing academic excellence and quality teaching, learning and researching. The National University of Laos is a center of published works including books, newspapers, journals and other printed documents available to students, teachers, and personnel of the University. Regarding the NUOLs policy, the University Central Library has been given the tasks of organizing, managing and maintaining academic works, scientific documents and all the other documents advertising its activities, distributing local and international information to the public, especially the students and teachers accessing information in the form of printed and non-printed materials.

III. LOCAL WISDOM RESOURCES DESCRIPTION

The University Central Library of NUOL has established the project of preserving the local information collection to serve the University academic programs and the community. The various local information or indigenous knowledge contained in such materials can be lost if they are not properly preserved. The goal of the local information project is the conservation of the Lao culture, historical heritage, research papers and other printed and non-printed materials related to social and economic history of Laos. However, to fulfill the above mentioned goal, funding is required for restoring and preserving the national heritage that exists in the country, as well as at the maintenance of national collections including research papers and documents of indigenous knowledge.

IV. MANAGEMENT OF INDIGENOUS AND RARE MATERIALS

The Library is a conventional system that all universities use as a mean to centralize information. With the emergence of new technologies everyday, a wider range of information becomes more and more available in various locations and formats. As the users demand to access more information on social, economic and cultural heritage is increasing rapidly, the demand to control and manage information efficiently is also growing rapidly. However, to manage indigenous materials for increased access, librarians need to know how to develop a strategy and process for deciding selection criteria, deciding upon the mission, goals and objectives of the project, raise financial support, and evaluating what to retain based upon the following aspects:

- value - usage - cost effectiveness - support of organizational mission, and quality of access.

As the local information is said to be the collective knowledge of ethnic groups of people of the country and institutions is a resource with value beyond measure. In Laos, there are 49

M. Shell is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).
J. Doe and J. Doe are with Anonymous University.
Manuscript received April 19, 2005; revised January 11, 2007.
officially recognized distinctly different ethnic groups, however many more sub-groups also exist (Vorakhoun 2008). This local information spans the fields of history, geography, law, demographics, natural resources, religions, and culture including literature and folktales, customs, clothing, ceremonies and the life stories of leaders and those who have inspired us. These forms of information should be safeguarded and readily available as reference source in a user friendly format and environment.

We have taken the following measures to effectively manage these local information materials: At first, these materials have been organized into the following categories:

**A. Lao Cultures**

- Vatanatham lae Heedkhong Paphenee an Dee Ngam khong Lao (The rich culture and traditions of the Lao Nation) Book describing the great variety and wealth of culture and traditions of the Lao Nation.

**B. Historical, Religious, and Ethnic Heritage**

- (Pavad Prabang Buddha) A book describing the history of Laos most celebrated Buddhist statue.

- Muong Lao Heun Nhao Paukado (Muong Lao Home) A book describing the home of one of the ethnic groups in Saravan Province in southern Laos.

**C. Fictions**

- Thao Khounlou Nang Oua The fictional love story of Thao Khounlou and Nang Oua.

**D. Fictions**

- Khatixon Phuen Muong Lao (Lao Folktales) A collection of Lao folktales and parables involving teaching the reader.

**E. Legends**

- Champa See Ton (The Four Frangipani Trees) The legend of the Lao National Flower, the Frangipani (Plumeria).

**F. Poetry**

- Thao Hung Thao Cheuang (Mr. Houng and Mr. Cheung) An epic poetic tale written in modern Lao prose.

**G. Proverbs**

- Kham Souphasit Phuen Muong Lao (Lao Proverbs) A collection of Lao proverbs.
H. Audio Visual Materials

Wat Phou Sy (Wat Phou Sy) A film about the famous Luang Prabang Temple Phou Sy.

V. Future Plans on Indigenous Material Collection Management

The next step is to fully integrate bibliographic records of all the materials contained in the Local Information collections into the Central Library's computer database. Once catalogue records have been created for all of these materials, we can look into further promoting and disseminating the local information. This can be accomplished through the establishment of a digital library, as well as through internal and external promotion of the collections and advocacy for local information resources. In the future, projects such as the digitization of these materials can improve not only our capacity to preserve local information, but also to promote and disseminate local information. We hope to soon have the possibility of using software such as DSpace to create a digital library.

VI. Obstacles

Indigenous knowledge presently is available in written and other media formats. However, some of them currently, are not available in print because they were published in limited editions, sold out and never reprinted. And due to the shortage of budget and of knowledge in preserving and conserving the local information or indigenous materials, some of these items were not kept in good conditions in the library.

Research has shown that the lack of management systems for indigenous knowledge perpetuates the low local content on the Web, retards buy-in from local communities into digital resources and inhibits digital skills development (Greylng 2007). This is especially true in the case of Lao PDR, where most of the internet content viewed by Lao people is of Thai origin, and very little content in Lao language is available.

At this time, the Central Library must first improve the functioning of the library's database and catalogue system before we can proceed with our goals of improving access to local knowledge resources. While the library staff certainly recognizes the value and significance of local knowledge resources, unfortunately we lack adequate financial resources to effectively promote or preserve these materials.

VII. Conclusion

Local information plays a significant role in maintaining our Lao cultural heritage. Nowadays, the great variety of local information available in different parts of the country increases the wealth of cultural resources available to scholars, and the general public. However, if proper attention is not paid to correctly documenting and preserving these local information resources, they may be lost forever. The government of Laos has recently encouraged the preservation of local information and heritage materials. This includes documentation of Lao architecture and the preservation of historic buildings and sites. The task of the library is to manage and maintain documentation about our cultural heritage. The emergence of new technologies, such as digitization, can facilitate the preservation and dissemination of this type of local information. In an era of globalization and international communication, the Lao culture must be preserved, and the Central Library is prepared to utilize this type of technology for such purposes as soon as possible, in an effort to safe-guard our national identity for future generations.

REFERENCES


Three Digital Archives Projects in Thailand

Extended Abstract

Neelawat Intaraksa
Punsarn Asia
NSTDA, Pathumthani
Thailand
neelawat@gmail.com

Pongtawat Chippimolchai
Punsarn Asia
NSTDA, Pathumthani
Thailand
pongtawat.c@gmail.com

Vilas Wuwongse
Sakon Nakhon Rajabhat University
Muang District, Sakon Nakhon
Thailand
vilasw@gmail.com

ABSTRACT
Although there have a number of digital libraries developed and deployed in Thailand, there are only a few digital archives in the country. Three on-going digital archives projects in Thailand that the authors are involved are described.

KEYWORDS
Access to Memory (AtoM), digital archive, ISAD(G) metadata schema

1 NDMI DIGITAL ARCHIVES

NDMI Digital Archives is the repository for historic and memorable records of National Discovery Museum Institute (NDMI). NDMI is an organization, established in 2004 under the Office of Knowledge Management and Development Public Organization (OKMD) to help set up “Museum Siam”, a prototype of discovery museum in Thailand. Apart from setting up a modern museum, NDMI takes part in promoting, developing, and enhancing standards for museum management, academic research and museum network collaboration.

NDMI Digital Archiving project was initiated in 2015 in order to serve as the institution’s corporate memory for 10-year work achievements. Original master files of documents identified as valuable records were gradually accumulated whereas some printed materials such as large exhibition design plans and research documents were sent to a professional scanning company to be digitized.

In 2016, NDMI digital archival repository was implemented using a web-based, open source Access to Memory (AtoM) archiving application. The departments and units under NDMI’s organizational structure were respectively mapped into fonds and subfonds. Born-digital and digitized items such as meeting reports, research documents, booklets, pamphlets, and photographs of events and activities have been gradually added to the system by NDMI staff members. All archival records were described using ISAD(G) metadata. The metadata descriptions are open and available to all while some original files are restricted for internal use only.

