2018-09-12 VAOH session

Presentation summary

The WorldCat Metadata Quality staff introduced themselves and an overview was given of each of the three sections that make up the Metadata Quality division at OCLC.

Metadata Quality is divided up into three primary sections: Metadata Policy, WorldCat bibliographic database (quality control, working primarily with bibliographic records and also authority records), and WorldCat knowledge base and registry.

- **Metadata Policy** works on setting policy for all of WorldCat, including both bibliographic and authority records with some involvement in Local Bibliographic Data (LBD) and Local Holdings Records (LHR). They are responsible for implementing policy within the WorldCat bibliographic database, and assuring that the database is updated when needed. The staff also represent OCLC to outside groups at ALA, IFLA, PCC, and other cataloging communities and are responsible for answering a majority of the questions submitted to askqc@oclc.org.
- **WorldCat Quality** improves, enhances, and enriches the quality of the WorldCat bibliographic database. Staff run multiple sessions of Connexion client to correct and enhance bibliographic records. OCLC staff as a group modify an average of 10 to 15 million bibliographic records a month. Staff also merge duplicate records in all formats and are involved in the Member Merge Project, an outreach program that trains select member libraries to merge duplicate bibliographic records. WorldCat Quality staff resolve user requests for corrections to bibliographic records, and support libraries who do not participate in NACO by creating and modifying name and series authority records.
- **WorldCat knowledge base** resolves issues relating to linking and holdings information, primarily for electronic resources, within the WorldCat knowledge base, enabling library to manage e-collections, provide access to resources, and receive MARC record delivery. The **WorldCat registry** contains information regarding institutions, such as the library's address and links to the catalog. Maintaining accurate information in the registry facilitates, among other things, access to library holdings in WorldCat.org.

Member questions

General Questions

If a provider neutral record adapted in our local catalog with an 843 for specific publisher and holding, an 843 is converted to an 853 as part of our data sync process, will that create a new non-provider neutral record.

Answer: We suspect that it probably will, given the way we process information in the 533 in our data sync processing. We pay attention to publishers that are in the 533 so that, for example in the case of microforms with different publishers, we wouldn't want to merge them together. So the fact that we

have a record that has a 533 that would include what looks like a publisher versus another record that doesn't, we would probably end up adding that at this point. We also have macros that we run on occasion to go after records for certain online resources, to make them provider neutral. So, it's possible that a 533 that once appears in the database might be removed as part of the process to make it provider neutral, and then the record could be subject to being merged after the fact through our duplicate detection algorithm.

Do you have a preference as to how we report duplicate records (online form, report error, etc.)?

Answer: Whatever method works best for you and fits into your workflow is fine.

Is there any clean-up plan for local headings and headings with unknown or invalid sources?

Answer: There aren't any specific plans, at this point, to clean up local headings. We encounter local headings as part of all the other work that we do, and we look to make sure that they are correctly formulated. So, if we have a pattern of some local heading that is problematic in that respect, we will sometimes go after it and fix it up. In some cases if we are encountering multiple forms, staff will do the additional step of establishing an authority record for that name. We know that a lot of local headings have been entering the database through our data sync processes, and when those are duplicating subject headings already in the record that are not local, we do have a macro established that delete those headings. We have not taken a systematic approach, but when we do encounter those we are deleting the local headings. We also have in process a correction to the way data sync works, so that few of these local headings will transfer from incoming records to master records.

What is the status of FAST headings? Is there any public interfaces that use them?

Answer: We don't know of any public interfaces that are using them (or specific ones), but they are being used by various institutions. Nathan said that he would have to talk to Jody DeRidder, who is overseeing the FAST process to figure out what specifically they are. We are in the process of creating an editorial board for FAST so that we can go through the process of updating them or adding terms, or various things like that, and not be solely dependent on the conversion of the LCSH. We have some announcements in the works that will go out in the next few weeks, or sometime in the future about those sorts of things. We are looking at creating a sustainable future of the FAST headings.

We've had experiences where we've manually corrected broken diacritics in OCLC records, only to have them return. Can you give us an update about this situation? Is manually correcting these fields when we find them the correct thing to do?

