Virtual AskQC Office Hours

Medley of popular topics

OCLC Metadata Quality

August 2021

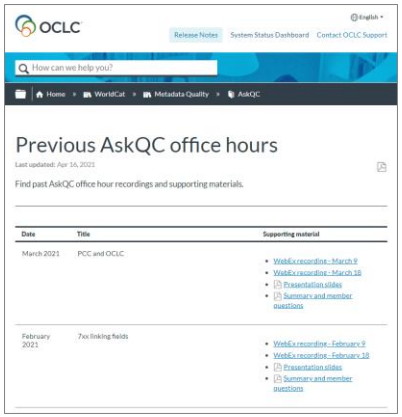
Housekeeping

This session is being recorded

Housekeeping

This session is being recorded

All session recordings, slides, and notes are available at oc.lc/askqc

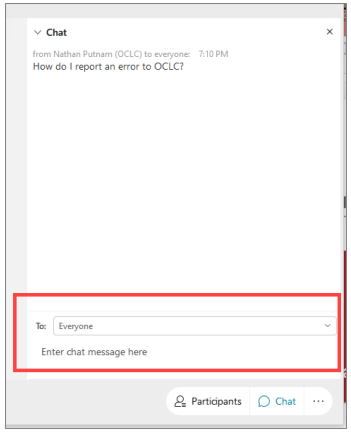


Housekeeping

This session is being recorded

All session recordings, slides, and notes are available at oc.lc/askqc

Enter questions in chat to "Everyone" at any time during the presentation



Housekeeping

This session is being recorded

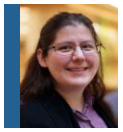
All session recordings, slides, and notes are available at oc.lc/askqc

Enter questions in chat to "Everyone" at any time during the presentation.

After the session, you will be directed to a brief, optional survey

On the call today

Robin Six
Database Specialist



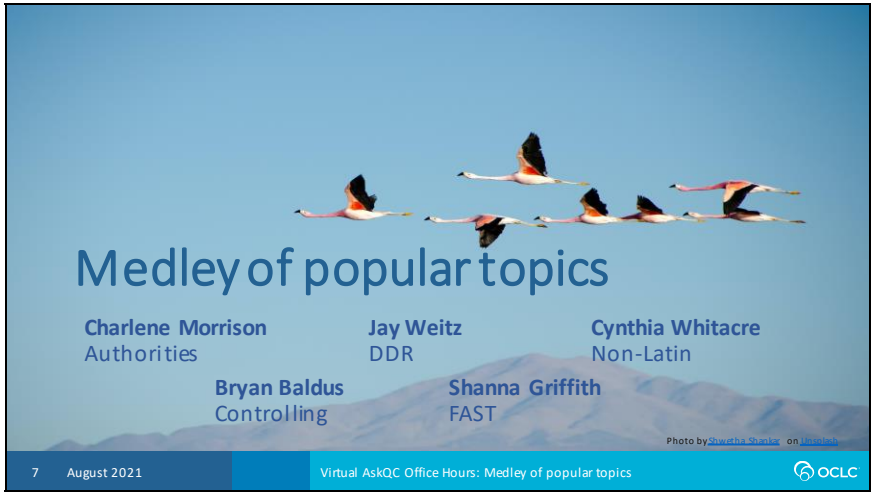
Charlene Morrison
Database Specialist

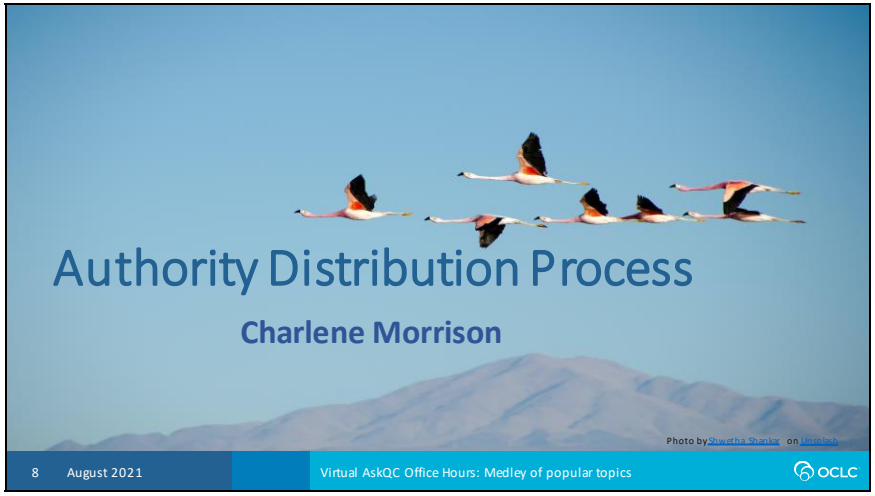
Bryan Baldus
Consulting Database Specialist

