2019-03-13 VAOH session

Presentation summary

Hayley Moreno presented on the topic of the Virtual International Authority File (VIAF).

VIAF virtually combines multiple name authority files into a single name authority service and is host to a variety of participants that include national libraries, special organizations, and other data providers.

URLs mentioned during the presentation:

- VIAF website: http://viaf.org/
- General overview of the VIAF program: https://www.oclc.org/en/viaf.html
- General overview of managing ambiguity in VIAF: http://www.dlib.org/dlib/july14/hickey/07hickey.html
- Download the VIAF dataset: http://viaf.org/viaf/data/
- Reporting issues with clusters or incorrect information: bibchange@oclc.org
- Reporting website issues or general issues not related to clusters: oclcviaf@oclc.org

Member questions

Where exactly do you find the choice "All VIAF" in the drop down? I'm looking above the cataloging tab and the authority tab and I'm not finding that choice listed.

Answer: If you go to VIAF.org, it is on the top of the page,

Does VIAF have a mechanism for ingesting ORCID numbers?

Answer: Right now, if an ORCID number is in the 024, we do use it to match between source records. We don't ingest ORCID records at this time because the last time we tried it, the match rate was insufficient. The ORCID data that's in the public download doesn't enough dates and titles, so we weren't making good matches.

In Record Manager, I think you have the Dutch and German authority files available for catalogers? Is that part of VIAF?

Answer: Yes, there is a Dutch and German authority file available for catalogers in Record Manager. VIAF and Record Manager are receiving some of the same source data but they are not talking to each other and are processing the data differently. Therefore, what you see in Record Manager is not exactly the same as what you will see in VIAF.

Does VIAF have any rules about how geographic names must be formed?

Answer: We try not to have any rules about how names are formed and just use what's sent to us. Which means there is quite a bit of forgiveness in our matching algorithms, because we realize that different national libraries have their own rules. We try to work with what we get. That's why those clustering numbers for geographic's tend to be lower, and tend to be what we call singletons (just means it's not clustered with anything else) because there are various rules that are used for geographics. They can be a little more complicated, as well as corporate names.

You said "identifiers are persistent for the most part", what are cases when identifiers change and what is the percentage?

Answer: The clusters get rebuilt every week. The membership within the clusters as records are added and deleted, and occasionally because the algorithm has changed. So, if two clusters become merged, the one of the VIAF IDs lives on and one becomes abandoned. Although if you had a link to the one that's abandoned, it will always get you to where you need to be. We always redirect old links to where they should go. We make every effort to not have dead links, even though we do necessary maintenance on the clusters themselves. We calculated the percentage for a couple of specifics, so like LC and ISNI recently, and it was in the 99% range of the IDs in those sources were in the same cluster a year ago that they are today.

Is there any communication between OCLC and contributors when contributors data is causing algorithmic errors?

Answer: Yes. If there is any information that is inaccurate in the entities VIAF page, we get that information from our contributors, so we need them to fix it in order for the next data harvest to get that correction in.

What redirection happens for cluster splits? Does it redirect to just one?

Answer: If a cluster splits that has four member, and splits into two clusters with two members, two of them are going to keep the old ID and two of them are going to get a new ID. So if you had a link to the old ID, it's going to take you to the old cluster that now only has two members. You can also save links to specific processed records, and that will always take you to the cluster containing that record. So if you particularly care where the LC record is or where the BnF record is, you can save a link to a particular BnF record and when you click on that link it takes you to the cluster containing that record always, regardless of where it's ended up.

Can you tell what kinds of traffic are coming into VIAF via the permalink ID? Is there evidence of the volume of automated calls on VIAF data?

Answer: Yes and yes. Another one of our team members runs Google Analytics on VIAF, and he can tell who's been downloading things, who's been searching, and how many searches are from people, and how many searches look they are from bots. About 75% of traffic is coming via permalink IDs in VIAF. Our conservative estimate is that at least more than half of the transactions done with VIAF data are by bots.

Do any end-user systems make use of VIAF?

Answer: We're using the VIAF IDs. They are certainly forming our data link projects. They're going into the prototypes we've been doing for linked data projects. Jenny is also on the team that does the FRBR clustering, and the VIAF IDs definitely play a big difference in the way the FRBR clustering works.

What redirection happens for cluster splits? Does it redirect to just one?

Answer: In a split, there is really not a redirect. Some of the cluster members will keep the old ID and some of the cluster members will get a new one. So there wouldn't be a redirect with a split.

How often does LC load to VIAF?

Answer: Weekly.

How quickly do staff respond to error reports?

Answer: We do get a large volume of VIAF requests. We do process them first come first serve. We do have a little bit of a backlog, but we are working through it. It's interesting just how many people use VIAF and actually, many are not in library-land. There's actually a lot of folks that edit Wikipedia articles that notify us with issues of clusters, as well as author's themselves are just general users. We will check them and get through them eventually.

Do you have any plans to work with subject authority files or other files (genre/form, etc.)?

Answer: Not at this time.

Does OCLC have any thoughts about the use of VIAF URIs in LCNAF authority 024s? Is it useful? Could/should the field be automatically populated based on LCCN presence in a VIAF cluster?

Answer: VIAF will use that. If you put the VIAF ID in your 024, we use it for clustering. 024 fields that have different control numbers do help to create suggested links. So that's definitely helpful for our side in clustering. On the flip side being the person who uses that, if you don't maintain it and make sure it's right, it can get to be a problem eventually. If you make these links you got to curate them. It was mentioned that there is a moratorium on adding 024s in authority records right now for NACO.

Has DNLM headings taken the place of MeSH headings? I came across a record that had DNLM headings but it didn't have MeSH headings in the record.

