Notes from the OCLC Cataloging Community Meeting on 3 February 2023

Access the recordings and presentation slides

Session: WorldCat Quality Update

Presenters

- Shanna Griffith, Senior Data Analyst, OCLC
- Robin Six, Senior Data Analyst, OCLC
- Nathan Putnam, Director, Data Quality, OCLC
- Jay Weitz, Senior Consulting Database Specialist, OCLC

Presentation summary

Shanna Griffith reviewed the upcoming Virtual AskQC Office Hours (VAOH) for the next six months.

- February’s topic is Bibliographic Formats and Standards (BFAS), the early chapters.
- March’s topic is Languages, non-Latin scripts, and the mysterious MARC 880 fields.
- April’s topic is Data and algorithms and bibs, oh my!
- May’s topic is What’s in a name? Descriptive access points overview.
- June’s topic is Cataloging rare materials defensively.

The dates and times are listed on the AskQC help site, along with previously recorded sessions and session notes. Each monthly topic is presented twice and is followed by Q&A regarding the topic or cataloging questions in general.

Nathan Putnam and Robin Six spoke about OCLC’s data modeling labeling project.

Nathan started with insight into OCLC’s work to develop a machine-learning (ML) model to identify duplicate records in WorldCat. The model needs source data to teach it to make appropriate decisions. OCLC created a tool to allow the library community to help validate the model’s understanding of duplicate records. He then presented stats from the tool.

- Top three contributing institutions: Cleveland Museum of Art, Indiana University, and Salisbury University.
- Top three individual review outputs: 4,136 pairs, 1,215 pairs, and 1,140 pairs
- Top five MARC fields used for decision-making of duplicates: 245, 300, 260, 100, and 264
- A graph of daily outputs, with a one-day high of over 700 record pairs being reviewed.

To finish, Nathan highlighted how community contributions have already affected the ML model. Using results from the data labeling tool, the Data Science team found new insights into the feature creation process. The original model assumed that different physical descriptions in the
300 field led to different items. However, in 75% of these cases, the community labeled them duplicates.

Robin presented a brief demo of the tool and described its functionality. She described the tool, its purpose, how many pairs were used, and how long it will remain open. OCLC identified a random sampling of 20,000 duplicate record pairs with the goal of achieving three answers from three different participants for each pair. Participants can choose pairs according to language of cataloging, material type, and record age. Two records are presented to the user, with field differences highlighted in yellow. Participants can select which fields were relevant to make their decision and indicate whether or not the pair is a duplicate. Having three reviewers for each pair will allow us to gather a consensus. The tool will be open through February 28th.

Robin concluded her portion by encouraging attendees to try the new tool and compete with other dedupers.

Jay Weitz summarized the recent and “thrilling” updates to Bibliographic Formats and Standards. Weekly meetings this past year have focused on providing additional examples, integrating MARC bibliographic updates no. 33 (November 2021) through no. 35 (December 2022), adding sections on punctuation, revising and expanding guidelines for individual pages, such as the Dates Fixed Fields, and adding editorial enhancements and improvements.

**URLs mentioned during the event**

- AskQC help site: oc.lc/askqc
- Bibliographic Formats and Standards: oc.lc/bfas
- Label Duplicates Tool: https://labelduplicates.worldcat.org/
- Label Duplicates Tool FAQ: https://www.oclc.org/content/dam/oclc/worldcat/data-labeling-faq.pdf
- Label Duplicates Tool Participation Instructions: https://www.oclc.org/content/dam/oclc/worldcat/data-labeling-participation-instructions.pdf
- MARC validation release notes: oc.lc/validation-release-notes

**Audience questions**

Q  Do you consider including specific fields when checking for duplicates?

A. The way that the machine algorithm works is that it compares the entire record. One of the unique things about using artificial intelligence and machine learning is that we don’t have to specify every rule set.

Q  What is the link to the BFAS webinar registration?

A. You can find registration information at oc.lc/askqc or https://help.oclc.org/WorldCat/Metadata_Quality/AskQC/Current_AskQC_office_hours.
Q I really like the examples and punctuation in BFAS! Do you consider suggestions for additional examples?

A. Yes, please let us know by contacting askqc@oclc.org.

Q Is there any plan to upgrade so I don’t have to enter subfield $e and subfield $4 in 100 and 700 fields? Will the system be able to generate the natural language role based on the subfield $4 and the language of cataloging?

A. There are no current plans, but we recognize the need for the code and human-readable forms, especially as we transition to linked data like BIBFRAME or other entity infrastructures.

Q Does OCLC regularly control LC Name & LCSH headings with $0 (whether single or multiple instances of subfield $0 in a single field)?

A. Currently for LC headings, subfield $0 is not sued and instead controlling works via the text string. Other authority headings, like those controllable in Record Manager, use subfield $0 and only use the text string when subfield $0 is not present.

Q When you find duplicates, how do you decide which record to keep? If you merge, how do you decide which one is the preferred record? What if we find a merged record where it would be better if a non-preferred record should be the preferred one? Is there a way to report that?

A. There is a hierarchy we use to determine the retained record. You can report questionable merges to bibchange@oclc.org.

Q When did the data labeling project become available to users?

A. The data labeling project was first rolled out to Member Merge Program participants in November 2022 and then to all community members in December.

Q Will MeSH subjects ever be controllable? I heard it was in process, but have not seen anything official.

A. MeSH is currently controllable in Record Manager. In Connexion, you will see a subfield $0 that contains the MeSH id in controlled fields. So even though Connexion doesn’t have the functionality to control all of the different vocabularies that Record Manager can, you still see the results in Connexion using subfield $0.

Q When a bibchange request is sent, what is the expected time frame for the change to be made?

A. It depends on the type of request and the volume of requests we receive. Generally, any needed changes are made within a week. If it’s a complex situation, it may take longer. Duplicates are different and go into a different workflow, so the time it takes to complete requests for duplicates depends on the backlog and the number of staff available to merge in that format.
**Q** Who to contact for Member Merge Program?

A. Please send a request to askqc@oclc.org.

**Q** Is it ok to control in Connexion, even if it wipes out the subfield $0 and turns it blue?

A. The controlling software will wipe out the subfield $0 but controlling headings, when available, is the preferred option. If the vocabulary is from the LC Authority File, that subfield $0 is maintained in the system behind the scenes. One thing to note is that subfield $1 will not get wiped out when controlling.

**Q** Should corrections and merges be included in the same Report error message in Connexion?

A. You may report these however it's convenient to you, whether it's the report an error drop-down option, the report an error form, or emailing bibchange@oclc.org for bibliographic record changes or authfile@oclc.org for authority record changes.

**Q** Could auto-generated RVM be allowed to be deleted so they could be repopulated, like with FAST?

A. Unless the RVM headings are incorrect, we ask you not to delete them. RVM headings are either added manually or generated via the enrichment process. Enrichment adds RVM headings on a 1-to-1 basis when (1) an LC heading is present in the record and (2) there is an RVM heading equivalent to the LC heading. Catalogers cannot tell by looking at the bibliographic record whether the RVM heading was manually assigned or added via enrichment. Instead of a 1-to-1 correlation, FAST headings are generated by evaluating all LC headings present in the record.