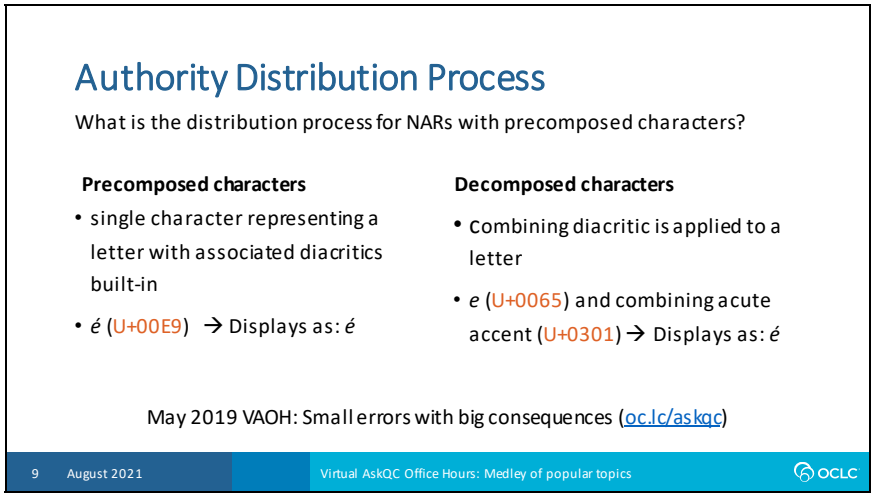
Jay Weitz
Senior Consulting Database Specialist

Shanna Griffith
Database Specialist

Cynthia Whitacre
Senior Metadata Operations Manager


Medley of popular topics
Charlene Morrison
Authorities
Jay Weitz
DDR
Cynthia Whitacre
Non-Latin
Bryan Baldus
Controlling
Shanna Griffith
FAST
Photo by Shwetha Shanker on Unsplash
7 August 2021
Virtual AskQC Office Hours: Medley of popular topics
OCLC


Authority Distribution Process
Charlene Morrison
Photo by Shwetha Shanker on Unsplash

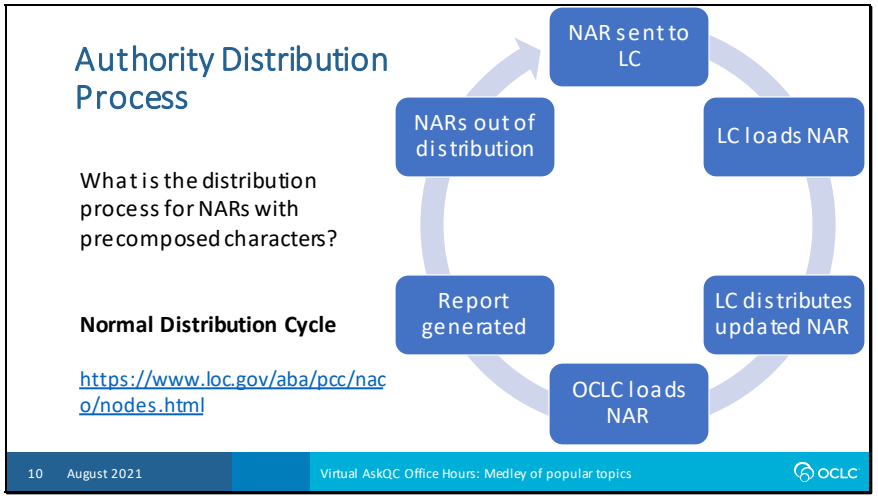


Questions about the distribution of authority records comes up from time to time including questions about processing precomposed characters in the WorldCat authority file. To answer this question let's first look at the definition of precomposed versus decomposed characters and why precomposed characters are problematic.

A precomposed character is a single UNICODE character representing a letter with associated diacritics built-in, while a decomposed character applies the combining diacritic to the letter. Both display the same to the user but precomposed only has one UNICODE character, while the decomposed has more than one.

The problem this poses is that precomposed characters are rejected by the LC/NACO authority file and they affect how indexing works within the WorldCat authority file.

My colleagues Bryan Baldus and Robert Bremer talked about these character in their May 2019 VOAH webinar entitled, Small errors with big consequences, if you wanted to check it out to give you more information.