Link: https://archives.museumsiam.org
2 NIDA DIGITAL ARCHIVES

NIDA Archives is an institutional archiving repository of the National Institute of Development Administration (NIDA), a leading institute of higher education in Thailand. Established in 1966 following His Majesty King Bhumibol Adulyadej’s vision of advancing Thailand’s development through advanced education, NIDA has now been expanded to support regional and international development apart from national mission.

In order to preserve memorable history of the institution, NIDA Archival unit was formed in 2013 under the Library and Information Center. The main duties are to acquire, organize, preserve, and disseminate the institute’s historic and valuable archival records to both interested individuals and public.

In 2014, NIDA Digital Archiving system was implemented with the aim to facilitate the management of the institution’s records and to provide online accessibility to public. Parts of archival records are digitized and organized according to the institution’s organizational structure. The filing consists of both textual and visual archival records such as administrative documents, institution policies, orders, research documents, photographs of events and activities, video records, etc.

3 PRASIDH SILAPABANGLENG’S ARCHIVES (https://archives.silapabanleng.hylib.org)

Prasidh Silapabangleng’s Archives is a personal archiving repository of Prasidh Silapabanleng (1912-1999), a Thai national artist of performing art in music composition. Born as a son of Luang Praditphairoh, one of the most renowned Thai classical music masters and composers in Thailand, he was trained to be a musician by his own father. Prasidh started his music study in 1934, at the Imperial Academy of Music, Geidai University, Tokyo. He then became the first Thai musician who has been graduated in music composition from a foreign university and the prominent artist who combined Thai musical composition with the western orchestra. Prasidh passed away on September 4, 1999 after suffering from Pneumonia at the age of 87.

In recognition of his father, Dr. Kulthorn Silapabanleng, the elder son of Prasidh, has initiated Prasidh Silapabangleng’s digital archiving project in 2016 with the aims to preserve his family history and to share his father's music compositions and plays as a cultural heritage to the nation. A small working group was setup with the collaboration of an archaeologist and librarians.

In 2017, a digital archiving system has been developed using AtoM (Access to Memory), an open source archival description software. Most of archival items, including manuscripts, letters, photographs, songs, plays, music notes and scores, are being digitized. All content cataloging is carried out based on ISAD(G) metadata schema in order to conform with ICA standard. This project was planned to be finished in the middle of 2018 with all contents made freely available to the public.

Link: https://archives.nida.ac.th
Enriching Museum Experience by a KKU Virtual Museum

Wanida Kanarkard  
Khon Kaen University  
Dept. of Computer Engineering  
Khon Kaen, 40002, Thailand  
+66 43 362 160  
wanida@ku.ac.th

Wanchat Boonmee  
Khon Kaen University  
Khon Kaen University Library  
Khon Kaen, 40002, Thailand  
+66 43 250 251  
bwanch@ku.ac.th

Yodsaya Sitthivong  
Khon Kaen University  
Khon Kaen University Library  
Khon Kaen, 40002, Thailand  
+66 43 250 251  
yodsayada@ku.ac.th

ABSTRACT
Museums are a place of heritage, a place to showcase the past, but that does not mean they cannot be interactive and futuristic. This paper shows the success story of our innovative accessibility enhancements by a KKU Virtual Museum that enable our customers to discover more about University History stories as they walk around the university campus. With the use of Beacon technology, our library can gain insights into where our customers went during their visit, trigger hyperlocal content and interactive geolocated notifications on their mobile devices.

CCS Concepts
- Information systems  Information systems applications  Digital libraries and archives.

Keywords
Virtual Museum, Internet of Things (IoT), Bluetooth Low Energy, Context Awareness.

1. INTRODUCTION
KKU Library has launched the interactive KKU history stories as a KKU Virtual Museum mobile apps [6] with Beacons, a low energy Bluetooth enabled IoT device for location and context based messaging and providing analytics in March 2016 (See Figure 1). This interactive mobile app aimed to create a solution that is both engaging and educational for customers. It has shown the benefit to put the advances in digital preservation technology in perspective and to recognise that the University history stories could be preserved for the benefit of future generations.

Figure 1. KKU Virtual Museum using BLE Technology.

2. RELATED WORK
For the museums, a Virtual Museum app is a wonderful tool that can add interactive content to the mobile [3]. It also invites customers to step inside the history content by themselves. Several museums have been utilised IOT sensing app. For instance, Wales’ National Slate Museum in Snowdonia was the first national museum in the world [8]. The Groninger Museum is the first museum in the Netherlands where beacons have been implemented [5]. New Museum in New York City hosted an exhibit that used Beacons to simulate a virtual minefield and let anyone experience the danger of landmines [4].

3. KKU Virtual Museum Mobile Application
Bluetooth Low Energy or BLE beacons are small wireless sensor devices that periodically broadcast a radio signal using Bluetooth Low Energy in which very low power usage to preserve device battery life[1]. Smartphones are able to interact with a beacon using the phone’s notifications feature [2]. The potential of beacons to add context to our virtual museum mobile apps, providing customers with navigation reference points and highly geo-targeted offers based on where the customer is at a given time in KKU campus illustrated in Figure 2.

Figure 2. System architecture

By using Beacons, we can provide fine grained customer location identification and tracking. We use this data to produce both location beacons on a live dynamic gallery map, but crucially have tied it to the location metadata from the KKU Museum System (See Figure 3) to produce collections search support and overlay to the map. We are currently indexing our university geocoded local history records into this IoT enabled mobile applications. The activities within the application started with a sweet look back at the story of the university and memories of alumni 50 years ago including historical new student inauguration ceremonies and new
student train ceremonies, dormitories for students inside the university, central cheerleading events, and the happiness from the old stories of the university. This story of the university tells about the background of important landmarks so that customers learn about their own background and creates pride in the institution.

Figure 3. KKU history stories within KKU Campus.

4. RESULTS
Beacons can be used for gathering data to gain insights on customer behaviour and use it to improve our virtual museum service and enrich the overall customer experience, illustrated in Figure 4. The age and gender report shows that 18-24 and 25-34 year olds together make up the majority of our customers, but the 25-34 segment contributes the most interest group. 54.15% are male customers. Since March 2016, this virtual museum app has been used by 658 customers with 3,825 sessions.

(a) Demographic Report

(b) Customer Engagement Report

(c) Customer Behaviour Flow

Figure 4. KKU Virtual Museum App Analytic Reports.

The customer behaviour flow report helps us to characterise the customer’s interactive relationship with our virtual environments, that show the contents that have the most popular exhibit by measuring the time spent by customers at different sections of our virtual museum contents on the individuals’ consumption experiences and also on the overall performance of the virtual app.

5. CONCLUSION
A KKU Virtual Museum Application helps the library to enhance our customer experience in accessing our university history stories that is highly relevant, convenient, and delivered in a timely and seamless manner using a technology that addresses the need for a low-cost, easy-to-implement solution for outdoor location-based services.

