Toward Semantic Metadata Aggregation for DPLA and Beyond: A Report of the ALCTS CaMMS Heads of Cataloging Interest Group, Orlando, June 2016

By Daniel Lovins

Submitted to Technical Services Quarterly 8/11/16

Introduction

The Heads of Cataloging Interest Group (IG) convened on Monday, June 27th, during the 2016 annual conference of the American Library Association in Orlando. Approximately 91 persons were in attendance. The session featured a panel discussion on the processing of local metadata for Web-scale cultural heritage discovery portals such as the Digital Public Library of America (DPLA).

The speakers were encouraged to address questions of interest to cataloging and metadata department managers, e.g.: What happens when metadata created for a specific library catalog are aggregated and repurposed for a Web-scale discovery service? What kinds of data modeling, mapping, remediation, and reconciliation are needed in advance of such aggregation? What happens when metadata from different domains (e.g., galleries, libraries, archives, museums), created with different standards and schemas, are forced to interoperate semantically?

NYPL as Content Hub

The first speaker was Josh Hadro, Deputy Director of NYPL Labs, coordinator of New York Public Library (NYPL) partnerships with the DPLA, and a board member of the Empire State Digital Network. His slides are available on DropBox.[1]

As a content hub, NYPL has a one-to-one relationship with DPLA. Having already shared more than 1.1 million records, NYPL is DPLA's fourth largest contributor. In its role as hub, NYPL collects metadata from its 26 research divisions and crosswalks them to the DPLA Application Profile,[2] and makes the normalized data available through APIs and bulk downloads.

A major milestone was reached on January 6th, 2016, when the NYPL released nearly 187,000 highest-resolution images to DPLA and to the general public. Hadro demonstrated an innovative NYPL Labs visualization that presented the entire set as thumbnail images tiled on a single Web page.[3] The images can be grouped by creation date (i.e., century), genre, collection, and even by color, and clicking on any one of them opens its resource description in the Digital Collections platform and provides access to higher and highest-resolution images.[4]

The challenges of a managing a content hub include the multiplicity of metadata creators, source systems, and practices, as well as a legacy of system migrations and a lack of complete documentation.

Metadata are harvested from NYPL's 26 research divisions mainly via the Metadata Management System (MMS), which imports data from other systems including through an "archives stack", The Museum System (TMS), and a Sierra ILS. Some categories of materials are described in up to all 4 of these systems. There is also the legacy of system migrations and legacy projects. Sometimes the hub staff are effectively doing "metadata archeology", e.g., screen-scraping rich resource descriptions from within HTML paragraph tags.

[1] http://bit.ly/ALA2016AN-HoC_JHadro
[2] https://dp.la/info/developers/map/
[3] http://publicdomain.nypl.org/pd-visualization/
[4] http://digitalcollections.nypl.org/

There is an ongoing effort to make terms-of-use explicit and then map them to the controlled vocabulary of www.rightsstatements.org. Over the past three years, NYPL has increased the proportion of records with rights statements from about 16% to about 80%.

The 6 members of NYPL's metadata unit perform back-office operations that support the Digital Collections platform and the DPLA content hub, among other duties. They use a Metadata Management System to assign or modify bibliographic and rights statements. These staff members also serve as liaisons, grouping the 26 research departments into 5 different areas, and try to speak with them with a unified voice about "minimal viable descriptions" for discovery of their collections on the Web.

They perform metadata audits on all incoming resource descriptions to ensure that at least 6 core elements are present, well-formed, and useful as access points. They make sure date representations are properly formatted, for example, and that subjects have URIs. Hadro showed a sample CSV output that indicated the presence, absence, and/or frequency of certain metadata properties, and then a statistical summary of those counts. With this information in hand, one can identify common errors, e.g., data corrupted during system migrations, or unit titles missing from EAD source records. Metadata liaisons can then meet with their assigned research divisions and consider options for remediation.

Hadro showed a "chart that launched a thousand questions," which plots average views per collection item against relative size of collections. Given two similarly-sized collections, one can ask, why does one have higher average views-per-item than the other? The answer may involve the nature or desirability of the collection itself or it may have to do with metadata quality. If it is the latter, some additional remediation may be warranted.