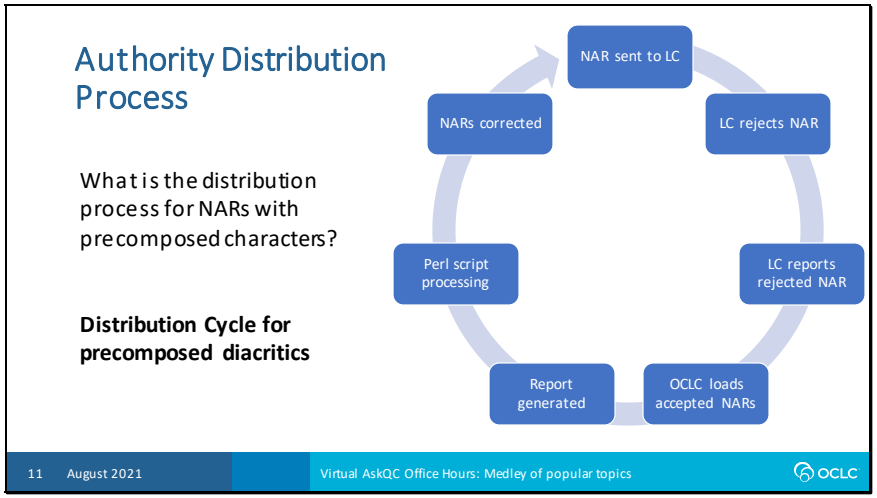


Now let's take a look at a normal distribution process.

In a normal distribution process, an authorized NACO participant updates or replaces a name authority record (NAR) in OCLC's copy of the LC/NACO Authority File, and at the end of the day these NARs would then be sent to the Library of Congress (LC). After the records are loaded, they are then distributed to all of the NACO nodes. OCLC then loads the distributed file into the WorldCat authority file. This portion of the process usually take about 48-72 hours or 2-3 days to complete from the time that the NAR was originally added or updated to the time that it's returned to the WorldCat authority file. At this time, the NARs should fall out of distribution and be available for further editing as needed.

I've included a link to from the PCC website that explains the NACO distribution process and

NACO nodes. I recommend checking that page out if you are interested in learning in greater detail how the process works.

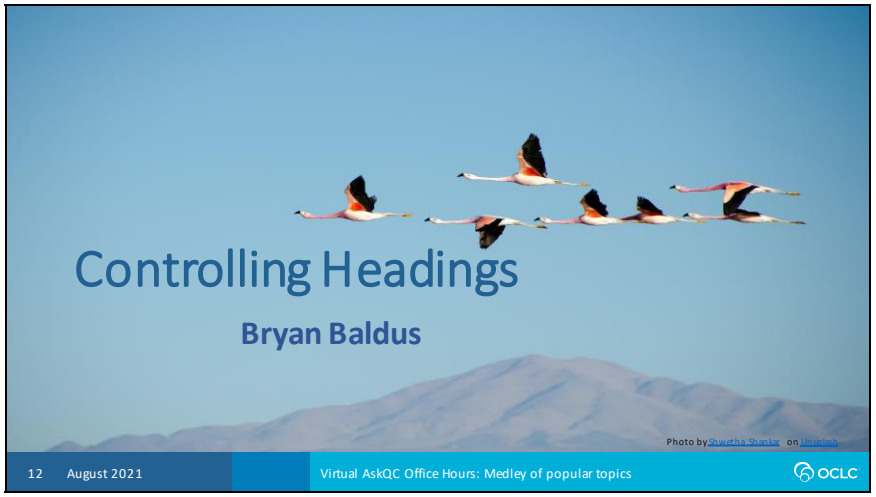


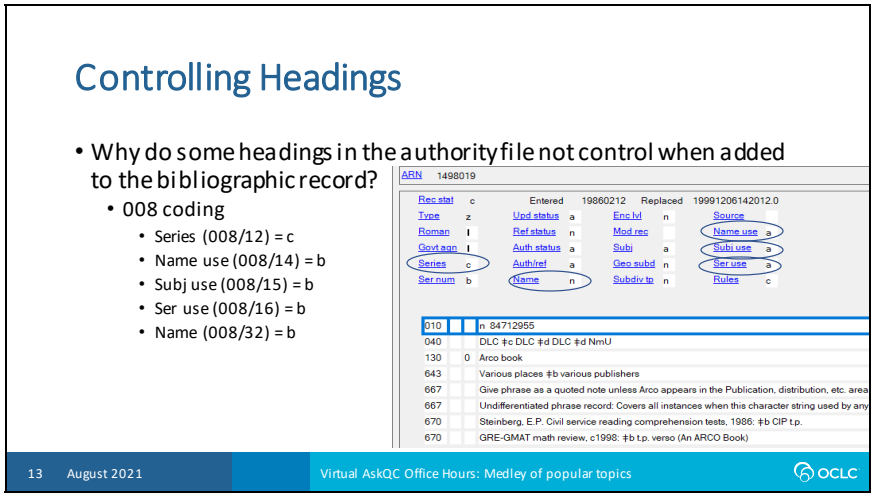
For those NARs that have precomposed characters, the distribution process remains the same but with a few additions.

