OCLC and Linked Data: An update on infrastructure testing and linked data quality

Presenters

- Anne Washington - OCLC Product Analyst
- Laura Ramsey - OCLC Sr. Metadata Operations Manager
- Charlene Chou - NYU Head, Knowledge Access

Presentation summary

OCLC hosted a virtual meeting on 27-10-2021 that highlighted OCLC's recent linked data work and the impact it will have on the library community. During this 1-hour meeting, NYU's Charlene Chou along with OCLC experts Anne Washington and Laura Ramsey shared insights into OCLC's Shared entity management infrastructure project.

Acronyms used during the presentation

- BIBFRAME - IFLA Library Reference Model, the successor to FRBR, FRAD, and FRSAD
- LCNAF- LC/NACO Authority File
- RDA - Resource Description and Access
- RDF - Resource Description Framework
- SEMI - Shared entity management infrastructure
- SNAC - Social Networks and Archival Context
- VIAF - Virtual International Authority File

URLs mentioned during the presentation:

- BIBFRAME: https://www.loc.gov/bibframe/
- OCLC and linked data: https://www.oclc.org/en/worldcat/oclc-and-linked-data.html
- OCLC and Linked Data: The transition to contextual metadata March 2, 2021 webinar https://www.oclc.org/en/events/2021/oclc-and-linked-data-the-transition-to-contextual-metadata.html
- OCLC Research Linked Data https://www.oclc.org/research/areas/data-science/linkeddata/linked-data-outputs.html
- SEMI (Shared entity management infrastructure) http://oc.lc/sharedentitymgmt
- VIAF http://viaf.org/

Member Questions and Discussion

1. **Questions about when and where.**

   This is not in production yet, so stayed tuned for future announcements.

2. **Questions about editing and quality control.**

   What we showed today was essentially view only functionality but in the coming months we'll be putting in the capabilities to edit, enhance and create new entities. The intent is to have this be ongoing so libraries can help edit and maintain them.

3. **Questions about data coming in and how it relates to VIAF.**

   VIAF data is used to populate person entities and VIAF identifiers are being added to them.

4. **Questions about BIBFRAME.**

   OCLC's intent with this project is to make sure that we have a resource that can be used as a reference point for a variety of metadata including BIBFRAME while also making sure it makes sense in the context of other RDF-based systems and MARC data itself.

5. **Questions about Q based identifiers and the usage of Wikibase.**

   Earlier in the project we were using Wikibase to see the linked data principles in action. Then moved away from that architecture but we still needed to generate IDs. So, in this interim period we've been using Q numbers which more or less go back to the original Wikidata Q number. But we will be minting our own. I think they're either 24 or 26 character unique IDs. When you see this in production, you'll fully see another number. We do still, in that external sources area, link to an entity in other systems, but you won't see the Q numbers for entities.

6. **Questions on Q codes.**

   All of the entities will be published with URIs. They will not use the IDs shown in the prerelease images you see here.

7. **Questions about the modeling relating to elements such as birthdate** (which in theory should not be duplicate but in many cases there is ambiguity or there are two values and we can't necessarily judge which one is correct.)

   Those of you who do authority work, like with NACO for example, we know we do have situations where we have two dates that differ and it just cannot be resolved. So in that case I believe we will just have to have a constraint violation in that situation. Hopefully we won't get into any editing wars as we sometimes do with NACO where it's back and forth but sometimes those cannot be resolved without the information to be confident of what date is correct.

8. **Questions about the usage of structured data in terms of describing special collections, rare materials, things that are maybe a little bit outside of the domain of traditional authority files.**

   I think including the special collection and the archive community will help, and I believe the RDA Steering Committee has a new guideline for this. Also, if we follow a lot of standard vocabulary, I think the result will be improved. It's a matter of the data in the record if we consistently use that terminology or follow the guidelines.

   One thing we're in the midst of right now is that the phase of the project is basically building this large data aggregation so of course we rely fairly heavily on established Library Name Authority Files which have a bias toward published materials. But as this project continues to develop and as the database continues to be updated, maintained and curated, we're expecting that a big value will be the ability to represent names and works that may not be well representative in those files. My favorite example is Winston Churchill who is represented in authority files because he wrote a couple of books but he's much more famous as an individual and as a topic than he is as an author.

9. **Questions about SNAC.**

   The SNAC project was built in consultation with OCLC using VIAF and WorldCat data so there is some overlap, but SNAC is a little bit different. It's mission is different, but obviously could be seen as complementary. This project is *really* focused on structured data instead of annotation and prose.

10. **Who can curate, mint new URIs in the system?**

    The vision here is for OCLC member libraries to use the system, and we hope that they will be able to the pull in not just catalogers but also reference librarians, books and special collections, rare books, and subject-specific librarians to make sure that we do represent some of those materials that are not commonly included. Pulling those together, I think is a big potential benefit of this project.

11. **There's a question about why not just use Wikidata and why not just integrate with plenary Wikidata?**

    We believe that there is big value in having a model where this resource is curated by librarians and is used for a specific domain in that it's the same combination that we've seen for 50 years in WorldCat. A combination of professional librarians, subject matter experts, enthusiasts at libraries do cataloging while OCLC Metadata Quality perform enhancements in quality control. We think that model will work well here too. We also see a big benefit of having a centralized data hub as a nonprofit organization that has proven stability and sustainability.

12. **Question about who it's available to.**

    We will be publishing URI data online for everybody to refer to, but we are also considering the editing and writing to be for OCLC member libraries.

13. **Will we begin to see Work entity ids in WorldCat bibliographic and authority records?**

    We plan to add references to WorldCat entities from WorldCat bibliographic records. We wouldn't do that to authority files without more consultation with the owners of those files.

14. **Given that much metadata starts life in vendor systems, are discussions and access to the infrastructure going to include the companies that sell us systems, content, and records? We could really leverage the multiple trading relationships to drive development.**

    Of course, we will be talking to vendor partners about this. Sometimes there are "chicken-and-egg" problems with talking about things before they can be demonstrated, but now that we are able to show a real system, we should be able to have substantial discussions.

15. **How do you see this being used to improve the user discovery experience?**

    We have some folks on the Advisory Group that are particularly interested in what's possible for discovery, so that's an active area of exploration.