Answer: We presume this is referring to the situation where the character typically shows up as a black diamond with a question mark. Fixing them is certainly an appropriate thing to do. That would involve deleting the fields because they are often duplicates of fields that are already there. That's because the character is considered a non-Latin character within the OCLC database, and that is partly the reason they have transferred in. We are working on the root cause of this problem before we take an approach to finally cleaning all of the up. We have, on occasion, gone back to get rid of large groupings of these but find that in some cases they transfer back in again. So helping us by cleaning them up when you

happen to see them is a good thing. Nathan added that the time frame for the fix thing is within a couple of months and hopes that with the October or November office hours to state where we are in the development work for that. The biggest thing is that we don't want them to continue coming in.

When records are merged together in error, what is the process of reporting this and how are these errors resolved?

Answer: A request can be submitted to the Bibchange inbox (bibchange@oclc.org) to let us know that an error in merging has occurred. As long as the records were not merged prior to 2012, they can be recovered. Once the records have been pulled apart, we can test them to see if subsequent changes to Duplicate Detection and Resolution (DDR) may have taken care of the problem that allowed them to merge in the first place. If DDR would still merge them for one reason or another, we can often work with the reporting institution to come up with a way to prevent DDR from merging them again. As we have mentioned in previous sessions, our de-duplication process is continually evolving. As we stumble upon an incorrect merge, we do go back and test it and change the algorithms that are merging it in hopes of preventing future cases like that.

Follow-up question: Are records recoverable since 2012 or 6 years before the present?

Answer: As of right now, it's since 2012. This date is fixed, so in 2020 we will be able to go back to 2012. As we look at data retention and the size of this file, because there is a lot of stuff in the journal history file that is kept, this decision may be reversed or changed or shortened sometime in the future.

We are in the transition to move to WMS, for authorities that are not in NACO or headings that are different from headings in bib records, how will OCLC control them?

Answer: Right now, within WMS, there is controlling for multiple sets of authority records. NACO being the prominent, and the one most used with English language of cataloging records. There is also controlling for LC subject headings, for MeSH subject headings, Maori subject headings, Dutch names (mostly used on Dutch language of cataloging records), and German names (mostly used on German language of cataloging records). Later this year we are going to be implementing a French language authority file from Canada. That will be controlling names in French language of cataloging records. If something is not in NACO, it will not be controlled. If an authority record is needed for a particular heading, where one does not exist, one of the things that OCLC can do is create an authority record for that heading. Requests for new headings can be sent to authfile@oclc.org and WorldCat Quality staff handle those requests.

If a library's records are provided to a vendor and the vendor distributes the records, why does the 040 subfield $a and subfield $c reflect the loading library? Both the vendor and the WorldShare library have ignored the MARC conventions for subfield $a and subfield $c.

Answer: This is an OCLC thing, not a vendor thing. If a vendor sends us records on your behalf and it's loaded under your symbol or a collection created for your library, our software here at OCLC (DataSync software) changes the 040 subfield $a and subfield $c to your OCLC symbol. This is a decision that was made when the DataSync system was being programmed.

Follow-up comment: There are records provided by a national library for MARCIV. Now we are looking at our records, and rather than match/replace our records, the master records are match and attach.

Response: These are reasons why we would want to pay attention to the subfield $a and subfield $c in the 040 field, so we will take that into account as we continue to look at that system.

Can you provide details on which tags can be edited or deleted and which organizations have the permissions? Specific example: 015 field.

Answer: There are edit restrictions that are built into the system for certain fields and in certain kinds of records that would prevent somebody from replacing a record after making certain kinds of changes. There is no edit restriction on field 015. Libraries can add them or delete them as they see fit. It seems pretty unlikely that an 015 field would be deleted, if it's legitimate. And in this case, since we're talking about records that have come from Library and Archives Canada, I would imagine that your 015 fields are safe.

What are the points used in duplicate detection? The records I have seen have the similar title but different authors.