The normal process takes place where the NAR is sent to LC, LC loads the NAR and updates their authority file, and then re-distributes the file back to OCLC. Here is where the difference comes into play. When a precomposed diacritic is sent in one of the NARs to LC, LC's system will reject that added or edited NAR. This means that that specific name authority record will not be updated with the change in the LC/NACO name authority file. When all of the records are sent back to OCLC a report is generated that identifies the NARs in the file as accepted and rejected by LC. Metadata Quality staff process the report through a Perl script to pull out the rejected NARs to process. Once the corrections are made, the NARs are sent back to LC to go through the distribution process again, which can take another 2-3 days. Once OCLC received the corrected records back as "accepted" from LC, then they will fall out of distribution and be available for editing.

One thing to note is that the distribution lock stays in place because the LC/NACO name authority file was not

updated with the changes for those rejected NARs. As long as there is a difference between the WorldCat authority file and the LC/NACO name authority file, the NARs will remain locked in distribution in the WorldCat authority file. The reason for this is that the LC/NACO name authority file is the official authority file while OCLC's WorldCat authority file is a Node or copy of that file.

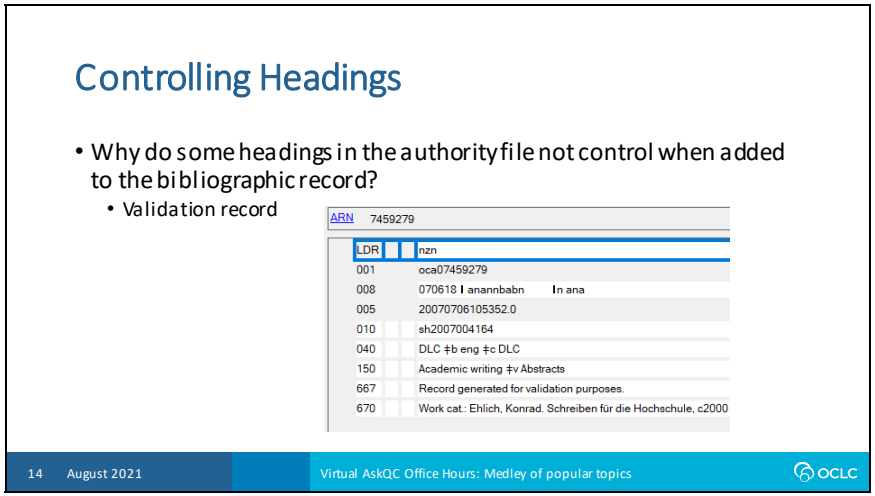

Controlling Headings
Bryan Baldus
Photo by Shwetha Shanker on Unsplash



Thank you, Charlene. Hello, I'm Bryan Baldus, and I'll be discussing a question related to controlling bibliographic headings.

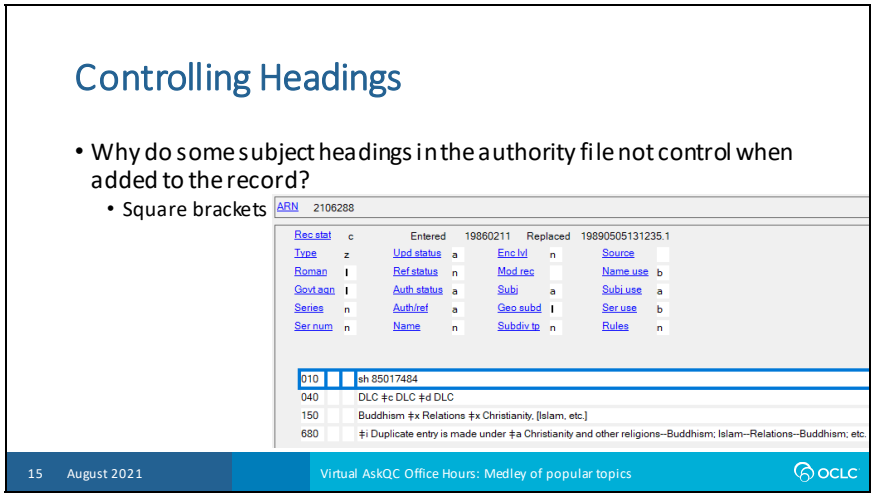
One of the most common questions we receive regarding controlling is why do some bibliographic headings that have an authority record not control to that authority? There are 5 primary reasons a heading cannot be controlled when the heading appears to be in a controllable field:

1. First, check certain codes in fixed field 008, depending on the nature of the heading. If the field is an 8xx series, and the authority record has Series (008/12) coded as "c" for series-like phrase, then controlling will be blocked. Similarly, if "Series use" (008/16) is coded "b", for "not appropriate", then the heading will not be controlled. If the field is for a personal name and Name (008/32) is coded "b" for "undifferentiated personal name", then that heading cannot be controlled. Similarly, controlling will be blocked for name headings with Name use (008/14), and for subject headings with "Subject use" (008/15), coded "b" for "not appropriate".