6. REFERENCES
## Authors' background

<table>
<thead>
<tr>
<th>Your Name</th>
<th>Title</th>
<th>Research Field</th>
<th>Personal website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wanida Kanarkard</td>
<td>Associate Professor Dr. and Director of KKU Library</td>
<td>Computational Intelligence, Data Mining, Big Data Analytics</td>
<td></td>
</tr>
<tr>
<td>Wanchat Boonmee</td>
<td>KKU Librarian</td>
<td>Digital Preservation</td>
<td></td>
</tr>
<tr>
<td>Yodsaya Sithivong</td>
<td>Assistant Director of KKU Library</td>
<td>Digital Preservation</td>
<td></td>
</tr>
</tbody>
</table>
DIGITAL HUMANITIES RESEARCH GROUP

Department of Information Science,
Faculty of Humanities and Social Sciences
Khon Kaen University

Kultida Tuamsuk
Khon Kaen University
Dept. of Information Science
Khon Kaen 40002, Thailand
+66 43 202 861
kultua@kku.ac.th

Kanyarat Kwiecien
Khon Kaen University
Dept. of Information Science
Khon Kaen 40002, Thailand
+66 43 202 861
kand@kku.ac.th

BACKGROUND

Digital Humanities research is currently recognized as an important issue at a global level. It has been stipulated as the research agenda of many countries, e.g., England, the United States, Australia, Japan, and Bulgaria. In these countries, the governments see the necessity to compile, maintain, and utilize humanity knowledge and cultural by means of high-competent information technology and communication that streamline management of information in the form of digital formats not only easy to obtain access, but also simple to learn and create meaningful dimension or compare with other information and more. As such, critical and creative thinking skills can be developed. The listed countries set aside tremendous amount of monetary support for research and for establishing research center or research lab, or research group in digital humanities in the world’s leading universities.

The digital humanities work of these countries are generally under the responsibility of their i-School or university’s library as can be seen in the following example: Digital Humanities Research group, King’s College of London; Center for Digital Humanities, University College London; Center for Digital Humanities, University of California, Los Angeles; Center for Digital Humanities Research, Australian National University; and Digital Humanities Center for Japanese Arts and Cultures, Ritsumeikan University, Japan. Besides, at the international level, digital humanities associations have been founded as a venue for exchange of knowledge and building digital humanities researchers’ network.

OBJECTIVES

The KKU-DHRG has as its objectives to:

1) Collect, analyze, synthesize, store, and systematize knowledge contents in humanities which exist in multiple forms. Most of the knowledge is rare, reflects histories, cultures and ways of living. Present storage risks

destruction and loss, both from man’s naivety and natural disasters.

2) Use high technology to manage knowledge contents in the digital format that will provide distance-, time, and place-unlimited access for all who will in due course enjoy learning through new multiple dimensions.

3) Develop multi-disciplinary research work among researchers and academics in the fields of humanities, and information technology and communication so as to create higher innovative research work in humanities.

4) Produce work outcomes and reputation in digital humanities research work at the national and worldwide levels.

FRAME WORK OF DIGITAL HUMANITIES RESEARCH

Diagram: Digital Humanities Research Framework


These associations also take responsibility in organizing many international conferences each year. Following are examples of these associations: Alliance of Digital Humanities Organizations (ADHO); The Canadian Society for Digital Humanities; The European Association for Digital Humanities (EADH); Australian Association for Digital Humanities (aaDH); and Japanese Association for Digital Humanities (JADH).

Faculty members and students have a chance to conduct quite a number of research studies in digital humanities, present their work at international conferences and publish their articles in international journals under SCOPUS base. The head of research group is also a member of the founding committee for CiSAP – Consortium of i-Schools in Asia-Pacific, and a program cochair of the ICADL 2014 – The 16th International Conference on Asia-Pacific Digital Libraries, in which the main topic is Digital Humanities. It can be seen that the faculty members of the Doctoral degree program in information studies are highly competent in development of KKU-DHRG up and in producing world renowned work outcomes.
Semantic Retrieval of I-san Proverbs (Pha-Ya) for digital Preservation

Kanyarat Kwicen
Khon Kaen University
Dept. of Information Science
Khon Kaen 40002, Thailand
+66 43 202 861
kandad@kku.ac.th

Krongkaew Kingsawat
Khon Kaen University
Personnel Division, Office of the President KRU, 40002 Thailand
+66 43 205 672
pun.kaew@gmail.com

Sompech Junlabuddee
Khon Kaen University
Dept. of Information Science
Khon Kaen 40002, Thailand
+66 43 202 861
jsompe@kku.ac.th

ABSTRACT
Pha-ya is a local wisdom and is a tacit knowledge using a dialect language, it is rarely found in formal information resources, is normally unrecorded, and now is gradually disappearing. This study, therefore, aims to use a theory of digital preservation to collect, preserve, disseminate and retrieve the knowledge on Pha-ya by using semantic technology. This paper revealed the scope and knowledge organization of Pha-ya can be divided into 6 classes, 15 sub-classes, and 39 divisions. The next step of this research is to develop the ontology of Pha-Ya for semantic retrieval system.

KEYWORDS
Knowledge organization, Thai Dialect, I-san proverb, Local verbal literature

1 INTRODUCTION
Language is one form of the nation’s culture. It is used for communication and it provides a tool for its people to convey their thoughts. Through proverbs and aphorisms, language can reveal the nation’s root and wisdom. Similar to language, dialects are part of the nation’s culture. They represent their local people’s cultural identity. Dialects give pictures of local people’s living conditions, ways of life, and beliefs. Dialects, however, have been influenced by the advances of modern communication. A greater number of people prefer to use the official language or national language, which is more prestigious. It can be said that local dialects are endangered of being lost, and currently, they are preserved and used by only some groups of people [1]. In conclusion, proverbs are important because they are cultural transmission that symbolizes social morality, customs, manners, and thoughts of people passed down from generation to generation. Apart from that, local proverbs help represent their people’s ethnic, identity, and life philosophy as well morality and local values.[2]

In the Northeast or the I-san region of Thailand, there are a number of local proverbs, or Pha-ya, which represent the I-san culture. “Pha-ya” is a folk wisdom demonstrating how I-san people use their language for verbal communication and compose poetry. Pha-ya means "philosophy" through which different aspects of the villagers’ thoughts, intellect, and wisdoms are vibrantly seen. There are many forms of Pha-ya: rhymes, short or long phrases, or sentences. All of them involve almost all aspects of Isan people’s lives, whether it is sin, merit, manners, gratitude, duties of membership in society, living one’s life, working, and so on. Through Pha-ya, the simple life of the I-san community that likes to be free and to have fun is shown. This makes the I-san society different from other societies in Thailand [3].

Pha-ya is diverse in terms of content. As mentioned, Pha-ya can be used to teach or deliver useful knowledge, which is up to date. This knowledge can be useful to people in other areas, not limited only to the I-san region. However, the number of those who have the knowledge of Pha-ya and can create it is diminishing. As such, a study adopting a digital humanities approach to develop semantic retrievals is needed. This approach can help increase the effectiveness of searches and reduce the limitations in regards to its dialectal form. This will help academics as well as those who are interested in Pha-ya to access the knowledge embedded in Pha-ya with no dialectal limitations.

Language preservation is the effort to prevent languages from becoming unknown. Russell mentioned that Information Technology can help preserve the vanishing native language[4]. Through this concept, the local wisdom of Pha-ya, can be systematically collected and searched. This in turn will help preserve Pha-ya to prevent lost as well as to spread the knowledge to other regions. To create semantic searches that are effective and meet the user’s requirements, the knowledge scope of Pha-ya must be examined. Such knowledge will also be used to design and develop an ontology on Pha-ya and a semantic search system. This paper, therefore, aims to present the structural knowledge of Pha-ya, which will increase access to this knowledge and reduce lexical barriers (e.g., words with different forms or word variations). With higher access to the knowledge and local wisdom, the user will able to use this knowledge more efficiently.
2. RELATED WORK

Based on the literature review, it was found that Thai researchers were interested in collecting and publicizing the knowledge on Pha-ya in books and review articles. Also, a number of research studies as well as master degree’s theses and PhD dissertations have been conducted to study Pha-ya to understand how it is used as well as to preserve it because Pha-ya symbolizes the Thai local wisdom and I-san community’s identity [3, 5-10]. Also, other studies examined the linguistic structure of Pha-ya [11] compared its meanings with Thai proverbs[12], and interpreted the meaning of Pha-ya by adopting religious and philosophical methods [13-14], and a cultural anthropological approach[15].