Answer: A bibliographic record that has NLM as the source of the record and for some reason it does not have MeSH headings on it, would be very unusual. Normally, you would expect to see MeSH headings there, so 6xx fields with a second indicator 2, whether it's a 650 topic or a 655 form/genre heading with a second indicator 2. If you have a record like that, that looks like it's from NLM and is lacking those kinds of headings, send it to us so that we can investigate what's going on. If something stripped those headings off, we would be concerned and would want to take some action on it. If it was just issued that way, we could maybe figure that out.

(Laura) Maybe these are the headings that appear to be MeSH headings because they have second indicator 2, but did not appear to be coming from NLM? So there was no DNLM indication in the subject heading itself. If that is what the question is about, we are aware of it and cleaning them up.

(Robert) The record that was at the heart of this question appears to be a record that originally came from a vendor and then has been upgraded with a bunch of different symbols, the National Library of Medicine not being one of them. So even though it has medical subject headings that have been assigned, they look like they came from one of the libraries that contributed to this record rather than from the National Library of Medicine itself.

If PCC libraries start putting URIs in 024 rather than numerical identifiers, will you be able to use those?

Answer: Yes.

If we report an error, do you send that information back to the original source institution/database?

Answer: If clusters get modified or changed, we are not letting or contributors or participants know because they're downloading the data. So those changes will eventually be found in the data file which is actually updated monthly, unlike the live database which is updated weekly. So they will get those changes. What we do report back out to our contributors is if there is incorrect data coming from their authority file or their bibliographic file. So if a work is incorrectly associated with an author, we will notify the institution that is contributing that incorrect data because until they fix it, then it's not going to be reflected in VIAF.

If the only LC call number in a record is very general (for example NA680 for a non-general architecture book, or just PN for literature), can that be replaced with a more specific one? Or should we just add a second more specific one?

Answer: If you are able to replace the record and there is a legitimate, more specific classification, you should replace the existing one with the more specific one.

Do any digital assistants like Siri, Alexa, Google, Cortana, or Bixby use VIAF data?

Answer: We are not sure.

If we find a controlled heading in a bibliographic record that is not the correct entity and change it to the correct entity and control it, will that eventually trickle down to VIAF so that the work is not associated with the incorrect entity?

Answer: No, that data does not flow in that direction. The controlled headings in WorldCat affect the way Identities pages are built so if you control it to the right name, the Identities pages will be corrected. The works don't flow directly into VIAF. The works that you see on VIAF.org for entities are coming from bibliographic files that are sent to us from the VIAF participants. They are not works that come from WorldCat. The WorldCat Identities, which is another project, that data is being harvested from WorldCat. For VIAF, those works are coming from bibliographic files of our participants, and not all of our participants have records in WorldCat. Many of them don't, and many of them send their records directly to us.

Can you take advantage of 672 and 673 fields in authorities to help with matching? For clarification, the 672 and 673 fields have control numbers unlike the 670 field.

Answer: We don't attempt to do that now, but we will investigate it further. We do pull some data from 670. We look for some very specific format citations and patterns to find titles or to find birth dates from title pages. Sometimes there is a birth date that came from the title page of a book, and we have some real specific patterns that it can look for but that's about it.

Are we allowed to use 386 fields in bib records? If yes, are they indexed in WorldCat?

Answer: Yes, it is part of the entity attributes and can be searched by en:.

What kinds of bibliographic records are sent to you as part of NACO?

Answer: Those are coming from the Library of Congress. The LC NACO file is a cooperative, there are many libraries, but the bibliographic records that are associated with any LC NACO records are just coming from LC and not from any other contributors.

Is any work being done to reduce the circularity between VIAF and ISNI? Sometimes badly-clustered VIAF data is used as a source in ISNI and then that ISNI source data reinforces the VIAF cluster. It becomes very hard to resolve errors.

Answer: We agree. At this point, we don't believe that ISNI has ingested VIAF data in probably a year. VIAF is continuing to ingest ISNI data. Otherwise, it kind of goes back to once you make a link, you got to nurture it and take care of it and make sure it's still good.

Does the Canada grouping include both English and French names for the same entities? With LAC contributing through NACO for English names, is there going to be duplication with LC and Canada?

Answer: When that whole transition was occurring, we did the work and now any LAC authorities that are English are in LC NACO and showing up there. The only LAC authorities that you would see in VIAF are from the French authority file.

Does OCLC have any plans to index the 008/24 Nature of Contents field in bib records in WorldCat? I am asking because it has the value 6 for comic books/graphic novels that can be hard to limit to in a WorldCat search.

Answer: We have no plans at this time, but can certainly be looked at and investigate to see if it is something that we can take on.

LC is a contributor to VIAF and VIAS is a resource for LC authority creation. What future relationships do you anticipate between LCNAF, the OCLC authority file, and VIAF?

Answer: At this time, the relationship between all of us is working and don't see that changing at this point in time. Right now VIAF is in a transition and it continues to change little by little, but at this point in time the relationship we have with LC and the other participants is pretty stable and don't see it changing.

Can you give specifics for the "weekly" reclustering? Is it always the same day? I'm thinking of checking back to confirm a change has "worked", etc.

Answer: If everything works beautifully, the clustering finishes sometime Monday morning and it's visible by Tuesday morning. Occasionally there are issues and it's later in the week, but it starts on Saturday afternoon and it's usually done Monday morning and usually visible by Tuesday morning. Please note that if it was a request that you sent and it hasn't been processed yet by WorldCat Metadata Staff, then that change is not going to appear because that change needs to be done by staff manually.

How does OCLC calculate the percentage of clustering errors?

Answer: We pull a sample of 100 random clusters and check them.

Most clusters are very small. Do they weigh equally in the cluster error measure?

Answer: Yes, and we did look for over-clustering and under-clustering.