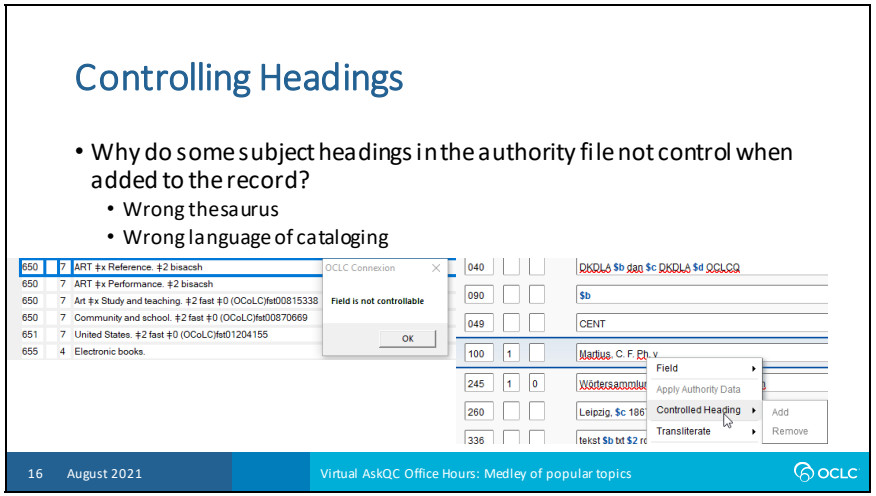


2. The second reason is that the authority is a “Validation record”: In 2007, the Library of Congress began distributing authority records created based on usage of free-floating subdivisions in bibliographic records. As seen in the image, these records are identified by the presence of a 667 note, “Record generated for validation purposes.”

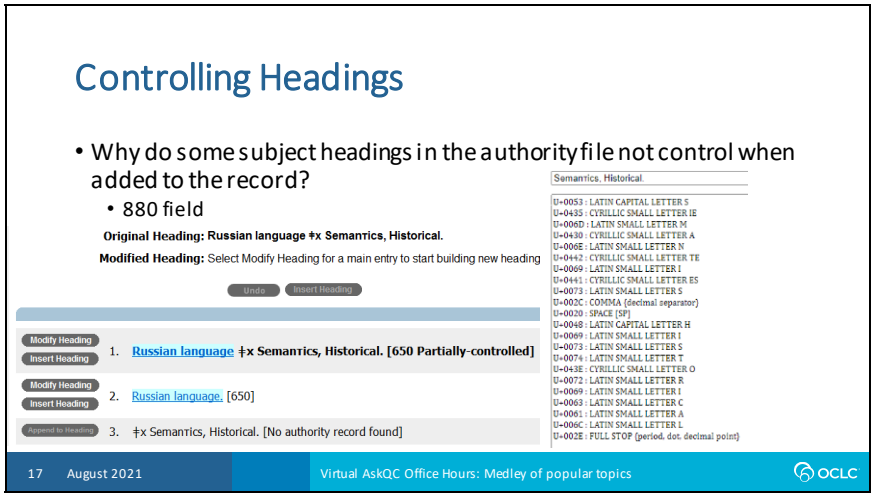
Note that in many cases, the bibliographic heading will control to more than one authority record. For example, Academic writing, plus $v Abstracts (authority records oca03542273 and oca04928257), rather than to the single validation record (authority record oca07459279).



3. A third reason is that the authority's 1xx field contains brackets: For authority files containing "multiple" subdivisions, including LCSH and RVM, controlling is blocked for any heading containing a square bracket. "Multiple" subdivisions are meant to be used as pattern examples for similar subdivisions and are not intended to be used as headings themselves. LC has a project underway to cancel the "multiple" subdivisions from LCSH and create individual authority records for each valid, complete, heading string that was created based on a multiple subdivision. Université Laval, who maintains RVM, has also been working on removing multiples from RVM.