3. RESEARCH METHODOLOGY

This study is a qualitative research study conducted to examine the meaning of Pha-ya (I-san proverbs) by adopting a content analysis technique. The scope of knowledge of Pha-Ya was collected from the information resources such as books, research report, theses and dissertations, and related website. 373 Pha-ya were collected and analyzed in this research.

4. RESULTS

The knowledge of Pha-Ya were classified by 3 steps. The first is information analysis and extraction. The elements and contents of Pha-Ya were analyzed in order to find correlations and then extracted to identify the knowledge representations based on concepts and terms. The following step is classification of data. We analyzed the knowledge system and reclassified the knowledge of Pha-Ya by placing similar knowledge and related knowledge close to one another according to knowledge management principles. Then, knowledge structure of Pha-ya was developed. Based on the knowledge structure, Pha-Ya was further divided into six basic classes as show in Table 1; each class has many different subclasses. Each subclass includes a loosely hierarchical arrangement of the topics relevant to the subclass, going from the general to the more specific ones.

Table 1. The scope of knowledge classification on Pha-ya

<table>
<thead>
<tr>
<th>Class</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-san wisdom</td>
<td>The knowledge and wisdom of I-san region, including health care, Thai herb and education system</td>
</tr>
<tr>
<td>Buddha and I-san beliefs</td>
<td>The Buddhism and beliefs of I-san regions, including what is concerned with Buddhism that influence their subsistence</td>
</tr>
<tr>
<td>I-san subsistence</td>
<td>The I-san’s subsistence, which includes occupation, food</td>
</tr>
<tr>
<td>I-san culture</td>
<td>The culture and customs of I-san region, including family system,</td>
</tr>
<tr>
<td>I-san Social norms</td>
<td>The rules and expectation of behaviors of I-san region</td>
</tr>
<tr>
<td>I-san value</td>
<td>The value system of I-san region, including seniority respect, women’s value, generous and empathy</td>
</tr>
</tbody>
</table>
Palm Leaf Manuscript Preservation in Asia: The Metadata Standards

Nisachol Chamnongsri
Suranaree University of Technology, Thailand
nisachol@sut.ac.th

Lampang Manmart
Khon Kaen University, Thailand
lamman@kku.ac.th

ABSTRACT
This paper aims to provide the current state of PLMs management in Asia, current use of PLMs metadata schema in working projects for the long-term preservation of PLMs, the challenge and constructive solutions, and the first draft of PLMs metadata core elements from the IFLA PLMs preservation workshop at National Library and Documentation Services Board of Sri Lanka held by IFLA on September 6-7, 2017 are presented in this paper.

KEYWORDS
Palm leaf manuscript, PLMs Preservation, PLMs Metadata,

1 PLMS CHARACTERISTICS; VALUES; PROBLEMS OF MANAGEMENT, ACCESS AND USE; DIGITIZATION AND METADATA
In the South Asia and South East Asia, Palm Leaf Manuscripts (PLMs) are the ancient document form that comprises a significant documentary heritage of the people in these areas. These manuscripts contain a vast amount of knowledge such as Buddhism, Tradition and Beliefs, Customary Law, Traditional Medicine, Astrology, History, Folktales, etc. PLMs vary in size and styles, and each country has their unique characteristic. For example, in Thailand a standard PLMs are generally 5-6 cm. in width and 50-60 cm. in length with 48 pages per fascicle (24 leaves written on both sides). PLMs can be as short as 15 cm. or as long as 80 cm. and can vary according to the number of pages. The ancient people used different sizes in different ways: for example, in Thailand the longer PLMs were used as a textbook to record Buddhist stories and doctrine, while the shorter ones were used as a notebook to record local wisdom related to daily life. The languages in which PLMs are written are both local and undergoing shift (Balinese, Thai, Lao, Khmer, Sanskrit and etc.), and the manuscripts are written in archaic orthographies, requiring expert translation. Because the length of PLMs is determined by its physical dimensions rather than its content, a single manuscript may record many stories, or a single story may require more than one manuscript. PLMs in the past were knowledge resource with recordings of stories and knowledge that had been passed on by the local wise men.

Currently, many PLMs preservation projects in South Asia and South-East Asia have attempted to collect PLMs which are scattered in many monasteries and have attempted to digitize the PLMs in their collections to make it easier for users’ access and use and for project staff to translate and preserve the original PLMs. India probably the largest collection of manuscripts in the world, The Ministry of Tourism and Culture, Government of India are established The National Mission for Manuscripts (NMM) in 2003 to preserve the vast manuscript wealth of India. They possess an estimate of ten million manuscripts, probably the largest collection in the world. The Mission has the mandate of identifying, documenting, conserving and making accessible the manuscript heritage of India. The mission has developed metadata standard and digitization standard for palm leaf manuscripts, and make it available on their website for public access. The participants from India said that their project have tried to collect the PLMs and then conserve the physical condition of the PLMs before microfilming for conservation purpose. To support user access, they digitize the PLMs and use open soft software, DSpace as an online database. National library of Laos also has a huge PLMs collection. They microfilm PLMs in their collection and then digitize the microfilm and develop online database to allow easy access from anywhere and at any time. In Indonesia, the National library of Indonesia has assigned many places in Indonesia to hold the collection. However, PLMs is a small collection, therefore, the national library of Indonesia put the PLMs metadata in Library OPAC. In Sri Lanka, the national library has a PLMs conservation and digitization project, but currently they have not set up any standard for the PLMs. In Thailand, there are two main groups working on PLMs preservation: universities and national libraries, where PLMs are both digitized and microfilmed. However, there are only 2 projects that develop online databases and open to public access.

2 CURRENT METADATA SCHEMAS USE TO DESCRIBE PLMS
The literature review revealed that there is no metadata standard to describe palm leaf manuscripts. Each PLMs preservation project has tried to develop its own description schema to elaborate the PLMs in their collection. Nevertheless, different projects created metadata in different ways and for different purposes.

In April 2017, we conducted a survey to investigate the current metadata or description schemas which PLMs preservation projects use to describe PLMs in their collection with these queries: Are there any standards readily and more frequently used than others and what is the important information present in PLMs bibliographic records? There are 16 projects included in this study; 10 projects from Thailand, 2 projects from British library, 1 project...
from Cornell University library, 1 project organized by Northern Illinois University library, 1 project from National Library of Laos, and the last one is NMM from India. Nevertheless, except Thailand and Laos, we could not find any PLMs databases from countries in South-East Asia and South Asia available on the internet, even if we knew that there is a vast amount of PLMs in these regions and they also established many of PLMs conservation projects. According to the 16 projects, we found that there is no standard used more frequently than others.

1. The National Library of Thailand schema, use only in the projects of National Library of Thailand.
2. The National Library of Laos schema; use by National Library of Laos; Lanna (Northern Thailand) PLMs preservation project, Chiang Mai University and Isan (Northeastern Thailand) PLMs preservation project, Mahasarakham University.
3. NMM metadata schema, use in their own projects.
4. KKUPLMMs, which created in 2009. The development of KKUPLMMs based on the study of requirement from PLMs users, PLMs preservation projects in Thailand, and the characteristics of PLMs in Thailand. Then, applying IFLA FRBR model to extract PLMs metadata elements. KKUPLMMs can support IFLA 4 user tasks in searching for information and administration function. The schema consisted of 76 properties (34 core elements and 42 element refinements) to describe all versions and formats of the PLM. After a few years’ implementation, it was revised in 2012 and 2015.
5. MARC21, a library standard, found only Northern Illinois University (NIU) Library use it to describe PLMs.

