Report on the Digital Library Federation Forum/Digital Preservation 2016

Jody L. DeRidder, 18 November, 2016

Sponsored by the Council on Library and Information Resources (CLIR) The Digital Library Federation (DLF) promotes work on open digital library standards, software, infrastructure, and best practices, including digital stewardship and curation, digital humanities, and data management. This year they have incorporated National Digital Stewardship Alliance (NDSA) and sponsored their annual meeting (Digital Preservation 2016) directly following the DLF forum. There are now 157 member institutions in DLF, and this was their largest conference yet, with 700 attendees, some from other countries. I presented in three sessions (one on digital library assessment, and two on intake and management of born digital content), and assisted with a working lunch meeting on digital library assessment. All were well-attended.

This report will summarize the information gathered from these two conferences which may be of interest to the University of Alabama Libraries. I have organized the primary topics into a table of contents, for easy reference:

Contents

Fedora

David Wilcox reported that 74 institutions are now participating in DuraSpace. They are now working on an API specification effort for core Fedora services (Fedora 4), which will allow for same interface over different implementations. This will enable import/export in standard formats using RDF metadata and BagIt. Currently they're investigating how to measure and benchmark performance and scalability. Extra hardware is required to run this at scale. Test migration tools are available on github, and feedback is requested.

I attended a 3-hour workshop provided by David on basic requirements for ingest into Fedora 4.6. There are 2 types of resources: containers (which can be nested) and binary files. Each is defined by specific JSON or RDF triples in accordance with the Portland Common Metadata Model. For a single-paged book within a collection, one must configure the collection container, the book container, and the page container, and then the page image(s) and metadata. XML metadata records, OCR and transcripts are also considered binary files. Each different type of object requires a complex organization of containers and triples metadata at each container level. Metadata option types include access, bitstream, descriptive, and technical.

Penn State has some tools for migration from Fedora 3 that may be useful. Data normalization is the hardest part. All interactions with Fedora should go through the APIs. More information about how your data is handled (for data integrity issues) can be discovered by studying the Modeshape documentation. Triples metadata is stored in key/value pairs in a database. Version 7 of Fedora will connect directly to the storage system, as opposed to using the Infinispan layer that is in version 6. Fixity checking and storage are the main preservation functions; Fedora is not a complete preservation system. It does give you versioning, but no format migration.

They are moving towards aligning with Memento for versioning. Via the Apache Camel messaging services, Create/Read/Update/Delete (CRUD) operations are supported via linked data, and for authorization, it uses WebAC (W3C web access control), which also uses linked data.

Batch operations are also known as transactions. They tend to be slow; bundled operations can be committed and rolled back. If any entry in a batch upload fails, the entire upload fails. On ingest each bitstream gets a checksum; if it doesn't match an existing one, the system creates a new entity with no warning if the identity is the same as something that is already in the system. Versioning, however, has to be specified. Metadata can be modified without versioning, which is recommended to avoid slowing the system further. Versions do not delete themselves. He was uncertain whether versioning at a high level automatically versioned all subsidiary containers and binary files. There is no good way to move a binary if it's created at the wrong level or in the wrong place. They're debating on adding a FileSet container for multiple binary files: DuraSpace wants it and Fedora doesn't.

More than a couple hundred thousand containers at one level will cause performance issues. An external triplestore (that supports SPARQL-update) can be used to index the RDF triples of Fedora resources and speed access. Fuseki, RDF4J, and BlazeGraph have been tested. Hydra does not use a triple store; instead uses a very robust query via SPARQL. Duraspace does use triple store.

It is possible to put SOLR on one server, Camel (the messaging service) on another, and Fedora on a third. By default, the system allows queries against your content using SPARQL. SOLR is much faster than Fedora, and provides search and retrieval via a discovery interface such as Blacklight. ***Fedora doesn't come with a triple store, SOLR or Camel, or the discovery interface – all those would need to be added.***

Research Data & Long-Term Access

ETD Research Data

Sam Meister & Katherine Skinner (Educopia) presented on preservation of ETD research data & complex digital objects. IMLS funded the ETDplus team. They performed a survey of 795 students on 8 campuses, and found that most ETD projects have more than just PDF, and for a 3rd, something outside the PDF was their most important output. Of staff members, 80% thought the PDF was the most important project.

80% of the students are planning to user or already using data types beyond text, such as tabular data, images, code, and digital text. Students wanted written advice, so the ETDplus team developed 6 guidance briefs, which can be branded by universities and modified as needed. These can serve to help educate students for future work as well. Oregon State has already adopted these.

The ETDplus Curation Workbench (demo online) is a customizable web-based tool designed to assist students in preparing and packaging ETD supplementary materials for long-term preservation and access. It will be released open source at the end of the grant.

