Title
What Is Our Place in the Metadata Ecosystem?

Journal
Technicalities, 39(3)

Author
Riemer, John J

Publication Date
2019-05-01

What Is Our Place in the Metadata Ecosystem?

Technicalities Column

March 9, 2019

by John J. Riemer

Head, UCLA Library Cataloging and Metadata Center

Time was when collaboration in cataloging took the form of going beyond the authoritative Library of Congress copy available to willingness to use the cataloging records created by peer institutions in a similar manner. Major microform sets were an opportunity for institutions to divide up the work of cataloging large workloads, contribute a share of the original cataloging, and then reap the benefit of complete record sets for one’s catalog.   When libraries subscribed to large aggregator databases and they desired sets of cataloging records to represent in online catalogs the full text serial titles available in those large packages, they were willing to communicate to providers such as EBSCO the specifications for record sets they would accept. All of those efforts represent the members of the library community collaborating with each other or delegating the labor to others working on their behalf.

The standards used were exclusively those developed within the library community. In retrospect, it seems quite pioneering for the LC Cataloging in Publication workflow to depend on some metadata produced in a completely different community, so that it could be built upon. This was the ONIX data created in the publishing world, which often gave libraries summaries and tables of contents data. This column will look at some recent developments that seemingly herald great prospects for a lot more cross-community metadata collaboration, as well as reflect on the implications.

Google and Its Partner Libraries

Google may not be the first name that jumps to mind when libraries imagine new untapped potential partnerships. In the days when I was chairing the University of California's Bibliographic Services Task Force,[1] it felt like Google and Amazon were eating our lunch. Users were clearly finding greater happiness using their discovery tools. Library online catalogs were diminished into a humiliating role of a last-minute known-item search checking to find out if the purchase price could be avoided.

Five years later, library views of Google palpably softened when Kurt Groetsch spoke at ALA Midwinter.[2] Google sufficiently valued library metadata to collect a hundred different feeds and it bemoaned that errors it corrected had to be corrected again in subsequent harvests.

In October 2018, I had the opportunity to attend a Google Books Library Summit.[3] Representatives from Google's partner institutions that have been loaning about million books annually for the digitization project that began 15 years ago assembled on the Google campus to take stock of various aspects of the project and discuss the latest developments.

Google Books Data Analyst Erin Dobias made a presentation on the utilization of library metadata at Google.   The corporate mission of Google "to organize the world's information and make it universally accessible and useful,"[4] certainly aligns well with the values of libraries and their metadata specialists. The "knowledge panels" that often result during Google searching are built from metadata provided to Google from libraries and publishers.[5] During the question and answer segment, with the error correction treadmill mentioned above in mind, I asked if Google had ever aspired to obtain the up-to-date metadata from the WorldCat master record. The long-established workflow has the library loaning the book for digitization also supplying its legacy cataloging record, which often dates back to when the book was first acquired by the library. It was startling to hear that all metadata feeds carried equal weight.

It struck summit attendees as odd that Google would have so little to do with metadata creation and enhancement. Like the books they needed to borrow from libraries, they seem equally dependent on the metadata creation efforts of others. This was strikingly different in that Google invests enormous amounts on Machine Learning, Natural Language Understanding, OCR technology, Search, etc. Right before a break, California Digital Library's Ivy Anderson suggested that this dialog pointed to a joint working group between Google and its partner libraries that could exist.

Sitting in the room were no fewer than four chairs of the Program for Cooperative Cataloging, including the current one. After just 10 minutes of energetic discussion, the Google Metadata Working Group was formed under Erin's leadership. The monthly meetings commenced in the new year and initial agenda topics have been dialogs on how Google might get more mileage out of existing MARC records and new data elements designed to support faceting, relationships, and FRBR clustering; commenting on a pair of proposed new schema.org data elements; sharing what happens with the metadata Google collects and clusters. Naturally, I hope we progress to discussions of what tools we need for creating and remediating metadata, and how Google might help develop them with us.