There are 52 elements found from 16 projects, can be separated into three groups; (1) physical characteristics is the biggest group with 20 elements, (2) bibliographic information with 19 elements indicating the content, and (3) administrative information with 13 elements. The highest number of elements in one project is 30 elements, the second is 25 elements, and the third one is 19 elements. The smallest number of elements in one project is 7 elements. The biggest group is 16-19 elements, there are 8 projects in this group and all of them are the active projects, have been working on PLMs conservation for 10 years. When look at the elements, it can have said that they have applied from the National Library of Laos schema and KKUPLMMs.

Even if staff of each project said that their users are happy to search the current online PLMs databases. However, all participants agree that creating the PLMs metadata standard is good for information sharing or interoperability. However, moving large amounts of PLM records to the new system is not easy, and it is time consuming. Conversely, the participant from new projects and upcoming projects would prefer to use the standard.

3 THE CONSENSUS WORKSHOP

The two days’ workshop at the National Library of Sri Lanka held by IFLA was arranged on September 6-7, 2017. The participant from PLMs preservation project in India, Sri Lanka, Indonesia, Laos, Thailand and Germany strongly agreed to jointly develop metadata standard for PLMs management. At the beginning, the participants suggested creating core metadata elements framework which can support two main functions of PLMs management: preservation and access. However, the important data for setting metadata requirement should come from the involved people’s behavior; these people include PLMs users such as historian, researchers in various fields, academics, graduate student and people who work with PLMs collection such as librarians, information scientists, curators and experts in ancient languages, etc. In order to initiate metadata standard by creating the metadata information framework, the participants decided to use metadata elements which were analyzed from the elements shown in metadata schemes used in 16 projects, as a guideline for selecting the core elements of metadata for PLMs management. The principle of selecting metadata elements based on PLMs management functions are: preservation and access. After discussion and consideration, the participation group created the first draft of metadata core elements which contains 39 elements divided into 3 groups by functions; 15 elements for supporting PLMs access; 15 elements for management; and 9 elements for supporting collection administration. The 39 elements provide definitions and explanation. The set of metadata elements is an initiated metadata framework for PLMs management: preservation and access. It is not a complete scheme, but it is the beginning for continuing development.

4 CONCLUSIONS AND SUGGESTIONS

1. PLMs preservation projects in the region should set up the regional working group; then arrange workshops, or seminars to discuss and develop metadata schema and other standards for the PLMs, both the original and reproduced versions by defining metadata standard continuing from the initiated metadata framework from the workshop. The standard should describe the PLMs in content, context, and physical characteristics. (2) The experts from each country should reconsider and try to implement the setting metadata core elements in their own PLMs management context, and consequently, making feedback problems and suggestion to the group. This information will be very useful for developing the metadata standard for PLMs management which is flexible for application in each country. (3) The PLMs users’ behavior in each country should be investigated to obtain the data for setting basic requirements of PLMs metadata. (4) Research should be conducted based on the research framework from this workshop or the element set we are going to set up, then PLMs metadata application profile and PLMs metadata registry should be established. Therefore, we can share and reuse PLMs information based on the same standard. (5) Development of PLMs metadata schema should be promoted to describe the PLMs. Therefore, PLMs can be accessed across the region and can be shared and reused including: all metadata, content, and full PLMs. Then, researchers, students and people who are interested in PLMs in the region and from other parts of the world can access, use, and share their memories, knowledge, and cultures; and to preserve the PLMs for the future.

ACKNOWLEDGMENTS

International Federation of Library Association and Institutions (IFLA), Strategic Programme on Preservation and Conservation (PAC)
ABSTRACT
The cultural heritage of northern Thailand is a rich of cultural expressions from previous generations. For educational purpose, Chiang Mai University Library (CMU Library) has continuously conserved cultural heritage. Lanna Documents Unit was found in 1981 for the purpose of conserve and service of Lanna information resources; Lanna was a kingdom in present-day upper Northern Thailand from 13th to 18th centuries. Cultural heritages in printed materials and several heritage manuscripts formats including palm leaf, khoi paper (tree-bark), and sa paper (mulberry) had been conserved. In 1982, Thai newspapers from the library’s holding was put on microfilms. In 1991, the unit changed its name to Northern Thailand Information Center (NTIC) and affiliate under Information Services Department, Chiang Mai University Library. NTIC with the area about 1,000 square meter locates on the 4th floor of the CMU Library building providing rare book collection as a closed-shelves service. In 2005, printed rare books (447 titles) and heritage manuscripts (875 titles) from the library’s holding were digitized. Nowadays (as of 2017) there are 3,730 titles with 4,488 volumes of printed rare books composing of Thai collection 3,555 titles (4,171 volumes) and English collection 175 titles (317 volumes). Those are available on CMU Library’s website. In 2018, 350 reels of digital heritage manuscripts microfilms will be available online. The collections of Lanna cultural heritage in both tangible and intangible in digital format, have been developed from 2005-present publishing online available on the library website with annual online usages are 1,200,000 users and 3,700,000 visits. The plan of 2017-2020, all newspapers from 1950-2015 on microfilms will be converted into digital archive collection.

KEYWORDS
Chiang Mai University Library, rare books, rare materials, heritage manuscripts, Lanna Cultural Heritage Collection

ACM Reference format:
https://doi.org/0000001.0000001
special occasion books such as funeral memorial book, platinum centenary book from government, (3) books owned by important or famous person, (4) books on historical events (5) Royal books.

2.2.2 Preservation and Digitization — Four steps of preservation and digitization includes: (1) remove mold and fungus inside the -15 degree Celsius cold room for 48-72 hours, (2) digitization by using either digital camera for old books or scanner for good conditioned books, (3) assign Dublin Core metadata, and (4) design database and record into the database.

2.3 Digital Lanna Cultural Heritage Collections

The collections of Lanna cultural heritage in both tangible and intangible in digital format, have been developed from 2005-present (6 Collections). It publishes online available on the library website which includes Lanna Photo Archive, Lanna Food, Lanna Tradition, Lanna Songs, Lanna Lacquerware, and Lanna Poetry. Dublin Core (DC) Metadata is mainly used in the collections. The average number of annual online usages are 1,200,000 users and 3,700,000 visits. According to the university innovation strategy, the Lanna Cultural Heritage Collection will expand a new collection every year.

2.3.1 Development Process. From start to finish, the deployment processes to create heritage collections are composed of five steps. (1) Define the scope of cultural heritage along with discuss with Lanna historians and gurus. Then the main area which are viable to develop into the creative economy aspects according to the university strategy of Lanna Innovation. (2) Identify law and regulations concerned including copyright, metadata, website standard, globalization, and accurate reference. (3) Collaborate with Lanna gurus to transfer their tacit knowledge into explicit knowledge. (4) Collaborate with Lanna historians in history, archeology, language, and literatures to gather information, and then compose into the digital format. (5) Design digital platform including database and website for dissemination information. Dublin Core (DC) Metadata is used as a standard for collection metadata.

2.3.2 Lanna Food — In 2007, the library working with Information Technology Service Center prepared food database to publicize the relevant information and to preserve, continue, improve and spread the cultural uniqueness and folk wisdom of the Lanna people. 183 Lanna recipes from 29 experts around Chiang Mai were collected.

2.3.3 Lanna Pictures — In 2008, historical Lanna photos were collected, to research and increase awareness in preserving the Lanna historical and cultural heritage among the people in the northern region in particular. The pictures are mainly those related to Chiang Mai, Lamphun, Mae Hong Son and Chiang Rai’s past. Most of it was the work of Mr. Boonserm Satrabhaya, who was recognized as a prominent figure who recorded the history of Lanna through photographs.

2.3.4 Lanna Tradition — In 2009, to promote cultural tourism in the north while preserving and embracing Lanna art and culture among the people in the north to take pride in their heritage. For 2009 there are three traditions to deal with: Yi Peng, Wedding and Songkran. In 2010, there are some major Lanna traditions in different months.

