REPORT ON THE DIGITAL LIBRARY FEDERATION 2014 FORUM

27-29 October 2014, Atlanta Georgia

Jody L. DeRidder

The attendance at this 2.5 day conference continues to grow, topping 400 this year (from over 150 organizations) plus a waiting list. DLF has 28 new members, 15 of which are from the Oberlin Group. As has become the norm, presentations were selected by majority vote prior to the conference. Many of those not selected became poster sessions with brief introductions and "snapshot" presentations lasting only 7 minutes each. This year, the presentations were on the whole less technical than usual, though the focus continues to be on developing community. As always, the program featured several opportunities for networking between, before, and after sessions. Internet access was freely available, and Google docs were created for shared note-taking for each session, linked from the schedule. The Director, Rachel Frick, has left, and they hope to announce the new Director in mid-January.

TABLE OF CONTENTS

IMPROVING DISCOVERY

FIXING GIS DATA DISCOVERY, PRESENTING GEOBLACKLIGHT

Geospatial metadata discovery tools have usability issues that alienate users and prevent wide adoption. GeoBlacklight is a collaboratively (MIT, Princeton and Stanford) designed and developed open source software based on Fedora. They developed a metadata schema based in part on Dublin Core that supports layers and datasets. Concepts and prototyping are online, with requests for feedback.

RDFA MARKUP, SCHEMA.ORG, AND DBPEDIA TOPICS: A CLOSER LOOK AT THE HOLY TRINITY OF STRUCTURED DATA AND THEIR IMPACT ON THE FINDABILITY OF DIGITAL COLLECTIONS

Jason Clark of Montana State University applied Schema.org markup to a digital collection, linking content to DBpedia topics, and compared the results to a similar collection that did not receive this markup. He used entity extraction (online demo) to classify content. He measured 100% more activity on the marked up collection this year versus last year.

LINKED DATA FOR LIBRARIES: FROM EXPERIMENTATION TO PRACTICE AT SCALE

"The Linked Data for Libraries (LD4L) project" (presented by Jon Corson-Rikert, Cornell University) is a collaboration of Cornell University Library, Stanford University Libraries, and the Harvard Library Innovation Lab. The goal is to create a model that utilizes linked open data to capture and leverage the intellectual value that librarians and other domain experts add to information resources. They are using linked data to describe people, organizations, places, subjects, events, works and datasets, using global and local authorities and the BIBFRAME (Bibliographic Framework Initiative) model. The challenges include that BIBFRAME is still evolving; we do not yet know whether non-MARC communities will adopt BIBFRAME; tensions between interoperability and this bibliographic focus; training; and scale.

"Linked Data at OCLC" (presented by Roy Tenant) focused on their efforts to mine the data in WorldCat to create entities that are related to each other: work, place, concept, event, organization, and person. They're creating a "knowledge card" similar to Google's sidebar, based on library data, which he says fixes the problem of the representation record, but they're some question as to whether this is relevant for users who do not seek to end up inside OCLC records. Google is crawling their triples (encoded relationships). There is hope that this will lead to better discoverability for libraries and integration with the web. It already works in WorldCat (example); so far they have over a 197 million work descriptions with URIs. Also, use of Schema.org markup to increase web findability is becoming more prevalent, and OCLC is working to extend with the W3C community and others to extend the markup to better represent library holdings (Schema Bib Extend). The Data Strategy for OCLC is available online.

Carl Stahmer, University of California, Davis, presented on BIBFLOW, an IMLS project of UC Davis library and Zepheira, an information management company. UC Davis has 40 different systems that interconnect on their 20 million records; a change in metadata for any of them impacts them all. He

estimates around 500 different workflows exist that touch the data; they're trying to document them all. They're using Kuali OLE but having to modify it to work with linked data. They use XC Metadata toolkit for transformation, and Zepheira tools as well. They are participating in LibHub, a Zepheira-backed effort to raise the visibility of libraries on the web. Participation in the experimental phase is free; once they are able to raise library search results to high page rankings, they intend to charge a subscription service to libraries.

