OCLC and Linked Data: The transition to contextual metadata

Presenters

- Shane Huddleston -- Product Manager, CONTENTdm (OCLC)
- John Chapman – Senior Product Manager, Manager Metadata Services (OCLC)
- Nathan Putnam -- Director, Metadata Quality (OCLC)

Presentation summary

OCLC experts shared insights into OCLC's recent linked data work, including the impacts of contextual (linked) metadata on the library community. Specific activities described include: OCLC's recent CONTENTdm Linked Data Pilot, an update on the Shared Entity Management Infrastructure , and OCLC's work model within the entity management infrastructure. Originally broadcast 2021-03-02.

Acronyms referenced during the presentation:

API

Application Programming Interface

BIBFRAME

Bibliographic Framework Initiative https://www.loc.gov/bibframe/

CrossRef https://www.crossref.org/

Digital Object Identifier (DOI) Registration Agency

FRBR

Functional Requirements for Bibliographic Records https://www.ifla.org/publications/functional-requirements-for-bibliographic-records

See also: *WEMI*, OCLC Research Activities and IFLA's Functional Requirements for Bibliographic Records https://www.oclc.org/research/activities/frbr.html

IFLA https://www.ifla.org/

The International Federation of Library Associations and Institutions (IFLA)

See also *FRBR*, *LRM*

ISNI https://isni.org/

International Standard Name Identifier

ISO https://www.iso.org/

ISO, the International Organization for Standardization

ISO Standards - “Standards are the distilled wisdom of people with expertise in their subject matter and who know the needs of the organizations they represent” – From https://www.iso.org/standards.html

LCNAF

LC/NACO Authority File

LRM

IFLA Library Reference Model, the successor to FRBR, FRAD, and FRSAD https://www.ifla.org/publications/node/11412

NAR

Name Authority Record

ORCID https://orcid.org/

Open Researcher and Contributor ID

PCC https://www.loc.gov/aba/pcc/

Program for Cooperative Cataloging

PID

Persistent Identifier

RDA

Resource Description and Access https://www.rdatoolkit.org/

ROR https://ror.org/

Research Organization Registry [primarily Funder IDs]

SEMI

Shared entity management infrastructure https://www.oclc.org/en/worldcat/oclc-and-linked-data/shared-entity-management-infrastructure.html

VIAF http://viaf.org/

VIAF® (Virtual International Authority File)

WEMI

Work-Expression-Manifestation-Item (From *FRBR*)

Other resources related to the presentation:

CONTENTdm Linked Data Pilot https://www.oclc.org/research/areas/data-science/linkeddata/contentdm-linked-data-pilot.html

> *Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project* https://oc.lc/transform-linked-data

OCLC and linked data https://www.oclc.org/en/worldcat/oclc-and-linked-data.html

OCLC Research Linked Data https://www.oclc.org/research/areas/data-science/linkeddata/linked-data-outputs.html

Member Questions

SEMI questions

How will the shared entity management infrastructure be made available? What is your business model?

Answer: In their grant award, the Mellon Foundation specified that OCLC provide free access to data, while also providing valuable services that earn the revenue required to keep the infrastructure sustainable. To that end, OCLC will be publishing entity data as linked data via Worldcat.org and will also be providing subscription access to user interfaces and APIs to work with the data.

Will the APIs be public or by subscription?

Answer: The APIs will require a subscription.

Will this work result in an OCLC linked open data service or will it be a proprietary system service?

Answer: We'll be committed to publishing the data on the web. The comparison I'd like to make is to viaf.org where there is a public facing web page and information on each and every VIAF identifier.

We also will need to keep the infrastructure going and so we'll be earning revenue based on subscriptions. The subscriptions will be focused on use of the tools and API access. But the goal here is to have URIs that you can put in your metadata that everybody can understand.

Will the data be downloadable -- like the LCSH and LCNAF? Much more fun than calling a remote API thousands of times.

Answer: We don't anticipate having bulk downloads available due to the size and the desire to avoid stale data circulating in the environment.

If you already have research data in a Wikibase, is there an integrated tool one can use to upload the data to Wikidata, or to the shared entity management system once it is operational?

Answer: Our APIs will provide the opportunity to add new entities to OCLC's entity aggregation from your local system. We don't provide tools to upload to Wikidata at this time, but do incorporate Wikidata identifiers, when appropriate in our entities.

What do you mean by "at scale"?

Answer: "At scale" is a phrase which is shorthand for the ability to provide real-time descriptive metadata infrastructure for hundreds of millions of resources, providing them to thousands of library customers and to the web in general.

A lot of discovery metadata is generated and managed outside libraries--by vendors, indexing agencies, publishers, etc. Is SEMI looking to engage those communities and get their participation?

Answer: Yes. We are looking for opportunities to do this very thing.