Answer: There are roughly two dozen comparison points for bibliographic records in DDR (Duplicate Detection and Resolution). That is misleading, in the sense that many of those comparison points actually draw from various parts of the bibliographic record and not simply one field. In many cases, the information gets manipulated in order first to see if things are the same that are transcribed differently or look different to see if they are actually considered to be the same thing or two, if they appear to be the same but really are different. There are roughly 300 fields possible in a MARC bibliographic record and there are roughly over 200 fields that we look at or otherwise consider. Most of those are the things that you would expect such as the title (245 field), places of publication, publishers, dates, series. But there are all sorts of other things where a comparison point is specific to a particular kind of bibliographic record. For instance, scale in Maps records, publisher numbers in Sound Recordings, and in Scores various elements of instrumentation. So there are lots and lots of comparison points. Jay has done a defensive cataloging presentation that helps you know which fields play in, so that if you wanted to create a record it's not merged into another record. We can maybe do a session sometime in the spring about what our merging process is.

Is there any way there could be an option to not include other national bibliography authority headings in the browse headings function? This would be similar to the limit by language of cataloging in the search function, but in the browse function instead.

Answer: This seems like a very worthwhile request. What we have in place, in terms of browsing headings in the bibliographic file, is that all languages of cataloging are included with all their headings integrated into the same index. So, sometimes you'll see variations in names that are legitimate names because one is the form that's used by the Germans, the other is the form that's used in English language cataloging and so there's a difference in qualifiers. Or, if it's personal names, one will have a date and the other one doesn't. It's a little bit confusing as you're looking at that display. I presume this is what you are asking about in this question, as opposed to searching related to an authority file where

you normally just pick the file that you want to search in. In browsing through headings in the bibliographic file, they are all mixed together. This is something that we ought to keep in mind for Record Manager, because it would be a desirable thing to have.

At Library and Archives Canada for CIP records, do we have to take the vendor records and enhance them or can we create new records?

Answer: Since you will be using WMS and WorldCat is your database for the Library and Archives Canada, we would hope that you would enhance what was there and not create a duplicate record. In many cases, there won't be a record already there when you are doing Canadian CIP.

Is anything being done about subject headings order being changed to field number order rather than the intentional order of the cataloger? Also including second indicator zero is taken a priority if that is how the cataloger ordered them.

Answer: This is a known problem in data processing. It is on a list of things to resolve, along with other things, but we are not sure where it ranks. It is a problem in that tag 600 is automatically going to float to the top rather than the 650 that a cataloger put there intentionally.

When trying to add the subject heading Shīʻah $z Lebanon $x History $x Press coverage, controlling moved the Lebanon piece to the end which of course isn't the same thing. Do we need to make a subject heading for validation purposes, can OCLC do this, or is that only LC?

Answer: As an LC subject heading, that would be only up to LC to establish a subject authority record for it that we could potentially control to. The re-positioning of the geographic is determined by what can be subdivided in terms of the subdivisions that were here. So, this heading had History and Press coverage, presumably Press coverage is a subdivision that can be subdivided geographically which is why Lebanon was moved to the end in this case.

Knowledge base question: When OCN chosen as override OCN in the KB is DDR'd (merged) in the WorldCat bibliographic database and one is chosen over the other, what happened to our 'held by' in Discovery? It seems to disappear or not appear on the remaining OCN. If it requires an action on our part, is there a way to be alerted?

Answer: Right now when you merge two records and the OCN that is in the Knowledge Base is not the OCN that is retained in the bibliographic database, there's a brief disconnect period where we have to wait until the OCN gets updated in the Knowledge Base. That will happen automatically eventually but if you need it to happen much sooner, then the OCN needs to be updated in the Knowledge Base. We are working on making that much more streamlined because, it doesn't do any good to not have the correct OCN in the Knowledge Base. At the moment there aren't any notifications as to when records are merged, but we should probably look into that more.

If two records were merged possibly not by an automated process but an actual person, can the records be recovered back to 2012 as well?

Answer: Yes, they can. It doesn't matter the process they were merged, as long as it happened after 2012 we can have them recovered.