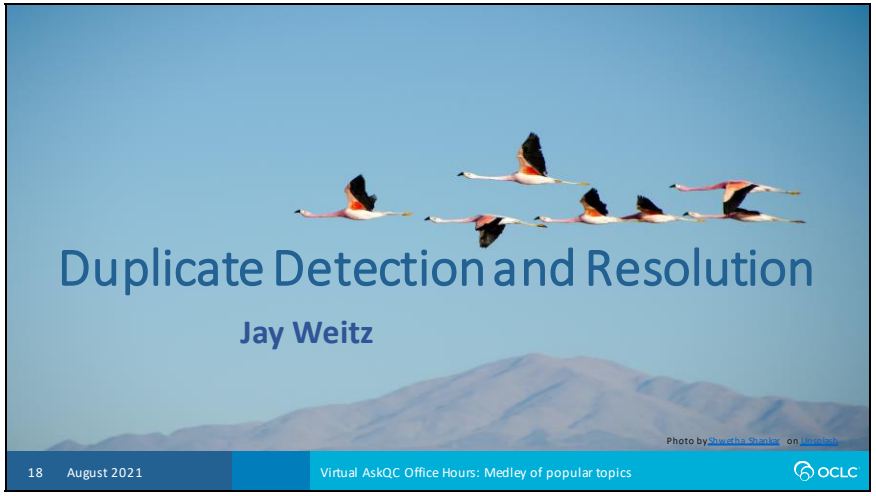


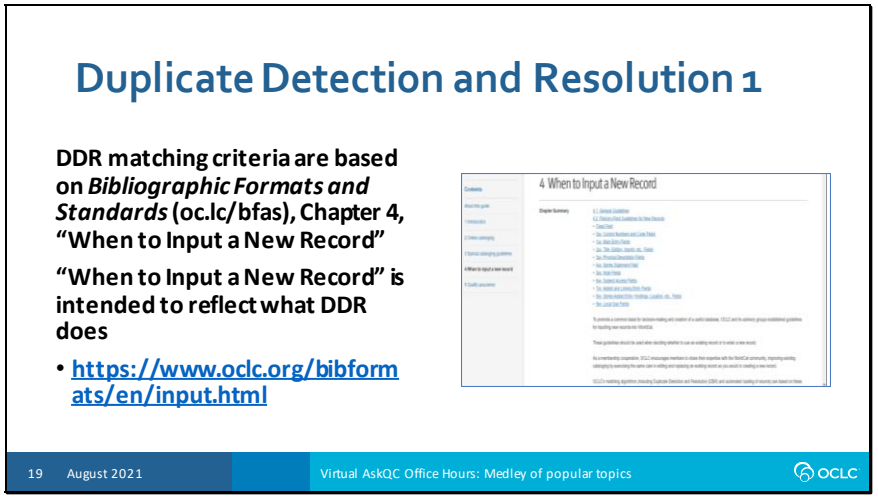
4. A fourth reason is that, for subjects, if the bibliographic field's 2nd indicator, or subfield $2 if the 2nd indicator is 7, does not match one of the controllable thesauri—LC in Connexion; and LC, AAT, Canadiana, Canadian Subject Headings, RVM, BNE, GND, Maori, and MeSH in Record Manager—then the heading cannot be controlled. Similarly, for descriptive headings, the language of cataloging will determine which thesaurus is used for controlling those headings: LC for English (in Connexion and Record Manager); and NTA Names for personal names in Dutch; GND for German; Canadiana for French; and BNE for Spanish.



5. The fifth reason is that the bibliographic heading contains non-Latin characters and is therefore coded as an 880 field masquerading as a controllable field. In the example shown in the image from Connexion, some of the letters in the subfield $x are Cyrillic rather than Latin, as seen in the image showing the characters translated into Unicode values using What Unicode character is this? (https://www.babelstone.co.uk/Unicode/whatisit.html?)

Now I'll hand it over to Jay.

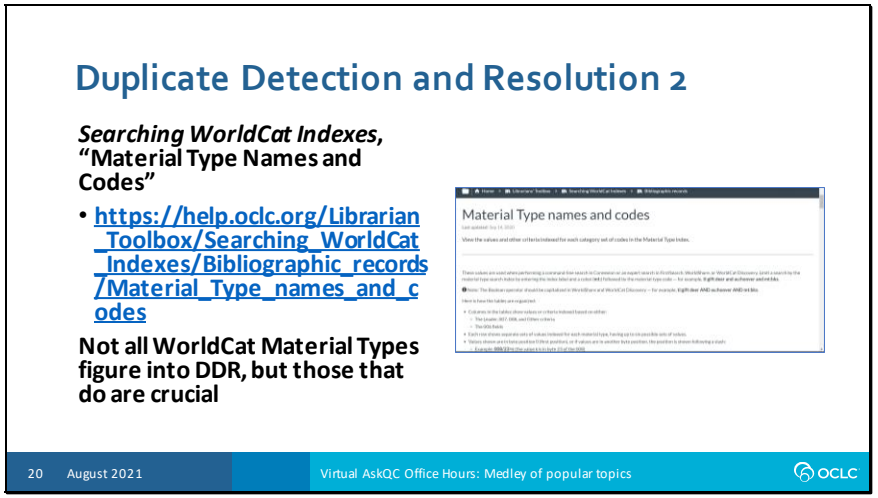

Duplicate Detection and Resolution
Jay Weitz
Photo by Shwetha Shanker on Unsplash
18 August 2021
Virtual AskQC Office Hours: Medley of popular topics
OCLC