AVALON MEDIA SYSTEM: IMPLEMENTATION AND COMMUNITY

Indiana University and Northwestern University, in collaboration with nine partner institutions, are completing a 3-year IMLS-funded effort to build a fully open source solution for managing digital audio and video collections. The Avalon Media System is based on Hydra repository software and includes Opencast Matterhorn, Fedora (version 3), Solr for search and Blacklight for retrieval. Avalon used to be well- known as "Variations," and a demo server is online; Northwestern's site is also available. This is integrated with their ILS. The system works with various authentications systems, and offers different levels of support for manager, editor, and depository for each collection. Item access can be configured for various levels as well, and items can be uploaded via the web interface, Dropbox, or Secure FTP. New features include bulk actions and batch processes. They used an agile scrum process for development, with a single blended team from two institutions, meeting daily via videoconference, face to face twice a year, and uploading code to a public GitHub site. They are hoping to integrate hydraDam to support more preservation functions. They are also looking for sustainability/governance/business model and exploring hosted models. They will build support for transcriptions and captioning, and structural metadata for navigation within files. There's a dedicated email listserv for the project. System requirements are fairly heavy; it seems that each institution where this is implemented had a dedicated server for processing incoming content and another just for load balancing, both separate from web delivery and the media server.

BUILDING A TEN-CAMPUS DIGITAL LIBRARY COLLECTION AT THE UNIVERSITY OF CALIFORNIA

Sherri Berger and Brian Tingle presented on the California Digital Library (CDL) effort to provide online access from a shared interface for 10 campus libraries and around 24,000 collections, each with thousands of items. Thus far, 680 collections are online. To date, each library worked independently, publishing content on different types of platforms, creating a silo environment for end-users. None of the platforms in use have worked at scale for providing access to digital content (this apparently includes XTF (Extensible Text Framework), California Digital Library's primary delivery software for digital content). The new platform will be based on Nuxeo (a commercial vendor, using "silver" level support) harvesting metadata from collections across the libraries (similar to how the DPLA- Digital Public Library of America - is functioning). The harvested metadata goes into a Solr index with open API (application programming interfaces) so libraries can build custom interfaces on top of the index to supplement the shared interface.

Their user studies indicated that most people arrive at their site directly on an image, so they incorporated a carousel of other images in the collection across the top of each image display, to advertise other content.

One of the reasons they chose Nuxeo was to leverage the broader technology of document management systems leveraging folders for complex objects. They incorporated Shibboleth for security, which was a real pain. Omeka has a Nuxeo connector plugin. They are moving everything to Amazon Cloud, and paying JIRA for project management support. This all replaces XTF in major areas; that software will still be used for finding aids and XML-encoded text (TEI). All cross-collection searching in XTF is going away, as well as delivery of METS-wrapped digital items.

DIGITAL PRESERVATION

RUNNING UP THAT HILL: THE ACADEMIC PRESERVATION TRUST: A COMMUNITY BASED APPROACH TO DIGITAL PRESERVATION

The Academic Preservation Trust (APTrust), a consortium of 17 institutions, was formed two and a half years ago to take a community approach to preservation of the scholarly record. Metadata is managed by Fedora with pointers to content preserved in Amazon S3 and Glacier with administrative functions built using Hydra and Blacklight. The repository is scheduled to go live in July and will become a DPN node. Diverse collections reveal that we are "way too software dependent." So APTrust remains repository and format agnostic by using the BagIt specification for content submission.

PILOTING A PEER-TO-PEER PROCESS FOR BECOMING A TRUSTED DIGITAL REPOSITORY

The University of North Texas and University of Florida engaged in a collaborative process to each complete a full self-audit using the Trusted Repository Audit Checklist (TRAC) for both institutions' digital repositories. In addition to the self-audit, each institution agreed to participate in a peer review process evaluating and scoring each other's self-audit and supplied documentation. Part of the intent was to pilot a review option that offers more rigor and external feedback than a self-audit, but which also does not have the same financial requirements as a full external certification by a third party. For communication, they used Basecamp, Drupal, phone calls and site visits. Multiple administration changes required continually revisiting agreements to have them re-signed.

AUDIO AND VIDEO AT SCALE: INDIANA UNIVERSITY'S MEDIA DIGITIZATION AND PRESERVATION INITIATIVE

Jon Dunn and Juliet Hardesty reported on an initiative started in 2013 to digitize and preserve over 300,000 audio and video assets from across the university. They partnered with a commercial vendor, Memnon Archiving Services of Belgium, who set up a facility in Bloomington and are producing as much as 12 TB per day. This collaborative effort is in response to a 2008-9 survey that uncovered over 569,000 A/V items on 51 different physical formats held in 80 different organizational units across campus, with significant quantities of rare and unique materials in danger of becoming inaccessible within 5-15 years. They intend for this stage of the process to be finished in 2020. There are four phases to the workflow: pre-digitization, prioritization, inventorying, and forming content into batches for high throughput. UI is also setting up a digitization lab on the Bloomington campus for unique, one-to-one formats that need attention rather than a mass approach.