OCLC's Project Passage

Project Passage is a linked data project conducted for 10 months ending in September 2018. Various parties in the library community have been exploring what the post-MARC metadata and tools can look like, how we can move into a linked data cataloging environment.   OCLC's focus in this project was "to improve our understanding of what it means to manage linked data as a community. Along with getting feedback from the pilot partners about needed services, OCLC is learning what kind of ecosystem is needed to support an entity-oriented data management system."[6]

In contrast to a cataloging focus on describing objects and establishing text strings for access points, the project focused on describing entities (persons, bodies, places, events, topics, etc.) and their relationships to the object. OCLC and its partners successfully relied on a copy of the open source MediaWiki software. By the end of the project, they concluded that they had "created an entity ecosystem."[7]   The presenter stated that "matching strings to entities worked well, but the entity store needs to be much more comprehensive, including relationships and mappings." The 1.2 million entities accounted for in the prototype "could represent billions of entities" when scaled to all of WorldCat.[8]

At a subsequent OCLC webinar, Stephen Hearn noted that University of Minnesota's participation in Project Passage raised in his mind questions about "what our place is going to be in a larger metadata ecosystem."[9] He went on to observe:

> *We currently have our own little community with various kinds of barriers but increasingly we are aware that we need to be able to borrow from other kinds of registries to bring in the necessary identification of persons and corporate bodies and topics and material types. We do that somewhat now with authority control, where we can cite a source for a term. But what we really want is to be able to offer the users the ability to navigate, and find out about the concepts, the persons, the bodies, and so forth, and to come back to our resources enriched by that knowledge. This may imply that we're going to be turning some of this kind of work to other communities, other entities, and other institutions, and then taking from them. It may also imply that in some cases we are going to be involving ourselves, and doing collaborative work in some of those other systems. And that could be a really interesting thing! Libraries are already looking into how they can contribute to the wiki community. ISNI is another project ongoing. We'll see more and more of that. We need to be thinking about where our role is in that larger system of metadata management.*

A summary of OCLC's linked data efforts to date has been issued[10] and a full report on Project Passage is expected in the May/June 2019 time frame.[11]

Program for Cooperative Cataloging Wikidata Pilot

The PCC's current Strategic Directions document includes this major goal: "Accelerate the movement toward ubiquitous identifier creation and identity management at the network level." It is further described:

> *We aspire to attain an environment where identity management work activity is characterized by much greater proportions and numbers of entities receiving identifiers; many non-NACO institutions participating; and strategic partnerships and collaboration existing among cultural heritage organizations, rights management agencies, Wikidata, and others. We expect to find ways in a linked data environment where collaboration on identity management can interoperate across multiple data sources. Attainment of this vision will increase both human and machine usage of this data and its overall value.*[12]

In 2017, the PCC launched the PCC ISNI Pilot.[13] In that experiment the PCC would gain access to another authoritative registry, larger than the LCNAF, to support cataloging work; experiment with batch matching and importing tools; and experience working in a different tool and setting. By agreeing to take on the PCC collectively in an "umbrella" membership, ISNI would benefit from the PCC'S high quality authority data, assistance in data maintenance, experienced trainers and documentation writers, and possibly even some skill developers of APIs, etc.

The PCC is currently gearing up to experiment with Wikidata. Wikidata started in late 2012 and underpins multi-lingual Wikipedia articles and other Wikimedia projects with a structured dataset. The members of the PCC Task Group on Identity Management in NACO are inclined to conduct the pilot in the production version of Wikidata.

Compared to the 10-12 million records found in the LC Name Authority File and the ISNI file, there are over 57 million established entities in Wikidata.[14] There are more than 20,000 active users in the Wikidata file.[15] It is tempting to think about how more could join them from among the ranks of catalogers at the 9,000 OCLC member institutions with a full-level cataloging authorization. If the upcoming Wikidata pilot is successful, how much greater a proportion of names going through our daily cataloging workflows could be covered by identifiers?

A significant advantage to possibly moving daily cataloging work into a Wikidata, if a pilot proves this is feasible and compelling, is that the resulting data is readily usable for linked data services. This seems significantly more advanced that stockpiling subfield $0 (Authority record control number or standard number) and

$1 (Real World Object URI) data in bibliographic record access points for eventual support of conversion to linked data.

Major Implications

If one embraces the much broader participation model for metadata creation and maintenance that the above developments suggest, it raises questions about what will become of professional role of the cataloger, original cataloging workloads, etc. From my speaking engagements in the past five years, where I address what new roles traditional cataloging units can and should embrace, I often conclude with this possible new outlook that catalogers could embrace: “My job is whatever nobody else is doing.” The professional of tomorrow will need to survey, assess, and complement the metadata creation efforts taking place in other sectors of the ecosystem.