Duplicate Detection and Resolution 1

Duplicate Detection and Resolution, or DDR, is OCLC's automated duplicate bibliographic record merging software. It acts behind the scenes, running most new or significantly edited WorldCat records though detailed and exacting matching algorithms, merging any duplicates it finds. We are often asked for a detailed explanation of DDR or about whether specific fields play into those DDR algorithms. The DDR algorithms are impossibly complex and have been developed and refined continually since the late 1980s. There is not a single document that is either publicly available or

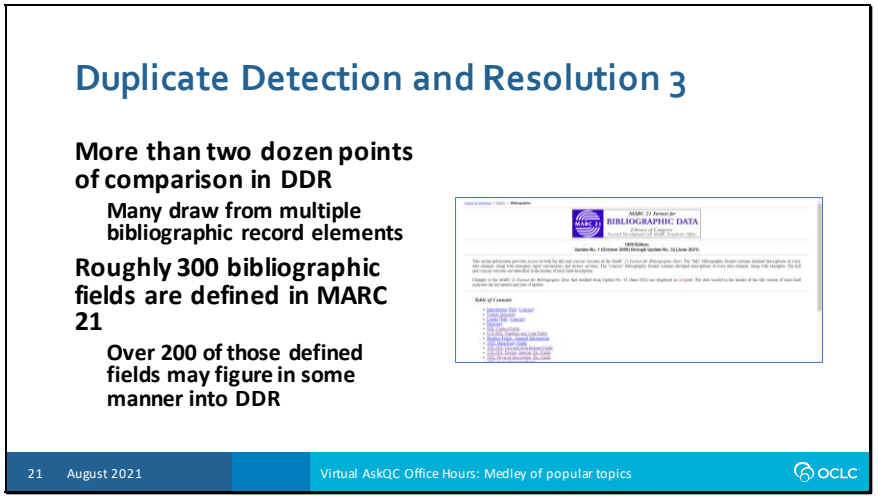
internal to OCLC that explains, or could even come close to explaining, all of DDR. The closest thing we can recommend is "When to Input a New Record," *Bibliographic Formats and Standards* (oc.lc/bfas) Chapter 4. DDR is based on the matching criteria set forth for catalogers to follow in "When to Input a New Record" and conversely "When to Input a New Record" reflects what we try to do in DDR matching.



Duplicate Detection and Resolution 2

Another document that offers insight into one particular aspect of how DDR makes its decisions is Searching WorldCat Indexes (oc.lc/indexes), specifically the section on "Material Type Names and Codes." Not every WorldCat Material Type is taken into consideration by DDR, but those that are considered help DDR differentiate between printed resources and electronic resources, between audio compact discs and audio vinyl discs, between streaming videos and DVDs, between microfiche and microfilm

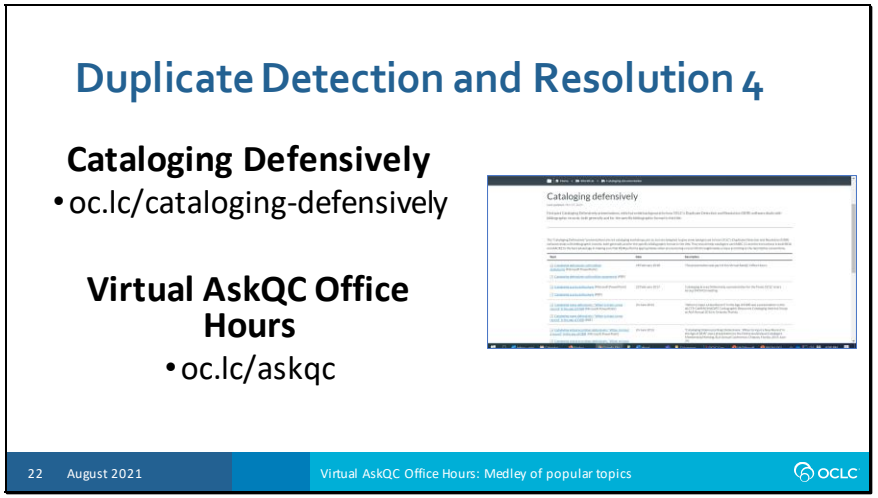
reels, and so on.



Duplicate Detection and Resolution 3

More than two dozen points of comparison are taken into consideration by DDR for bibliographic records. That number is misleading, however, because many of those comparison points actually draw from various parts of the bibliographic record, not simply from a single field. There are roughly 300 fields defined for MARC bibliographic records. According to an analysis I did some years ago, more than 200 of those 300 defined bibliographic fields may play some part in DDR. Many of those fields are the ones you

