



Published on *Ariadne* (<http://www.ariadne.ac.uk>)

CURATEcamp iPres 2012

13 December 2012 - 12:21pm

Table of Contents [hide]

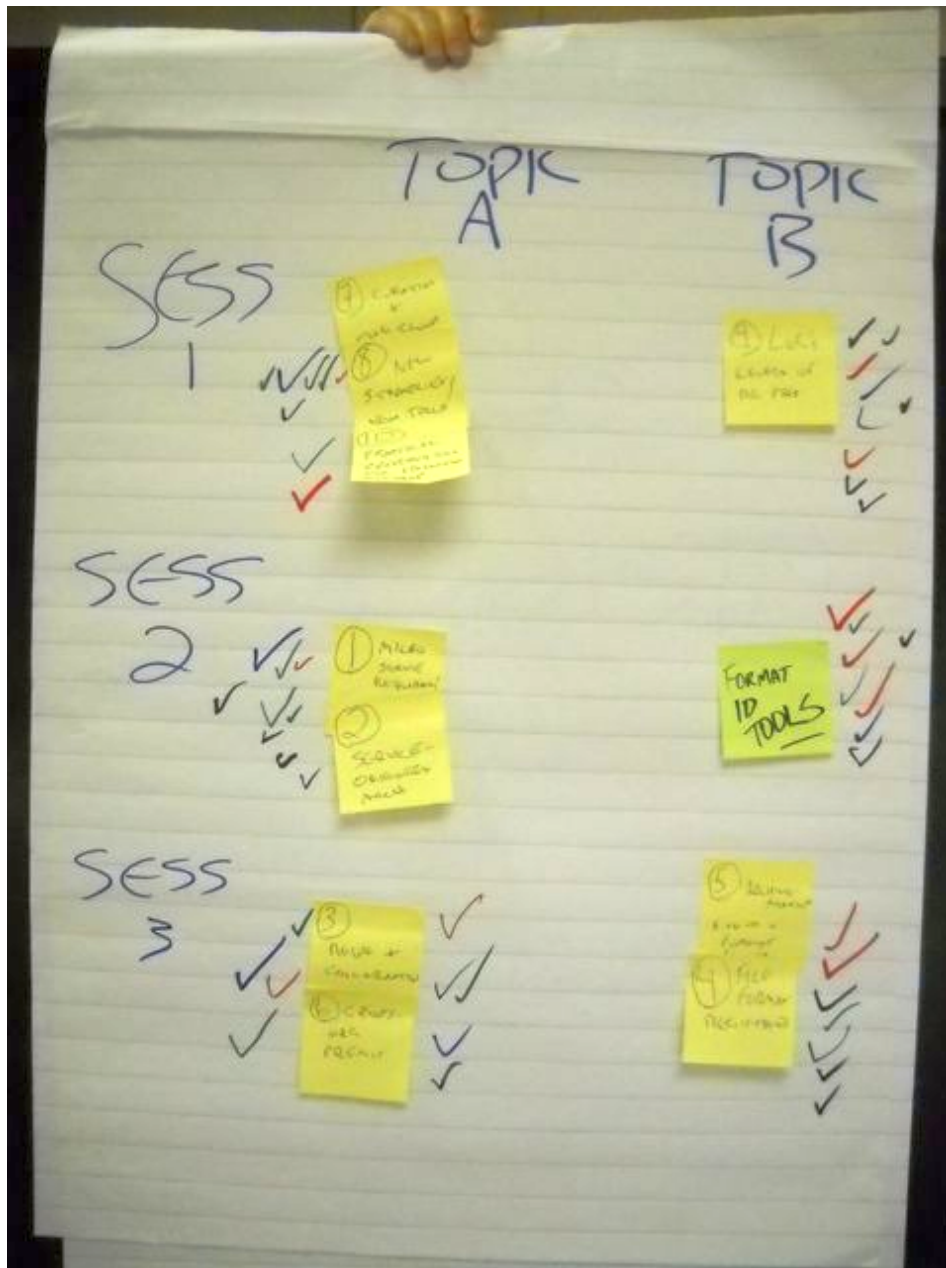
1. Feedback on National Digital Stewardship Alliance's Levels of Digital Preservation
2. Curation and the Cloud
3. Requirements for Digital Curation Micro-services
4. Requirements for File Format Identification and Registries
5. Lunchtime Lightning Talks
6. Conclusion
7. References
8. Author Details