2.3.5 Lanna Song — In 2010, the collection of 547 Lanna songs were collected and made available online on MP3 format.

2.3.6 Lanna Lacquerware — In 2014, the collection about local lacquerware were collected. The collection includes background and history of lacquerware production, methods and designs of different locations as well as photographs reflecting the roles and functions of lacquerware from the past to the present. Several interviews were collected in five areas from Thailand (Chiang Mai, Lampang, Phrae, Nan) and Myanmar (Kengtung).

2.3.7 Lanna Poetry — In 2017, the collection about local poem were collected. The collection includes background and history information. 101 media were recorded along with poem lyrics written down. The profiles of poem artists were collected as well.

2.3.8 Statistics and Usages – The collection usages have been tracked by JavaScript program via Google Analytics service. The average number of annual online usages are 1,200,000 users and 3,700,000 visits as seen in Fig. 2. The visitors come from more than 150 countries which, the top 5 visitors’ origins are the United States of America, Australia, Japan, South Korea, and the United Kingdom respectively. The visitors are also from more than 11,000 cities around the world. 64% of them visit the website via Google search.

**Figure 2:** Lanna Collection Usage Statistics

ACKNOWLEDGMENTS

The authors would like to thank you M.R.Rujaya Abhakorn, a former director of Chiang Mai University Library as a pioneer of the rare books and Lanna Cultural Heritage Collection. This provide a solid guidance for us to create digital of rare materials for education, research, and maintain Lanna culture.

REFERENCES


Challenges of Personal Digital Archiving in Social Networking Service

Reflections on Archiving Mobile Dating Apps

Papatsara Arsa  
Faculty of Arts,  
Chulalongkorn University  
Bangkok, Thailand  
papatsara.a@chula.ac.th

Songphan Choemprayong  
Faculty of Arts,  
Chulalongkorn University  
Bangkok, Thailand  
songphan.c@chula.ac.th

ABSTRACT

Background. While personal digital archiving is largely considered to be personal choices, understanding strategies and motivations in preserving personal digital collections is crucial for cultural heritage community to understand individuals’ perspectives and provide appropriate recommendations. Social networking services particularly on mobile platform are now ubiquitous exhibiting challenges in preserving personal digital collections. Mobile dating apps have become one of the most popular services presenting a unique preservation context. Relationships with other is a part of human experience that may worth to remember. Many would like to keep these experiences for a short- or long-term period. Using mobile dating apps as a context, we conducted a qualitative study to investigate strategies and underlying motivations in organizing personal digital collection either for short-term or long-term period. This study aims to explore certain challenges in preserving personal digital collections in social networking services context from user’s perspective.

Methods. Fifteen men age between 20-40 years old residing in Bangkok metropolitan area who were active users (using more than 3 times a week) of mobile dating apps for gay men were recruited through advertising profiles and direct messages in apps as well as snowball sampling technique. All participants were asked to write personal diaries, either in paper or electronic format, documenting their engagements with apps for one week during January-March 2016. Semi-structured interviews were followed upon receiving participants’ personal diaries. The data was analyzed according to thematic analysis protocol using Nvivo, a qualitative data analysis software, version 11.

Results. Participants have had different levels of experiences using mobile dating apps, ranging from 1 to 4 years. Many of them used these apps primarily to find sexual partners, while a few were looking for serious relationship, friendship, and conversation partners. Although many of them were looking for short-term relationships and perceived information to be short lived, they deliberately engaged in archiving activities to certain extent.

Archiving activities can be classified into three categories including active keeping, passive keeping, and withdrawal. For active keeping, a user actively and intentionally engages in keeping digital objects emerged from interactions with partners and apps including saving content in local or cloud storages, screen capturing, organizing images into albums, sharing with friends, pinning profiles, and backing up.

For passive keeping, app users purposefully leave digital objects in the apps and expect that the objects would be available for reuse in the future. This group of activities applies when a user perceives that he already remembers information or that information is temporary and contain no or less value (i.e. contacts or conversation transcripts with uninterested profiles).

Those who maintain passive keeping activities would change to active keeping activities or withdrawal when, for instance, there is a limitation on storage space.

Two withdrawal activities found in this study include deleting an object (e.g., files, images, etc.) and blocking a profile. Deleting an object is a common activity when the storage space is running out and information perceivably no longer contains value, for example, a relationship has ended or no progress, conversations have been shifted to a more secured platform (e.g., Facebook, Line, WhatsApp, etc.). Blocking a partner profile (i.e., preventing both sides from interacting to each other) is applied when a user calls for attention from his partners as well as to deny contacts from those whom they have no interest.

Users’ motivations to engage in archiving activities include 1) to recall memorable experiences, 2) to ameliorate his feeling, 3) to protect oneself from potential harms, 4) to prevent loss, 5) to express their feelings and emotions, and 6) as a part of selection process. Lack of knowledge, information becoming irrelevant and useless, and benign neglect are the main reasons for disregarding keeping activities.

Discussion. In this transitory context, appraising objects to be kept or withdrawn becomes a dynamic process depending on numerous factors. Mobile dating app users assess value of information mostly from utility perspective (i.e., how it can be used in the future). Screen capturing becomes a new normal for saving information to local devices. Keeping activities involve other services and platforms (e.g., messaging services, storage services) Privacy concerns lead to activities that may affect
stewardship of personal digital objects. Such objects involve anonymity, misinformation, and hidden information.

Apparently, archiving in mobile dating apps heavily focuses on the interplay of various trusted agents. It can be represented through different personas of app users. For example, those who have past experiences of losing digital information tend to have low level of trust on service providers and, therefore, address concerns as well as actively engage in archiving activities.

In most cases, oneself becomes one of the most trusted sources in organizing personal digital collections. Users who believe that they have remembered information and have a good memory would not actively perform keeping activities. In addition, those who perceive themselves for lack of knowledge also rely heavily on passive keeping activities.

Trust in service providers results in benign neglect. Additionally, users who trust service providers would favorite (aka., pin) profiles expecting they would be able to go back and retrieved pinned profiles and relevant information later.

Some mobile dating app users share personal information either of their own or their partners to their social networks (e.g., one-on-one conversation, discussion group), not only for social purposes, but also as a duplication activity as well. They perceive friends in their network to be reliable and trustworthy. Sharing is also considered to be a faster way to duplicate (i.e., distribute) information in multiple locations.

Selection of storage options (i.e., local and cloud-based) are not only based on trust, but also availability (i.e., capacity). Users actively engage in certain archiving activities when their local device storages are running out of space. For cloud storage, users would not adopt such a service when they perceive that the Internet connection is still an issue.

Although none has address expectations on cultural heritage communities in archiving personal digital collections, particularly in this context, the results from this study leave open opportunities for this group of agents to understand individuals in context that could help promote knowledge on appropriate approaches and techniques in archiving personal digital collection.

CCS CONCEPTS
- Human-centered computing → User studies; • Information systems → Digital libraries and archives

KEYWORDS
personal digital archiving, mobile dating apps

ACM Reference format:
https://doi.org/000001.0000001
ABSTRACT
Singapore started its Digital Preservation efforts in 2008. The paper aims to provide an overview on how the National Library Board Singapore, is playing its role in preserving the digital materials and its journey in doing so. It also gives a glimpse of how the institution is going forward with the change in the Act so as to catch up with the advancement of technology and web publishing.

KEYWORDS
Singapore, Digital Preservation, Web Archives

1 OVERVIEW OF DIGITAL PRESERVATION EFFORTS IN SINGAPORE

National Library Board (NLB) Singapore is the institution in Singapore that is responsible for preserving and making accessible the nation's literary and publishing heritage, and intellectual memory. Its statutory function is stipulated in the National Library Board Act (Chapter 197, 1996 Revised Edition) to acquire and maintain a comprehensive collection of library materials relating to Singapore and its people; and to take appropriate measures to maintain and preserve library materials deposited with the Board. With the advancement of technology, more literary materials are published in digital form, including websites, ePublications etc. These materials have a shorter average lifespan, and hence there is a dire need to preserve them.