Collecting Researcher Publications to Seed IRs

Lisa Spiro and Shannon Kipphut-Smith of Rice University presented on their method of populating their IR, in the face of minimal faculty participation. They studied the Symplectic Research Information Management System (RIMS), which harvests publication data from multiple systems to reduce data entry; this also makes it easy for faculty to create profiles and CVs, and to facilitate research analytics. However, the use of the Symplectic Elements at the UC system reduced faculty participation. So they selected Thomson Reuters' Converis as a system. Unfortunately, they had difficulty connecting this to DSpace, and concluded that many out-of-the box systems are not user friendly.

Assessing Risk to Drive Decisions

OhioLINK (Meghan Frazer, Judith Cobb) has 121 member libraries, 18 TB of centrally provided digital resources including 26 million journal articles, 160k items in DSpace, 60k ETDs, and 58k eBooks. 10 people to do everything. It's a $300 million investment, and they needed preservation strategies so they brought in Liz Bishoff as a consultant. To develop organizational understanding, they created a SWOT environmental scan (tech and non-tech) and developed a list of recommendations/strategies. To explain to central IT why backups are not digital preservation, Liz walked them through evaluating their current status with regards to the NDSA levels of preservation and JISC's AIDA tool which is now the Assessing Organizational Readiness toolkit. People care about "long term access" not preservation; adjust your language to your stakeholders.

Lessons learned:

- Risk analysis worked to drive the process
- Stakeholder participation is critical
- Inventory existing risk and disaster related policies and procedures

There will always be implementation challenges.

Enterprise-wide Digital Preservation

Northwestern University Libraries (Carolyn Caizzi) and their central IT unit brought in Amy Rudersdorf (AVPreserve) as an outside consultant to address digital preservation strategies for the university as a whole. The campus was building IT audits into their portfolio: a high risk area was digital preservation of library assets. A request for funding to manage this had been turned down by the dean. The action plan from this audit allowed the dean to obtain funding from the higher administration, and they agreed.

There were 3 main recommendations with discrete steps:

1. Establish governance
2. Build new infrastructures to support it
3. Training and education programs for faculty, students, researchers.

They developed a campus-wide preservation policy and established funding lines.

Methodology to assess included interviews, peer institution review and interviews, lit review, campus-wide data surveys, and an analysis of the current state of readiness against existing standards: OAIS, Trusted Digital Repositories, NDSA levels of preservation and the UK Research Data Archive.

Outcomes include:

- Organizational documentation to provide guidance on how to manage data and where it can be stored (not just who does it);
- A leadership body with authority to develop the campus wide preservation program, align approaches and define cross-institutional roles

- Cheap or free central IT-sponsored preservation storage system informed by library curation knowledge and skills, to reduce independently managed storage of data. (Faculty want this to be free.)
- Scalable outreach and training program
- Phased approach to implementation

Only 14% of faculty members store their work on university servers (less than 10% in recent UK survey). IRs need to either broaden their scope and lower barrier to inclusion, or work in parallel with campus data storage systems.

Born Digital Content

Data Curation Network

Lisa Johnston at the University of Minnesota is helping to plan a network of expertise model (the Data Curation Network) for curating research data in academic libraries, supported by Alfred P Sloan Foundation (2016-2017). They will be sharing staff across universities: UMich, Penn State, UIUC, Washington U in St. Louis, Cornell, and U of Minn. Their guiding light is the Findable, Accessible, Interoperable and Reusable (FAIR) principle. It's not enough just to keep the files; reusability is not trivial, and there are many Data Curation activities.

The question is how to scale up across all disciplines? Skills to curate the data may require discipline expertise around types of data, software, code, etc., and file format expertise as well. There is a frequency issue, as there is much demand in some areas due to centers of excellence or certain departments.

They want standards-driven data curation techniques, and they need to span multiple types of infrastructure. ***They also want to expand beyond the current partners to build an innovative community that enriches capacities for data curation.*** Curated datasets are measurably of greater reuse value.

In planning phase now, developing written model for deployment, then pilot phase will test the model and plan how to grow the network (requesting more funding). Each institution will continue to keep their own policies, including their long-term preservation methods.

Curation activities covered include opening/inspecting the files, arrangement and description, editing and creating metadata, corresponding with the author, and transforming file formats.

They performed a baseline study, and are seeking input from researchers now. In Jan 2017, they are releasing an ARL Spec Kit survey. In the spring they will develop financial/governance models and share draft data curation model with stakeholders for feedback. Already they have found via gap analysis that versioning and identifiers are important and not being done. More information is available on their website.