CURATEcamp is 'A series of unconference-style events focused on connecting practitioners and technologists interested in digital curation.' [1] The first CURATEcamp was held in the summer of 2010, and there have been just over 10 Camps since then. The activity at CURATEcamps is driven by the attendees; in other words, 'There are no spectators at CURATEcamp, only participants.' [2] Camps follow the 'open agenda' model: while organisers will typically build the activity around a particular theme within the field of digital curation, and sometimes (but not always) collect topics for discussion, there is no preset agenda. The event's structure is determined at the beginning of the day by having participants propose and vote on topics of interest. The selected topics are then placed in an outline or 'grid' that records the day's activity.

18 people attended the iPres 2012 camp. The group selected four projects to work on from the list compiled prior to the Camp [3]. In addition, some participants gave lighting talks during the lunch break.

Feedback on National Digital Stewardship Alliance's Levels of Digital Preservation

In September, a working group from the National Digital Stewardship Alliance (NDSA) [4] released for public comment a document describing levels of digital preservation [5]. This document arose as a result of a perceived lack of guidance on how organisations should prioritise resources for digital preservation; the levels do not address what to preserve, workflows, preservation platforms, and other policy or operational details.



The CURATEcamp iPres 2012 grid

The document, which is presented in the form of a matrix of columns and rows, defines four levels:

- Level 1: Protect Your Data
- Level 2: Know Your Data
- Level 3: Monitor Your Data, and
- Level 4: Repair Your Data

and across these levels, five functional areas

- Storage and Geographic Location
- File Fixity and Data Integrity
- Information Security
- Metadata
- File Formats

In general, each level builds on the amount of organisational resources required to perform the preservation functions.

One of the CURATEcamp groups assembled to provide feedback on the document. Their feedback has been posted to the NDSA blog as a comment. [6] In summary, the group saw a need for a 'Level Zero', one identifying 'something you can point at that you suspect you are responsible for.' Content identified at this level may not even necessarily be preserved; the point is that an organisation needs to identify that it has content that needs to be evaluated for preservation. Furthermore, the group felt that the level captions (such as 'Know Your Data') were not very useful and recommended removing them from the chart. Finally, contrary to the general pattern of each successive level requiring more resources than the previous one, the group felt that this was not the case in the 'File Formats' functional area but that a case could be made for making file format requirements more substantial as preservation levels increased.

Curation and the Cloud

This group attempted to tackle a lot for a single session: curation and the Cloud, new strategies and tools for new technology, and practical digital preservation solutions for production entities. Ultimately, the group focused on tools and strategies that could be used to preserve cloud-based services and social media, and simultaneously preserve authenticity - the myriad of issues the advent of the Cloud brings to digital preservation.

Preservation of email is not a new issue, and there are many documented workflows around it. However, preservation and curation of Web-based email brings up new issues. Grabbing the actual email is fairly straightforward if credentials are supplied. But, once the email is curated, its presentation can be difficult. How do we, or should we, present how a given user flags items? How his or her email is organised? Content is king, but what about context?

The group focused on curating email from Hotmail, Gmail and similar services. Questions raised included:

- What is in a SIP (Submission Information Package), and what is in a DIP (Dissemination Information Package)?
- What kind of tools and strategies could we use to preserve cloud-based services and social media while still preserving authenticity?
- Did we have to come up with different preservation strategies and use different digital forensics tools (compared to what we use for physical media)?
- Additionally, what tools did we use to provide user access to electronic material, such as emails, social media, Google docs, etc.?
- Should we have levels of email preservation?

Finally, email is not the only cloud-based service that digital curators will need to work with - Dropbox, online backup services, Google Docs, and social media services will also pose their own challenges. In the very limited amount of time remaining in the session, the group raised similar questions to these other cloud-based services as they did with Web-based email, and also noted that standards and best practices for archiving social media and other cloud-based content needed to be developed by the community.