They developed an in-house system to manage the project, and described the process in detail. The resulting content (over 9 petabytes) will be stored in a scholarly data archive, an access repository, and a preservation repository, which they don't have yet. For out-of-region storage they hope to use APTrust / DPN and data swap agreements. Challenges include rights issues at scale; descriptive metadata and discovery; quality control strategies for mass digitization; strategies for born digital content; how to manage out-of-region preservation storage; and what approach to use for film (more stable as an analog, and carries larger storage challenges). There is a website for the project, and reports are online. Although the original press release said this would cost 15 million, the in-kind time and effort makes the actual cost far higher.

DIGITAL LIBRARY ASSESSMENT

Building on groundwork laid at the 2013 DLF conference, we continued the effort to engage community in developing best practices and guidelines in digital library assessment. Nettie Lagace presented on the current NISO effort to standardize altmetrics (alternative measures of use). Joyce Chapman of Duke presented a beta version of an online digitization cost calculator. I presented on our faculty researcher study, and Ho Jung Yoo presented on a usability study at the University of California, San Diego. We then split into working groups, each with one topical area: altmetrics, cost assessment, and user studies. Each group was tasked with two questions: "What are the critical areas we need to address?" and "What are the next steps we can take?" Each group reported back to the larger audience, and contact information was gathered from over 30 people who want to work on these issues over the next few months. Since the conference, I have organized 4 working groups to follow up, one on each topic and a fourth on Google Analytics. We hope to have results to report at next year's conference.

LEARNING MORE ABOUT OPEN SOURCE OPTIONS

DIVE INTO KUALI OLE (OPEN LIBRARY ENVIRONMENT)

Kuali OLE (Open Library Environment) is a community source next-generation library management system developed through a partnership of research libraries with the Andrew W. Mellon Foundation. Operating since July 2010, Kuali OLE is the one of the largest academic library software collaborations in the US. Kuali is written in Java using a Spring framework, and is based on RICE middleware for use in managing library acquisitions. OLE does not include a discovery interface and is not meant to be a customer-facing system. There's an online demo and test drives available. The demo provided at CNI is online as well as a 2013 slideshare. For this presentation, Jeffrey Fleming of Duke University simply walked us through the online documentation.

DEVELOPING WITH HYDRA

This workshop hosted by Bess Sadler of Stanford walked participants through an online tutorial to demonstrate basic functionality and build a sense of capability with this set of open source tools. The workshop, like the "Hydra Install fest" that preceded it, moved far too fast for participants to absorb what exactly they were doing and how the tasks could form the basis for future development. However, the tutorial content is online and available for further exploration.

CONCLUSIONS

A number of the presentations by major research libraries referred to collaborations with vendor services, indicating a surprising shift away from wholly open-source implementations, and perhaps reflecting a realization that it is smart to combine forces and leverage others' capabilities where it is to our benefit. I remain skeptical, however, of the Zepheira-backed effort to raise page level rankings of library content for a price, leveraging linked data. And I was a bit shocked that California Digital Library has set aside XTF for a vendor-based delivery system. XTF has long been the only viable competitor to Acumen for a web-directory based delivery system (Acumen is simpler and far less difficult to use). This places Acumen at the forefront in terms of open source, easy-to-use web delivery systems for digital content. We should consider engaging other institutions in support of Acumen and release it again as open source software.

It is clear that we need to develop support for better findability of digital library content in web search engines such as Google; schema.org implementation would be the lowest cost method with the highest payoff. (We have already begun implementation of this at UA, and should expand upon it.)

While the Avalon system for audio and video delivery seems ideal, it is also very costly in terms of hardware and effort. I do not advise adoption for UA until such time as we have a tremendous quantity of such content for web delivery. The Indiana media digitization and preservation initiative lays the groundwork for mass capture of content that is quickly becoming obsolete; but again, this is at a high cost, which must be balanced against the value of the materials.

The peer-to-peer approach for becoming a trusted digital repository is one worth considering, as it greatly increases the preservation capabilities of each institution at a far lower cost than the certification process, via accountability. Perhaps we should consider some form of this process in coordination with other institutions in ADPNet.