A question recently arising in PCC Policy Committee circles is the extent to which we feel it is important for PCC output to conform to Resource Description and Access (RDA). I recently wrote down my thoughts on the matter to share with colleagues:

Part of the benefit of changing over to creating linked data natively is that we would be able to collaborate with others working in the same or a similar schema, such that we could frequently build on the metadata provided by others, as opposed to create it all ourselves in the library community according to standards that are relatively unique to our community.

Strategically, we will come out much better if we go the route that enables lots of collaborating with others and frequent capitalizing on the efforts of others. We will end up with fuller metadata on a much greater array of resources. An implication of this practice is that we may well end up with metadata whose degree of RDA influence is significantly less than what we would have if we created the metadata largely ourselves or if we dictated that any collaboration with us had to involve RDA adherence.

I believe it was the prior PCC Strategic Directions document (2015-2017) that said we sought to influence others as well as to be influenced by others. I would want to say I accept the implied outcome of the above route I recommend. While RDA would have a diminished role, we could say it still has an important role. We could say it represents our best thinking on what practice we think should apply in a vacuum. I just would not say it is the only standard or best practice we are willing to live by.

Conclusion

When I look back at my previous columns in this space, I notice that I once viewed a successful effort to open up the LC NACO file to many more new contributors as a real sea change in the cataloging world. In comparison, that is now feeling like a small-scale dream. If our place in the metadata ecosystem were to move in the direction that the above recent developments are suggesting, that would be a tsunami!

[1] University of California. Bibliographic Services Task Force. "Rethinking How We Provide Bibliographic Services for the University of California: Final Report" (December 2005) http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf

[2] Eric Hellman. "Google Exposes Book Metadata Privates at ALA Forum." Cataloging Futures blog entry, January 18, 2010. http://go-to-hellman.blogspot.com/2010/01/google-exposes-book-metadata-privates.html

[3] Google Books Library Summit, October 3-4, 2018, held on Google's Sunnyvale campus.

[4] Google. About Google page, viewed March 10, 2019. https://about.google/

[5] Searches for titles ("Select Orations of Marcus Tullius Cicero with explanatory notes") or authors ("Cicero") on google.com will produce examples of these "knowledge panels." A search result somewhat like a series ("Jean Auel books," "Ski books," or "Cupcake club books") will result in a "book carousel" display.

[6] A quote from the project partners web site, in Joyce Bell. "OCLC's Project Passage" a presentation made at the Middle East Librarians Association 2018 Annual Meeting, slide 3. https://bit.ly/2UNQ0uM

[7] OCLC. "Project Passage Partner Meeting," September 27, 2018. Powerpoint available at https://www.oclc.org/content/dam/research/presentations/pace/PassagePartnerMeeting-20180927 (2).pptx (slide 10)

[8] Ibid. (slide 11)

[9] [OCLC] Works in Progress Webinar: Lessons Learned from a Linked Data Prototype for Managing Bibliographic Data, October 30, 2018. https://www.oclc.org/research/events/2018/103018-linked-data-prototype-managing-bibliographic-data.html

[10] OCLC. "OCLC and Linked Data," January 2019. http://oc.lc/linkeddatasummary

[11] OCLC. "Connecting with Your Cooperative," webinar hosted by Meryl Cinnamon, February 13, 2019.
https://bit.ly/2HJheQw (comment made approximately at the 41:30-minute mark)

[12] Program for Cooperative Cataloging. "Strategic Directions, 2018-2021." Revised January 24, 2019. http://www.loc.gov/aba/pcc/about/PCC-Strategic-Directions-2018-2021.pdf (page 5, item SD4)

[13] Program for Cooperative Cataloging. "PCC ISNI Pilot Home," updated March 12, 2019. https://wiki.duraspace.org/display/PCCISNI/PCC+ISNI+Pilot+Home The pilot began mid-2017 and continues throughout 2019.

[14] Association of Research Libraries Wikidata Task Force. ARL Wikidata Task Force White Paper, November 30, 2018 https://docs.google.com/document/d/1ZsOyw2sOD3a7xJQ6XCSYDGjZUPxGGl8tuvC7vvtlJRU/edit (page 8),

[15] Wikidata Dashboards. Wikidata Site Stats, viewed March 17, 2019 https://grafana.wikimedia.org/d/000000162/wikidata-site-stats?orgId=1&from=now-1y&to=now