Central Toronto, venue for iPRES 2012 events

Requirements for Digital Curation Micro-services

A popular design pattern in digital curation and preservation is micro-services, which are single-purpose, loosely coupled tools that are combined together into workflows. Another CURATEcamp iPres 2012 group examined the feasibility of formulating a standardised set of requirements that could aid micro-service users, developers, and integrators. Functional requirements worth exploring included:

- That a micro-service can be used synchronously (i.e., run in parallel) or asynchronously (run in a particular sequence) depending on the workflow
- The benefits and disadvantages of allowing micro-services in the same workflow to be written in the same programming language (as one participant put it, 'Why not just write all your services in Python?')
- Categorising micro-services using a standard taxonomy, for example, the PREMIS Data Dictionary eventType vocabulary
- Standardised input formats, output formats, and error messages
- Shared guidelines for installation, user testing, and evidence-based evaluation of micro-services

Examples of 'good' and 'bad' tools that can be integrated into workflows as micro-services are (as an example of the 'good') bagit by the Library of Congress [7], which is well documented (both within the source code and externally), can be included in Python scripts as a library or used as a standalone script, and is easy to install and use; as an example of a 'bad' micro-service, participants nominated FITS [8], which they saw as not well documented and difficult to install.

Some guidelines for developers of digital preservation micro-services already exist in the form of David Tarrant's 'Software Development Guidelines' [9]. Furthermore, participants familiar with work being performed at SCAPE (SCALable Preservation Environments), [10] revealed that that organisation was planning on producing Debian packages for its preservation tools, making them easy to deploy on standard Linux infrastructure.

Requirements for File Format Identification and Registries

File format identification and characterisation are important tasks in digital curation workflows, since successful application of downstream processes (like validation and normalisation) rely on accurate identification of a file's format. Many tools exist that attempt to identify file formats -- too many in some people's opinion. JHOVE, FITS, FIDO, Apache Tika, and the recently released Unified Digital Format Registry (UDFR), which unifies two other services, PRONOM and the Global Digital Format Registry, are all services or pieces of software that perform functions related to format identification and characterisation, but they all do it in slightly different ways.

Building on the work of Andy Jackson [11], Paul Wheatley [12], and the Archivemática Format Policy Registry Requirements [13], this CURATEcamp group discussed limitations of existing tools and approaches, the need for better use cases, clearer functional requirements, and performance optimisations of popular tools.

This discussion continued after the CURATEcamp and resulted in the organisation of the 'CURATEcamp 24 Hour Worldwide File ID Hackathon' [14], which was held on November 16, 2012 and coordinated across time zones from GMT +12:00 to GMT -8:00. The day began in New Zealand and crossed continents with the rising sun, joined intermittently by participants interested in enhancing best practice for format identification and validation. The group communicated on Twitter via the event hashtag (#fileidhack) and in the IRC #openarchives chatroom. This event was a resounding success and ended with summaries provided by the Vancouver, Canada team in the final time zone. Highlights of the outcomes included (with names of principal contributors in parentheses):

- Forking of the open-source FITS (File Information Tool Set) application to OpenFITS [15], so that its performance could be improved substantially (Gary McGath, Independent Consulting Developer, USA).
- Creation of an OpenFITS Debian package so that OpenFITS can be installed easily on standard Linux operating systems (Artefactual Systems, Canada).
- Detailed testing of OpenFITS to determine if the performance enhancements worked; the preliminary results showed that more effort is needed to increase OpenFITS' processing speed (Artefactual Systems, Canada).
- Establishment of the 'Format Corpus' [16], a collection of openly licensed files that exemplify a wide range of file formats (Andy Jackson, British Library).
- Updates to the Precipio [17] and Fidget [18] file format signature generation and testing tools (Andy