Moving beyond Wikibase to what, exactly?

Answer: In terms of the infrastructure it uses behind the scenes, OCLC won't be replacing Wikibase with a single new platform, but instead moving to a set of different components – for example, for its triple store, user interface framework, indexing engine, etc.

Data/identifier questions

How do your persistent identifiers for Persons relate to the persistent identifiers in the LCNAF?

Answer: Many Persons will have LCNIs included.

Is ISNI one of the name identifiers you provide?

Answer: ISNI identifiers appear in some of the entities because they are in the source data. Investigations are ongoing for further integration with ISNI ids.

How would the definition of "Creative Work" map to either BIBFRAME Work or the RDA /FRBR / WEMI definition of "Work"?

Answer: It will be related to these, but to what extent is still being evaluated.

What data model is behind "work"? BIBFRAME? LRM? Something else?

Answer The data model behind work is OCLC's works model, similar to what is used in WorldCat bibliographic catalog. It has or will have relations to BIBFRAME and LRM.

Has the creative work model been tested out on serials with their title changes, splits, supplements, etc.?

Answer: This model works reasonably well right now with books and other monographic type things, so, what we will need to do is look at how it does play out with continuing resources. I know the people with RDA have done all sorts of stuff. With diachronic works and various things like that and making sure that the model ends up working in that way is definitely a priority for us.

Digital resources question

How does this apply to cultural heritage?

Answer: When we think about digital materials, a characteristic thing about them is their uniqueness, and that often applies to the descriptive entities that are being used with them. Often, they're unique. They're not well-known individuals, let's put it that way, but they might be reused. Today, they would apply to a whole bunch of materials, especially within a given archive that has a lot of the same kind of thing. There's certainly plenty of cultural heritage material that does refer to well-known individuals. Even the types of individuals that you'd see often and would be in entity management infrastructure make sense to link to. You get some, some lift from that. But playing further forward into the future, you can apply the same concepts. The same sort of linked data concepts to sort of local entities, or regional entities, or that kind of thing, lesser known entities. The same sort of technical principals applies and gives you the management benefits and the discovery benefits, regardless of whether they're sort of authoritative. Persons or just ones relevant to a smaller subset of data. For me one of the things I'm most interested in is sort of that person-as-subject relationship where we see some of that in the standard authority file structure. But I think it's an area that we can explore more. Whereas someone like Winston Churchill is represented in authority files because he wrote a couple of books, but he's much more valuable and interesting and well-connected as a topic than he is as a creator.

Incorporating other standards:

Are there plans to incorporate existing PIDs: Orchid, ROR, CrossRef

Answer: We focused in this discussion about the creative work model just because there have been lots of questions about it, and we don't have as much prior work like we do on person entities to kind of point out what we're doing. To give some background, yes, we plan to incorporate other identifiers for persons. One of our hopes is to expand past the kind of relying on the authority files. sort of, over represents authors of monographs and underrepresents people who are involved in serials literature, periodical literature, licensed material. o that's an important aspect of it as well.

How about cooperation with ISO-standards? ISNI, ISBN, ISSN ...

Answer: We will be incorporating ISNIs where appropriate. Currently, the level of granularity for creative works is not always the same as represented by ISSNs and ISBNs, but we will continue to look for opportunities to provide useful relationships to other identifier schemes. The WorldCat bibliographic database will continue to refer to ISBNs and ISSNs and will begin to link to WorldCat entities.

Is the data in WorldCat then open in LOD format? in RDF?

Answer: it will be published via worldcat.org, in RDF, and in a few different encoding formats.

Entity Scope:

You're focused on people and subjects, correct? So, no linked data for controlling 830 series information?

Answer: We are also developing some other types of its kind on an as needed basis. Sometimes it helps to have an additional point of perspective, or triangulation on creating or defining entities. OCLC has established a few tens of thousands of place entities in addition to the works and persons. As we go on, we expect to find other situations where we may build out other incidental types of entities.

Is the scope of the Person entities limited to people represented as authors in WorldCat only or is it open to identity management for academic researchers more broadly?

Answer: In addition to identity management for those in academia, it is also designed to be used in the context of representing persons mentioned as topics or agents in a broad range of contexts, including that of special collections, archives, and the printed record.

Will the infrastructure eventually incorporate timespans as entities? (they seem as integral as geographical entities for purposes of identification)

Answer: We're already modeling events. We haven't been publishing those, but that's something that is definitely on our radar. This is a really interesting subject, though, the notion of Time spans versus events. That's something that we'll be looking at and making sure that we're clear on. I think in some

cases that the relationship is really a relationship between 2 entities, but then that relationship is defined as existing within a certain time span.