Figure 1: eResources of NLB

2 BACKGROUND

NLB started a review of the preservation and access of Legal Deposit materials and national heritage materials in 2005. In 2008, the team looking into Digital Preservation System referenced NLNZ’s implementations and performed Gap Analysis for NLB’s special requirements. In 2010, the Digital Preservation System was implemented for Singapore. By 2011, test ingestions of the digital materials were completed and the files were ingested in the Permanent Repository.

Figure 3: National Library Building of Singapore built in 2005

In 2012, a new team was formed to look into digital preservation. Besides NLB contents, public access materials from the National Archives of Singapore1 (NAS) were also included in the ingestion into NLB Preservation System. The Digital Preservation System would preserve materials that were published in Singapore that are deposited with the Legal Deposit in fulfilment of our statutory obligation. In addition, content on and about Singapore, but not acquired via Legal Deposit, are to be preserved. These content can be either published locally or overseas. Most importantly, the completeness of the items must be there, be it print (digitised) or digital (born digital).
These digital content, or object, would be made up of files and metadata and may include any type of content such as images, text, sound, video, etc. These content requires identification and description to be captured as metadata. Most importantly, there must always be at least two copies in at least two different places. NLB also felt that it is important to ingest metadata in the digital preservation system, as we use these metadata to uniquely identify the digital objects, as well as to ensure the digital objects are understandable in the future. It also serves as the tracking means to ensure the object is authentic over long-term preservation.

3 NOW AND FUTURE

NLB will have 500,000 items to preserve by end of 2017, and the number might increase substantially in 5 years’ time, with the anticipation of the change in NLB Act to include legally depositing of digital content such as websites, ePublications. With the change in Act, it is mandatory for publishers to submit their digital publications to NLB. Up till now, publishers with born-digital content submit their publications on voluntary basis. The number of digital publication is not significant as of now. The revision will include harvesting online .SG materials (domain archiving) without requiring written consent from owners and circumvent technological restrictions such as robot.txt, as well as to enable online material archived by NLB to be accessible from NLB platforms. In anticipation of the increased contents, NLB would be constantly reviewing and improving the processes to meet the needs of both NLB and NAS.

NLB is looking forward to the Act being changed in one to two years’ time.

In the near future for web archives, NLB would be preserving the digital content in WARC format. The team would also be converting the current ARC format to WARC format, and preserving the ARC format in the Digital Preservation System.
## Authors Index