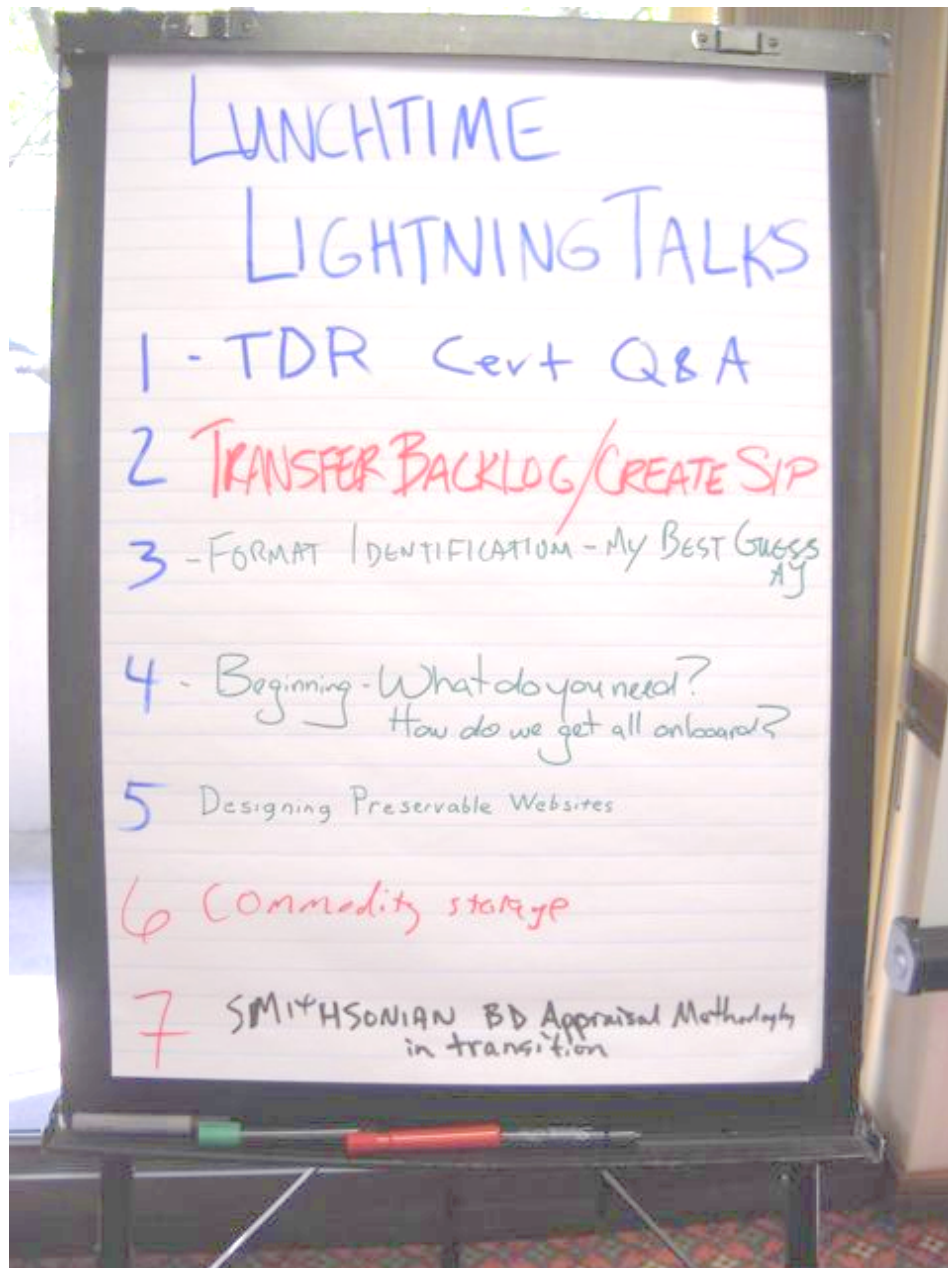
Jackson, British Library).

- Creation of 13 e-book file format signatures (Maureen Pennock, British Library).

These accomplishments all happened within the *same day*, and involved people from organisations ranging from independent consultancies to national libraries. Overall, the 24 Hour Worldwide File ID Hackathon proved to be an excellent example of the digital curation community's ability to tackle a specific set of problems in a loosely coordinated yet highly focused burst of activity.

Lunchtime Lightning Talks

The Camp's organisers offered participants the opportunity to deliver a 5-minute lightning talk during the lunch break. These impromptu presentations covered a broad variety of topics, including one institution's experience of becoming certified as a Trustworthy Digital Repository, workflows for transferring content into digital preservation systems, specific work done to identify file formats in a large Web site archive, and low-cost disk storage options.



Lunchtime Lightning Talks at CURATEcamp iPres 2012

Conclusion

Participants at this CURATEcamp felt it was a success: it generated useful discussion, resulted in some concrete outcomes, and provided a venue for digital curation practitioners and researchers to meet face to face. Moreover, unconference events like this one are spreading in the library and archives community, and are proving to offer a productive alternative to traditional conferences and other forms of collaboration. Part of the appeal of CURATEcamp is that it is easy to organise one; interested readers need only consult the CURATEcamp 'How it works' Web page [2] for more information.