Digital Acquisitions Tool

Bertram Lyons (AVPreserve) shared information on "Exactly: a new tool for digital acquisitions" which he developed with help from Doug Boyd (Louie B. Nunn Center for Oral History, U Kentucky). The tool provides packaging, delivery and notification, using Bagit, ftp4j, JSch, java mail, and derby db. There's not enough time in the day to manually acquire all the digital content desired; we need depositors to help. This makes it easy for them. Exactly is available from github and from the AVPreserve website.

Exactly will package the metadata, create a local copy on the person's computer, integrates with Dropbox folders, optionally supports ftp/sftp, and has an email notification feature and an independent mechanism to verify fixity. Exactly is a client application, not a command-line tool, and is cross-platform. Users need to download it, double click the download, and fill in three things: the title, the location of the content on their computer (they can browse to it, and select as much as they want, including folders), and where to save the copy on their computer. The interface is configurable, and provides zip options as well. When the user clicks "Transfer" the content is sent. Requests from the audience include rsync, a transfer agreement and institutional branding (other requests welcome).

Donors, who are actively collecting info in the field on behalf of others, love the tool. Scalability is limited by the processor on the sending computer and the network (it can take a long time to send bigger packages). The tool itself works really well at a TB scale, with hundreds of thousands of files. There's a google group available for discussion.

Audio and Moving Images in the IR

Susan Perdue, (VA Foundation for the Humanities) found the Columbia University Audio and Moving Image Survey Tool (AvDb) (an Access database) extremely helpful for tracking media types and assigning basic metadata for prioritization.

Managing Incoming Digital Content

Alissa Helms and Jody DeRidder presented the results of a survey on how cultural heritage institutions are managing the intake of digital content (the results are also published in D-Lib). While many institutions are still developing their processes, workflows and policies, some highlights are clear: the amount of digital content being collected is increasing, and the top target formats are TIFF, WAV, PDF/A, MPEG-4 and CSV or TXT. Few are able to extract content from older and more obscure media, and email and executable programs are the least collected content types. The top technical metadata collected are file dates, file type, size, and checksums. Analysis and selection are primarily manual processes, and spreadsheets are widely used to organize and track content. Forensic Toolkit and BitCurator are the lead tools for identification; and Archivematica, Forensic Toolkit and Adobe Acrobat Pro are most useful for processing. Notably, most of the tools identified in our survey are open source. Similarly, in the 2012 ARL survey, 74% (31 of 42 respondents) use open source tools; 50% (21) stated they use commercial tools for digital processing; 43% (18) use home-grown tools, and 29% (12) use outsourced services such as Archive-It.

Results from the survey have been used to seed spreadsheets in an Open Science Framework Project, which are organized to reflect the framework of the AIMS methodology (developed by Harvard, Yale, University of Hull, and UVa). Members of SAA, DLF, and NDSA have been invited to help build a collaborative resource for easy identification of the best tools and workflows for various types of content.

Digitization, Prioritization & Project Management

Project Management

At Northwestern, Jennifer Young and Dan Zellner have developed a points process (points = hours) for project management. This helps them determine their capacity, and then they divide up the points: 60% are committed to scheduled projects, and 40% to ad hoc projects. They meet with curators (and occasional faculty) to determine their needs, and assign the number of points necessary to complete each project. This helps everyone understand the limitations on capacity, and helps determine priorities. They're considering Smartsheets to track projects.

Assessment

The DLF Assessment Interest Group is currently hosting working groups in analytics, costs, cultural assessment, metadata and user studies:

- The Analytics Working Group is developing an annotated bibliography of resources on how web analytics are used to support digital libraries.
- The Cost Assessment Working Group has developed a Digitization Cost Calculator to assist in estimation of the costs for various types of digitization projects. They are actively seeking additional data to support the analysis.
- The Cultural Assessment Working Group is developing an annotated bibliography, and investigating metadata and description practices, selection and prioritization for digitization, levels of digitization and preservation, and how collections are promoted.
- The Metadata Working Group aims to build guidelines, best practices, tools and workflows around the evaluation and assessment of metadata for digital libraries and repositories.
- The User Studies Working Group has two subgroups:
    - User/Usability Studies seeks to develop guidelines and best practices for user & usability studies in digital libraries
    - User/Reuse has submitted a grant proposal to IMLS to obtain funding to develop a toolkit for assessing the reuse of digital content.

Prioritization

Robin Pike at U of Maryland described their prioritization rubric (see slide 7) for digitization and preservation projects. They weight proposals based on multiple aspects, such as copyright, preservation need, department prioritization, relation to teaching/research priorities, funding, etc.