Hadro described the possible "skills trajectories" of staffers across the organization working with metadata. Starting with a record-by-record approach, the goal is to transition to working in batch mode with tools like csv files and Excel, and from there to bigger and more versatile tools like OpenRefine, Python or other scripting languages, and beyond.

Finally, Hadro presented a diagram of "The NYPL Registry", which envisions a future of metadata obtained from various repositories getting merged and reconciled, and their results stored as RDF entities in a Registry database. Some identifiers could be hosted and name-spaced locally, and the full graph exposed through a Web interface and APIs. Enhancements to the entity descriptions could be fed back "upstream" to the original repositories.

MDL as a Service Hub

The second speaker was Jason Roy, Director of Digital Library Services at the University of Minnesota Libraries and project manager for the DPLA's collaboration with the Minnesota Digital Library (MDL). His slides are available on Dropbox.[5]

As a DPLA service hub, the MDL aggregates metadata for some 450,000 images, audio, video, newspapers, maps, documents and 3D works from 180 institutions, and conforms them to the DPLA application profile. The hub aggregates records from Minnesota Reflections,[6] the University of Minnesota Libraries, the Minnesota Historical Society, Minnesota Public Radio, and many other institutions, as harvested through OAI, RESTful APIs, and data dumps.

[5] http://bit.ly/ALA2016AN-HoC-JRoy

[6] http://reflections.mndigital.org/

The MDL is responsible for multiple programs, including state/regional metadata aggregation, digitization projects, "digital storytelling," audio/video documentary initiatives, and community-centered engagement and outreach. Only the first of these was addressed in Roy's presentation.

As a service hub, the MDL collects metadata and conforms them to the DPLA application profile. Several Minnesota institutions do not want or need to use the centralized platform of Minnesota Reflections, but they do want the MDL's help getting their metadata into DPLA. This puts MDL in dual role, acting as a full-stack central digital repository for many smaller organizations while still providing single-service aggregation for larger institutions that can provide repository functions for themselves.

Having provided about 450,000 items thus far, the MDL is ranked 8th among DPLA contributors. Of these, about 200,000 are contributed by Minnesota Historical Society,[7] while the Minneapolis Institute of Art provides about 30,000.[8] The latter will be able to contribute additional records, but still needs to standardize terms-of-use statements (as with NYPL, mapping them to rightsstatements.org),[9] and then flag those records with images that have been cleared for sharing and reuse.

These MDL partners expose their normalized metadata through OAI-PMH, RESTful APIs, and data dumps. Roy showed a screenshot of a system administrator's panel with configurable import jobs, with which data can be loaded and transformed "at scale", and where OAI encoding is converted into more human-readable JSON.

With their understanding of application profiles and controlled vocabularies, catalogers contribute to the extract/transform/load (ETL) process by guiding the application of transformation rules from source metadata schemes to the DPLA application profile.[10] They also advise on lookup tables, whereby variant terms can be treated as synonyms and mapped in a many-to-one relationship to a preferred term. Once the ETL is complete, records are loaded into Blacklight, where catalogers can use facets to identify anomalies and flag them for correction.

The MDL's aggregation work requires carefully-worded data exchange agreements,[11] which establish rules for harvesting and sharing of member data. Hub participants are able to establish common ground on policy and permissions, for example, by consenting to the Creative Commons Zero (CC0) Waiver.[12] Roy posed a question that might be raised by a partner organization: "What happens to our metadata and previews if we decide to terminate the contract?" The answer, based on the exchange agreements, and to paraphrase Roy, would be something like: "We will stop harvesting you, but, given the nature of the Internet, there's no way to reel back in the records you've already shared."

MDL's mandate includes support for small institutions, which often have rare and interesting collections that would otherwise not get discovered. For example, an image of a Dakota quiltwork leather vest was contributed by the Minnesota Historical Society via the MDL, and was featured in the *Guardian* newspaper after being found in the DPLA. Another example is an obscure image from the Blue Earth County Historical Society, which, after being featured on the WGBH Digital Mural in Boston,[13] gave the Society a chance to further publicize (and generate support for) its collections.