Data quality:

I'm also concerned about incorrect clustering. It has a whole downstream effect of bad data (there is a NAR in NACO that got incorrectly merged based on a bad VIAF cluster for example)

Answer: (Nathan) We are building tools to monitor and correct if needed clustering issues. It will look different than the current work-clustering in the WorldCat biblographic catalog.

Good point! How will bad data be reported and fixed?

Answer: (Nathan) Metadata Quality teams at OCLC will be responsible for responding to bad data reports. This mechanism has not been determined but will likely be similar to reporting errors in Connexion and Record Manager or emailing a specific address.

I imagine that some of this depends on quality data in WorldCat, and yet there are more and more vendor-created records that don't follow standards, don't use authorized access points, etc. Is OCLC doing anything to work with metadata creators that are subpar?

Answer: OCLC has a lot of strong relationships with publishers and with yes, I think what someone was referring to in their earlier question was sort of upstream metadata, creators, and providers. Of course, we're looking to leverage those in addition to the sort of third-party application creators that we would want to partner with. We're also looking to our connections with those who create metadata to see if we can get data that would be better suited for this type of work earlier in the process.

How will you vet your clustering algorithms for creative works and the entity above that? What percentage of them need to be correct? Do you see a level of human review after these identifier links are made - perhaps based on size of cluster or allowing users to flag something for review?

Answer: Yes, the UI's that we're building, for SEMI, and for this project, are going to be fairly simple, but we are already embarking on planning for future products. We're going to be building out services and features so that there is flagging. There's viewing of our quality measures, that sort of thing. So, yes, the idea here is that just as we've provided the infrastructure for the last 50 years to do shared metadata work involving bibliographic records and MARC that we will have, not identical, but a different set of tools to hopefully bring that model into a cooperative and community effort to work on this stuff as linked data. To extend what Shane and I have already worked on in terms of native linked data creation. And now it's going to be native linked data management.

... so much depends on the creator of the so-called "canonical work." Is there willingness, or implementation energy, given to a cooperative qualitative effort for record contributors?

Answer: We are working on the shared entity management infrastructure with the goal of providing a basis for cooperative work on the entity data. However, we also need to seed the infrastructure with data to begin, so that hundreds of millions of entities do not need to be created by hand. In the future, we hope to be able to identify metadata that likely needs human intervention and make it prominent to users in order for them to improve it.

Will data be given a public quality score based on the upstream provider?

Answer: Yes. The quality metric is really supposed to be an inclusive score that's very visible when you first pop up an entity, so that you can judge the quality of it at a high level, then dig down as to how the score will actually be done. In theory because you're asking about the upscale or upstream provider. As soon as the entity is within the system, the score would be applied or created immediately, or relatively immediately. If no one goes in and changes that entity, that would be a way to judge the upstream quality of those entities. And so there would definitely be a way to do that. And just for a little bit more background we're looking at things like completeness, disambiguation, currency of the information, timeliness of the information, things like that. As we make this more of a realization, we'll have that information publicly available as to what goes in to creating those different scores.

Other OCLC products:

This is really great! Do you know how this is likely to affect WMS Collection Manager and the Knowledge Base?

Will extant OCLC products employ these developments seamlessly, e.g., Wise/Bicat?

Answer: We aim to integrate this information into OCLC products both extant and future. We'll be working with those products to make sure that that that makes sense. But, yes, the goal here is to have these entities provide a layer of context and consistently available and consistently referenceable set of definitions for what it is that we're talking about, if it's a person or a creative work.

How do you use this API with MARC21? What do we do with authority files?

Answer: The API can be used to look up identifiers to be included in MARC data, or in DC data, or in BIBFRAME data. We expect that cataloging practice will continue to evolve, but the entity management workflow does not replace the work to maintain national authority files. Instead, it can add context and coherence across these different systems in order to improve discovery outcomes for our patrons and users.

Is there support for the API within MarcEdit?

Answer: Not at this time.

VIAF

How are you able to avoid some of the problems in VIAF with incorrect clustering?

Answer: These are services that will allow you to query the entity database and get information about the entities and be able to disambiguate them and work with them. Our vision is to have these APIs available for libraries to work with but also for 3rd parties to implement into their products.

I for one would like to see an easier way of reporting wrong or incomplete VIAF clusterings.

Answer: Anytime, if you see something incorrect about them you have 2 choices: you can send them to bibchange@oclc.org, which is the same email address you use to report cataloging bibliographic errors. Or there's a link on every single VIAF page down at the bottom that says, send a comment, and you can do it that way as well.

Miscellaneous:

Is there a demonstrator service available that we can look at now and get a taster of the user experience?

Answer: Currently, no, we don't have a demonstrator service as we approach the launch date. At the end of the year we'll have more and more information out. I believe that it'll be largely in the area of screenshots, wire frames, and documentation about the data that we'll be publishing, but we're going to be working right up to that date. We want to make sure that we pack in as many features as possible.