<table>
<thead>
<tr>
<th>Family name, Given name</th>
<th>page number</th>
<th>Family name, Given name</th>
<th>page number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akazawa, Hisaya</td>
<td>282</td>
<td>Hojo, Fuko</td>
<td>282</td>
</tr>
<tr>
<td>Amano, Eriko</td>
<td>254</td>
<td>Houssaye, Jordan de La</td>
<td>177</td>
</tr>
<tr>
<td>Aoki, Toru</td>
<td>288</td>
<td>Humblot, Jean-Philippe</td>
<td>229</td>
</tr>
<tr>
<td>Arsa, Papatsara</td>
<td>317</td>
<td>Imker, Heidi J.</td>
<td>285</td>
</tr>
<tr>
<td>Aziosmanoff, Florent</td>
<td>37</td>
<td>Intaraksa, Neelawat</td>
<td>304</td>
</tr>
<tr>
<td>Bailey, Jefferson</td>
<td>265</td>
<td>Jansen, Adam</td>
<td>237</td>
</tr>
<tr>
<td>Bertrand, Elodie</td>
<td>229</td>
<td>Jelicic, André</td>
<td>263</td>
</tr>
<tr>
<td>Blumer, Eliane</td>
<td>263</td>
<td>Johansen, Kathrine Hougaard Edsen</td>
<td>244</td>
</tr>
<tr>
<td>Boock, Michael</td>
<td>84</td>
<td>Johnson, Duff</td>
<td>150</td>
</tr>
<tr>
<td>Boonmee, Wanchat</td>
<td>306</td>
<td>Junlabuddee, Sompetch</td>
<td>311</td>
</tr>
<tr>
<td>Bredenberg, Karin</td>
<td>270</td>
<td>Kajita, Shoji</td>
<td>285</td>
</tr>
<tr>
<td>Brown, Geoffrey</td>
<td>64</td>
<td>Kamitsuna, Shuji</td>
<td>288</td>
</tr>
<tr>
<td>Burgi, Pierre-Yves</td>
<td>263</td>
<td>Kanarkard, Wanida</td>
<td>306</td>
</tr>
<tr>
<td>Byrne, Andrea</td>
<td>194</td>
<td>Karadkar, Unmil</td>
<td>261</td>
</tr>
<tr>
<td>Caron, Bertrand</td>
<td>17, 177, 194</td>
<td>Kasahara, Shoji</td>
<td>250</td>
</tr>
<tr>
<td>Channongri, Nisachol</td>
<td>313</td>
<td>Katre, Dinesh</td>
<td>194</td>
</tr>
<tr>
<td>Chan, Peter</td>
<td>141</td>
<td>Kazama, Yoshiyuki</td>
<td>290</td>
</tr>
<tr>
<td>Chen, Chu’Yin</td>
<td>37</td>
<td>Keating, Stephen</td>
<td>145</td>
</tr>
<tr>
<td>Chen, Shu-Jiu (Sophy)</td>
<td>299</td>
<td>Kelly, Heidi</td>
<td>89</td>
</tr>
<tr>
<td>Chippimolchai, Pongtawat</td>
<td>304</td>
<td>Kingsawat, Krongkaew</td>
<td>311</td>
</tr>
<tr>
<td>Choemprayong, Songphan</td>
<td>317</td>
<td>Klar, Jochen</td>
<td>52</td>
</tr>
<tr>
<td>Clark, Jonathan</td>
<td>268</td>
<td>Khindt, Marco</td>
<td>121</td>
</tr>
<tr>
<td>Cochrane, Euan</td>
<td>17, 99</td>
<td>Koivunen, Kimmo</td>
<td>164</td>
</tr>
<tr>
<td>Di Cosmo, Roberto</td>
<td>74</td>
<td>Krabbenhoeft, Nick</td>
<td>159</td>
</tr>
<tr>
<td>Dangerfield, Marie Claire</td>
<td>60</td>
<td>Kwicien, Kanyarat</td>
<td>309</td>
</tr>
<tr>
<td>Dappert, Angela</td>
<td>168, 270</td>
<td>Kwicien, Kanyarat</td>
<td>311</td>
</tr>
<tr>
<td>Davis, King</td>
<td>261</td>
<td>Ledoux, Thomas</td>
<td>17</td>
</tr>
<tr>
<td>Dillo, Ingrid</td>
<td>9</td>
<td>Ledoux, Thomas</td>
<td>17, 177</td>
</tr>
<tr>
<td>Doubrov, Boris</td>
<td>150</td>
<td>Lee, Christopher</td>
<td>190, 235, 272</td>
</tr>
<tr>
<td>Edwards, Glynn</td>
<td>141</td>
<td>Lee, Jin Ha</td>
<td>145</td>
</tr>
<tr>
<td>Engelhardt, Claudia</td>
<td>52</td>
<td>Lee, Kee Siang</td>
<td>319</td>
</tr>
<tr>
<td>Enke, Harry</td>
<td>52</td>
<td>Lehtonen, Kuisma</td>
<td>164</td>
</tr>
<tr>
<td>Ertl, John</td>
<td>286</td>
<td>Li, Chunqiu</td>
<td>27</td>
</tr>
<tr>
<td>Esquivel, Alicia</td>
<td>252</td>
<td>Lindlar, Michelle</td>
<td>110</td>
</tr>
<tr>
<td>Farquhar, Adam</td>
<td>168</td>
<td>Ludwig, Jens</td>
<td>52</td>
</tr>
<tr>
<td>Fauduet, Louise</td>
<td>37</td>
<td>Lukose, Jones</td>
<td>194</td>
</tr>
<tr>
<td>Fauduet, Louise</td>
<td>229</td>
<td>Maeda, Yukio</td>
<td>275</td>
</tr>
<tr>
<td>Fu, Honghu</td>
<td>297</td>
<td>Mandelbaum, Jane</td>
<td>257</td>
</tr>
<tr>
<td>Galloway, Pat</td>
<td>261</td>
<td>Mannmart, Lampang</td>
<td>313</td>
</tr>
<tr>
<td>Gerrard, David</td>
<td>246</td>
<td>Mason, Sarah</td>
<td>94, 233</td>
</tr>
<tr>
<td>Goethals, Andrea</td>
<td>257</td>
<td>McGovern, Nancy</td>
<td>257</td>
</tr>
<tr>
<td>Goto, Makoto</td>
<td>286</td>
<td>McGuinness, Rebecca</td>
<td>150</td>
</tr>
<tr>
<td>Gotoh, Haruyoshi</td>
<td>284</td>
<td>Mears, Jaime</td>
<td>47</td>
</tr>
<tr>
<td>Hakala, Juha</td>
<td>268</td>
<td>Meyer, Willibald</td>
<td>131</td>
</tr>
<tr>
<td>Hall, Nathan</td>
<td>84</td>
<td>Minamiyama, Yasuyuki</td>
<td>254</td>
</tr>
<tr>
<td>Halvarsson, Edith</td>
<td>94</td>
<td>Minato, Kotaro</td>
<td>250</td>
</tr>
<tr>
<td>Hangal, Sudheendra</td>
<td>141</td>
<td>Miura, Kazuki</td>
<td>293</td>
</tr>
<tr>
<td>Hara, Shoichiro</td>
<td>1</td>
<td>Mooney, James</td>
<td>246</td>
</tr>
<tr>
<td>Hara, Shoichiro</td>
<td>277</td>
<td>Morikawa, Kaichiro</td>
<td>279</td>
</tr>
<tr>
<td>Hockx-Yu, Helen</td>
<td>206</td>
<td>Moro, Shigeki</td>
<td>286</td>
</tr>
<tr>
<td>Family name, Given name</td>
<td>page number</td>
<td>Family name, Given name</td>
<td>page number</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------</td>
<td>-------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Myers, Steven</td>
<td>64</td>
<td>Tilbury, Jonathan</td>
<td>99</td>
</tr>
<tr>
<td>NDiaye, Nola</td>
<td>37</td>
<td>Tomioka, Tatsuji</td>
<td>282</td>
</tr>
<tr>
<td>Nagamasu , Hidetoshi</td>
<td>284</td>
<td>Tornroos, Juha</td>
<td>164</td>
</tr>
<tr>
<td>NAKAMURA, Akinori</td>
<td>14</td>
<td>Tuamsuk, Kulthida</td>
<td>309</td>
</tr>
<tr>
<td>Nakanishi, Tomonori</td>
<td>293</td>
<td>Tunnat, Yvonne</td>
<td>110</td>
</tr>
<tr>
<td>Nat, Joost van der</td>
<td>154, 259</td>
<td>Vatanen, Mikko</td>
<td>164</td>
</tr>
<tr>
<td>Neuhard, Mads</td>
<td>244</td>
<td>Veenendaal, Remco van</td>
<td>60</td>
</tr>
<tr>
<td>Neuroth, Heike</td>
<td>52</td>
<td>Vries, Denise de</td>
<td>131</td>
</tr>
<tr>
<td>Ng, Joshua</td>
<td>225</td>
<td>Wang, Hsiang-An</td>
<td>299</td>
</tr>
<tr>
<td>Nishioka, Chifumi</td>
<td>282</td>
<td>Waterman, Kees</td>
<td>216</td>
</tr>
<tr>
<td>Nojima, Hideo</td>
<td>250</td>
<td>Weatherburn, Jaye</td>
<td>248</td>
</tr>
<tr>
<td>Okamoto, Naosuke</td>
<td>293</td>
<td>Wilcox, David</td>
<td>57, 267</td>
</tr>
<tr>
<td>Omura, Akemi</td>
<td>282</td>
<td>Wilson, Carl</td>
<td>17, 110, 150</td>
</tr>
<tr>
<td>Ozeki, Katsuhisa</td>
<td>293</td>
<td>Windleharth, Travis</td>
<td>145</td>
</tr>
<tr>
<td>Pang, Natalie</td>
<td>225</td>
<td>Woods, Kam</td>
<td>190, 235</td>
</tr>
<tr>
<td>Pattanakiatpong, Wararak</td>
<td>315</td>
<td>Work, Lauren</td>
<td>89</td>
</tr>
<tr>
<td>Potter, Abigail</td>
<td>47</td>
<td>Wu, Zhenxin</td>
<td>297</td>
</tr>
<tr>
<td>Praetzellis, Maria</td>
<td>265</td>
<td>Wuwogse, Vilas</td>
<td>304</td>
</tr>
<tr>
<td>Pretlove, Lee</td>
<td>233</td>
<td>Yamada, Taizo</td>
<td>275, 288</td>
</tr>
<tr>
<td>Ras, Marcel</td>
<td>60, 154, 259</td>
<td>Yamaji, Kazutsuna</td>
<td>285</td>
</tr>
<tr>
<td>Reecht, Stéphane</td>
<td>177</td>
<td>Yoshiga, Natsuko</td>
<td>185</td>
</tr>
<tr>
<td>Rossenova, Lozana</td>
<td>227</td>
<td>Zacchiroli, Stefano</td>
<td>74</td>
</tr>
<tr>
<td>Saika, Tadahiro</td>
<td>280</td>
<td>Zhang, Xiaolin</td>
<td>297</td>
</tr>
<tr>
<td>Schaefer, Sibyl</td>
<td>257</td>
<td>Zhou, Peter X.</td>
<td>11</td>
</tr>
<tr>
<td>Schneider, Josh</td>
<td>141, 231, 274</td>
<td>Zierau, Eld</td>
<td>196, 257, 270</td>
</tr>
<tr>
<td>Seles, Anthea</td>
<td>194</td>
<td>Zwaard, Kate</td>
<td>47</td>
</tr>
<tr>
<td>Shimadzu, Yoshiko</td>
<td>286</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shu, Wang-Jung</td>
<td>299</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sicre, Adèle</td>
<td>37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sierman, Barbara</td>
<td>216</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sim , Ju Wei</td>
<td>319</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singvongsa, Sisavanh</td>
<td>301</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sitthivong, Yodsaya</td>
<td>306</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soh , Lin Li</td>
<td>319</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Somerkoski, Pauliina</td>
<td>164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Srirphan, Kochaporn</td>
<td>315</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stobbe, Oleg</td>
<td>99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strong, Dena</td>
<td>104</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suchodoletz, Dirk Von</td>
<td>131</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sugimoto, Shigeo</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sugimoto, Shigeo</td>
<td>275</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tadaki, Shin-Ichi</td>
<td>185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tarashima, Satoshi</td>
<td>295</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tarkhani, Zahra</td>
<td>64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technologies, Gail Truman</td>
<td>257</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Center for Integrated Studies of Cultural and Research</td>
<td>277</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thomas, Céline</td>
<td>37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thornton, Katherine</td>
<td>17</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>