[7] http://www.mnhs.org
[8] http://www.mnhs.org
[9] http://www.rightsstatements.org
[10] An interesting detail of the ETL process is the retrieval and insertion of geo-coordinates from GeoNames.org, based on place names found in the source metadata.
[11] https://drive.google.com/a/nyu.edu/file/d/0B8MTWFC5wAk2RVcwbFpoanV6b0U/view
[12] https://creativecommons.org/choose/zero/waiver
[13] http://www.wgbh.org/blogs/digital-mural/

Roy went on to describe a project called “Umbra”, which reuses much of the same infrastructure from the MDL service hub. Umbra aggregates African American cultural heritage data from various repositories, effectively reversing the flow of metadata aggregation by getting about 60% of its content from the DPLA.

The first step in improving discovery of African-American collections is to digitize them, after which associated metadata can be enhanced to highlight aspects of special importance to Umbra. Students can click on drop-down menus within the Umbra interface and modify records on a case-by-case basis.

One common type of problem is the false positive. For example, some 300 Justin Bieber related records managed to slip through into Umbra from the Internet Archive. Part of the problem was the phrase “hip-hop,” which matches against Umbra’s controlled vocabulary and was flagged for inclusion. As a solution, some of the false positives can be removed by students, and the remainder can sink in relevance far enough in the results to be effectively invisible. MDL developers are also looking into automatic classification techniques that improve collections filtering.[14]

Questions from the Audience

Following the two presentations, audience members were invited to ask questions.

The first was about tools that make it easier to correct metadata without having to consult the original artifacts and take other time-consuming measures. Hadro acknowledged that there is no “silver bullet”, but he noted the value of exposing resources and metadata to the public, since any viewer can then contribute corrections and identifications (e.g., “that’s a picture of my grandfather”). He said NYPL receives multiple messages of this type every day.

A second question concerned the funding and sustainability of DPLA hubs. Roy noted that the MDL is in good shape since it receives funding from the state through an arts and culture legacy amendment grant, which sets aside a percentage of sales tax receipts for conservation of land and cultural heritage. This provides a baseline, which they try to supplement with grant funding. Regarding NYPL, Hadro noted that it is one of the largest research libraries in the country, and the single largest circulating library in the country, so even though their staff seems large, it is small relative to the work at hand. Being a DPLA hub is a kind of “unfunded mandate”, but also one that fits neatly within the NYPL’s organizational mission. Also, as a content hub, the aggregation efforts are less resource-intensive than for the service hubs.

A third question was whether there are problems in communications across different communities of practice, e.g., metadata librarians working alongside archivists on hub activities. Hadro replied that no one is forced to change their practices or standards. The aggregation and enrichment is layered on top of the source metadata, so there is no interference in specialized metadata work. Also, milestones like the bulk release of public domain images make everyone feel good about collaborating.

A fourth question was about opportunities to return enhanced data to the original repositories. Both hubs are interested in pursuing this.

Conclusions

DPLA content hubs and service hubs face similar challenges in aggregating metadata. These include quality assurance, reconciliation of terms, and conforming source data to the DPLA application profile.

[14] Cf. https://github.com/UMNLibraries/hamster and https://github.com/UMNLibraries/gerbil

An area receiving special attention is the clarification and mapping of rights statements. In some cases, there is no information in the record and it needs to be supplied. In others, there may be notes with vague or irregular wording, and these need to be mapped to a controlled vocabulary in order to be useful in discovery systems (e.g., through faceting and filtering). Rightsstatements.org is helping to make this possible by providing unambiguous statements backed up by persistent URIs.

For both the NYPL and the MDL, serving as a DPLA hub aligns with their institutional missions. By aggregating and enriching cultural heritage data from hub participants, they make their collections more discoverable on the Web and provide a valuable public service. And in order to provide additional value, both the NYPL and the MDL hubs are considering ways to push enhanced metadata (e.g., place names enriched with geographic coordinates) back to their original repositories.

Practitioners and managers of cataloging and metadata services have an important role to play in large-scale aggregation. They can ensure that when data sets from multiple sources are combined and normalized, that the underlying data semantics are preserved. Knowledge of resource description standards and controlled vocabularies continue to be highly valued, but must be applied at scale. An understanding of schema crosswalks continues to be important for aligning metadata with target applications. Metadata audits and index-based faceting can expose problems, while tools like Open Refine and Python can be used for programmatic remediation.