Rapid Scanning

Scott Eldridge at Brigham Young University has developed a low-cost rapid scanning method. Günter Weibal had developed a conveyer belt scanning system which is very expensive and huge; most of us need a low-cost method instead. Priorities were safe handling, image quality and practicality. He engaged the engineering Capstone team at his institution. They built a prototype: a table with a stepper motor attached (controlled by Arduino), and a framework for holding the camera. There are 2 webcams (one inside hood, one outside) which center the documents under the camera. Also included are a hood, a Phase One Captureback camera system, the Capture One software, and a custom script that locates the image and takes the pictures. They use laser light to orient the documents, which helps with automated cropping. First the worker calibrates the light to get the white balance right. Then she rotates table to set things up. She sets down a document, turns the table a bit, lays down another, etc. One person can operate this, but it's more efficient with 2 (using someone to take documents off); that process is more careful with the documents as well. Auto cropping and deskewing is done in the Capture One software, and they have a glass table top so you can do transparencies also (covered with a cloth for other content). The process takes 4-6 seconds per item, and the results are almost 4 star compliant with FADGI standards. They are patenting the process, but will make their plans available so others can build their own; it costs $3k not counting the camera system.

Additional Tools and Resources

Rights Statements for Interoperability

Elliott Williams & Laura Capell (U of Miami) performed a systematic review of 52k digital items in their legacy collections which contained little rights-related information, and assigned rights categories from Rightsstatements.org. These include "copyright undetermined," "copyright not evaluated," "unknown copyright holder," "in copyright" and more. ***DPLA is moving towards requiring this of all submissions***.

They created a rights decision matrix to standardize and simplify decision making. There are a multitude of rights within collections, and many lack critical information. They reviewed deeds of gifts for the collections. This step just provided context, but occasionally was significant. The main challenges were that status assigned was the best guess with the information at hand. For example, orphan works were undated, unpublished, with no information about the creator (labeled as "copyright undetermined"). One question that came up was "what counts as a publication?" And for the international documents, they used "no copyright" and said that it was public domain in the US.

They've added two new fields to their metadata; one is a local rights statement, and one is a standardized rights statement containing the link to the rightsstatement.org. When asked how long this took, 90% was completed in a year by two people doing this on top of their regular work. The majority of the effort went into developing the decision matrix, which is available via a poster in their IR.

IIIF for Image Access

Sheila Rabun (International Image Interoperability Framework) presented on the Open Online Newspaper Initiative (Open Oni Project), a consortium of 38 member institutions. Over 80 institutions, software companies and projects are using their standards now, and they have multiple community groups. A new API for audio/visual content is coming soon. Open-oni takes Chronam (the Chronicling America software from LC) and develops it further, to make it easier to implement and maintain.

IIIF has 4 APIs, all URL based: image API, presentation, search and authentication. Image API transfers pixels and image manipulation. Presentation allows annotation, share and reuse as well as presentation.

To use this standard, you need a JSON file with your content, and a compatible server and back end system. Hydra in a box will support it; also Mirador, Universal Viewer, Internet Archive book reader, diva.js, a plugin for Omeka, and the latest ContentDM.

Newspapers are more complex, as they need to support OCR and annotations. In Mirador, using IIIF, you can compare newspapers side by side.

Carol Kassel presented the implementation of IIIF at NYU Digital Libraries. This was a user experience-driven process. They wanted to support image reuse, remixing, annotation, access control and citations. They chose a 4-part solution:

- Data portability (IIIF)
- Robust viewing and annotations (Mirador)
- Easy to use tools/pedagogical (Omeka)
- Collation tool (Viscoll)

For their image server, they are using Loris, and for their image client, OpenSeadragon. Their next phase is integration: setting up a plug-and-play method of using IIIF. They are interested in creating IIIF-to-go, for smaller institutions without sufficient tech staff, and provision via a cloud-based system. They are talking with people in Canada now on how to make this happen.

Authority Control

Kevin Clair (U. of Denver) presented on the Colorado Local Authorities project, which is publishing local name authorities as linked open data. They've mapped the properties in their data dictionary to existing linked data standards (more on github). They are developing persona-based use cases for the database and reaching out to potential contributors, developing tools for them to add data. So far they are working mostly with public libraries. The goal is to enable the addition of local name authorities for use by others.

Christina Harlow (Cornell) presented on an experimental project to address the issues around managing local and external authorities via linked data. They're pushing towards using ORCID identifiers, and building this with Vitro (the technology underlying their famous VIVO project for showcasing scholarly work). Issues include maintenance, persistence, document-focused metadata structures, inconsistent

and lossy data modeling, inflexible tools/models, and scalability problems. They're currently looking for how best to model entities (foaf:Person / madsrdf:authority / schema:Person).

Library Workflow Exchange

Anna Neatrour and Liz Woolcott of the University of Utah presented on a community documentation of workflow best practices that has been in progress for a year. This is an outgrowth of the DLF Electronic Resource Management Initiative, which resulted in a NISO standards and best practices discussion paper in 2009. Everyone is invited to submit their workflows to their database, which they are promoting via social media.