References

1. CURATEcamp home page <http://curatecamp.org/>
2. How It Works <http://curatecamp.org/pages/how-it-works>
3. CURATEcamp iPRES 2012 Discussion Ideas
http://wiki.curatecamp.org/index.php/CURATEcamp_iPRES_2012_Discussion_Ideas
4. National Digital Stewardship Alliance www.digitalpreservation.gov/ndsa/
5. Help Define Levels for Digital Preservation: Request for Public Comments
<http://blogs.loc.gov/digitalpreservation/2012/09/help-define-levels-for-digital-preservation-request-for-public-comments/>
6. Comment by Courtney Mumma, 20 November 2012, 6:11 pm in respect of "NDSA Levels of Digital Preservation: Release Candidate One" by Trevor Owens, 20 November 2012
<http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/#comment-9542>
7. edsu/bagit <https://github.com/edsu/bagit>
8. File Information Tool Set (FITS) <https://code.google.com/p/fits/>
9. Software Development Guidelines <http://wiki.opf-labs.org/display/PT/Software+Development+Guidelines>
10. SCALable Preservation Environments <http://www.scape-project.eu/>
11. Biodiversity and the registry ecosystem
<http://openplanetsfoundation.org/blogs/2012-07-06-biodiversity-and-registry-ecosystem>
12. Don't panic!: What we might need format registries for
<http://openplanetsfoundation.org/blogs/2012-07-05-dont-panic-what-we-might-need-format-registries>
13. Format policy registry requirements
https://www.archivematica.org/wiki/Format_policy_registry_requirements
14. CURATEcamp 24 hour worldwide file id hackathon, 16 November 2012
http://wiki.curatecamp.org/index.php/CURATEcamp_24_hour_worldwide_file_id_hackathon_Nov_16_2012
15. OpenFITS <https://github.com/gmcgath/openfits>
16. Format Corpus <https://github.com/openplanets/format-corpus/blob/master/README.md>
17. Percipio <https://github.com/anjackson/percipio>
18. Fidget <https://github.com/openplanets/format-corpus/downloads>

Author Details

Mark Jordan

Head of Library Systems
Simon Fraser University
Burnaby, British Columbia
Canada

Email: mjordan@sfu.ca

Mark Jordan is Head of Library Systems at Simon Fraser University. His interests include digital preservation, repository platforms, and metadata reuse and exchange. Mark is a contributor to several open source applications and the chief developer of several more. He is the author of *Putting Content Online: A Practical Guide for Libraries* (Chandos, 2006).

Courtney Muma

Archivematica Product Manager/Systems Analyst

Artefactual Systems, Inc.
New Westminster, BC
Canada V3M 3L7

Email: courtney@artefactual.com
Web site: <http://artefactual.com/>

Courtney manages **Archivematica** system requirements, product design, technical support, training, and community relations. She has been a researcher and co-investigator on the **InterPARES 3 Project**, researcher on the **UBC-SLAIS Digital Records Forensics Project**, and is a member of the Professional Experts Panel on the **BitCurator Project**. Courtney has been published in **Archivaria** and has delivered many presentations on the practical application of digital preservation strategies.

Nick Ruest

Digital Assets Librarian
York University Libraries
Toronto, Ontario
Canada

Email: ruestn@yorku.ca
Web site: <http://www.yorku.ca/>

Nick Ruest is the Digital Assets Librarian at York University. He oversees the development of data curation, asset management and preservation strategies, along with creating and implementing digital preservation policies. He is also active in the Islandora community, and occasionally contributes code to the project. He is currently the President of the Ontario Library and Technology Association.

This article has been published under **Creative Commons Attribution 3.0 Unported (CC BY 3.0) licence**. Please note this CC BY licence applies to textual content of this article, and that some images or other non-textual elements may be covered by special copyright arrangements. For guidance on citing this article (giving attribution as required by the CC BY licence), please see below our recommendation of 'How to cite this article'.

Ariadne is published by UKOLN. UKOLN receives support from JISC and the University of Bath where it is based.

© Ariadne ISSN: 1361-3200. See our explanations of [Access Terms and Copyright](#) and [Privacy Statement](#).

Source URL (retrieved on 14 Dec 2012 - 19:20): <http://www.ariadne.ac.uk/issue70/ipres-curatecamp-2012-rpt>