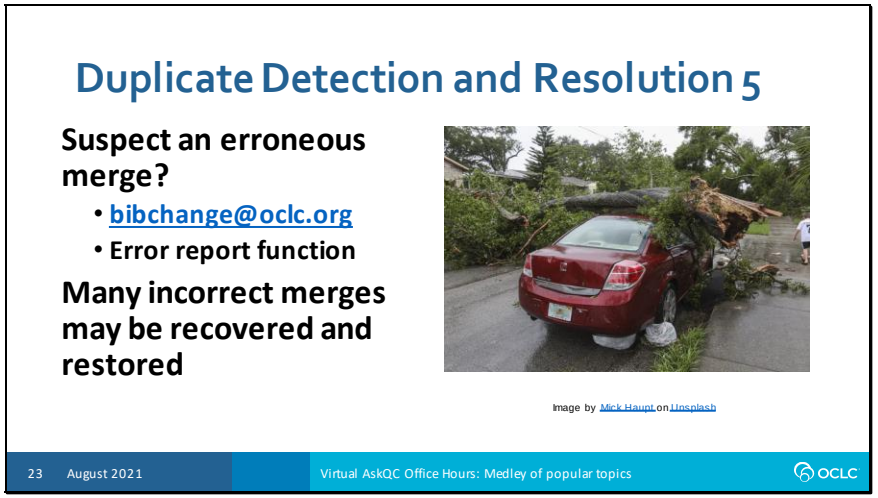
would expect, such as the title (245 field), places of publication, publishers, dates, series. But there are all sorts of other cases where a comparison point is specific to a particular kind of bibliographic record. For instance, scale in Maps records, publisher numbers in Sound Recordings, various elements of instrumentation in Scores, and occasionally even the names of cast members for Moving Images.



Duplicate Detection and Resolution 4

In addition to Chapter 4 of BFAS, "When to Input a New Record," and Searching WorldCat Indexes, we highly recommend the Cataloging Defensively series. These presentations offer background in DDR along with concrete examples of how catalogers may use the instructions in both RDA and AACR2 to their advantage to help ensure that DDR will deal appropriately with records that are legitimately unique according to the descriptive conventions. Cataloging Defensively with Edition Statements

was presented as one of the Virtual AskQC Office Hours in February 2018. Other VAOH sessions related to DDR include "Unraveling the Mysteries of a Merge and DDR Improvements" from August 2020 and "Merging Duplicate Bib Records and the Member Merge Project" from June 2019.



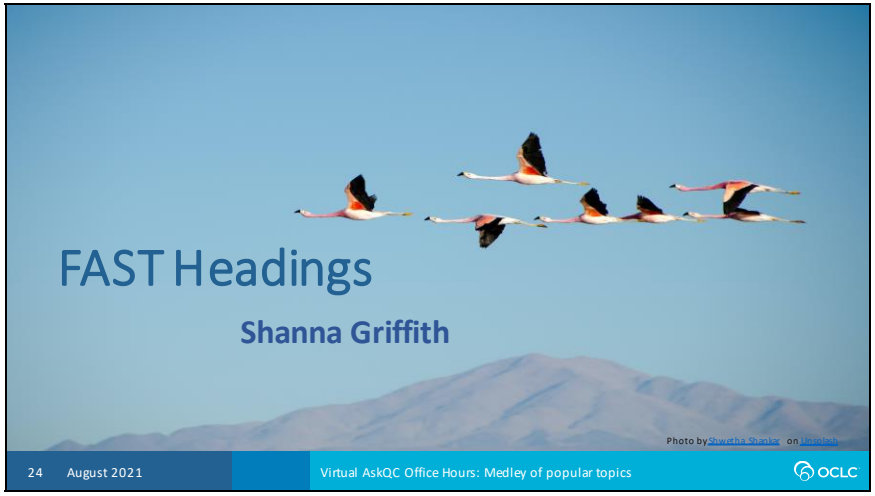
Duplicate Detection and Resolution 5

If you suspect that an incorrect merge may have taken place, please report it to us via bibchange@oclc.org or the error reporting function in Connexion or Record Manager. Metadata Quality staff will investigate the merge using an internal tool that enables us to review all record transactions that have occurred going back to 2012. If we determine that a merge was incorrect, we can roll it back using another internal process that reinstates the WorldCat records involved as well as all of their associated holdings, Local Bibliographic Data (LBDs), and Local Holdings Records (LHRs). More information about this can be found in the August

2020 Virtual AskQC Office Hours mentioned in the previous slide.

Once records are recovered, we can usually make suggestions that will help differentiate them so that DDR doesn't incorrectly merge them again. For records that were merged due to incorrect coding, we will make corrections and use yet another internal tool that runs the records through a DDR test to assure they will not be merged incorrectly again.

Your reports of incorrect DDR merges have been vital to the improvement of DDR over the years. The reports enable us to identify when our algorithms could use a little tweaking in the hopes that the same situation will not occur in the future.


FAST Headings
Shanna Griffith
Photo by Shwetha Shanker on Unsplash
24 August 2021
Virtual AskQC Office Hours: Medley of popular topics
OCLC

FAST Headings

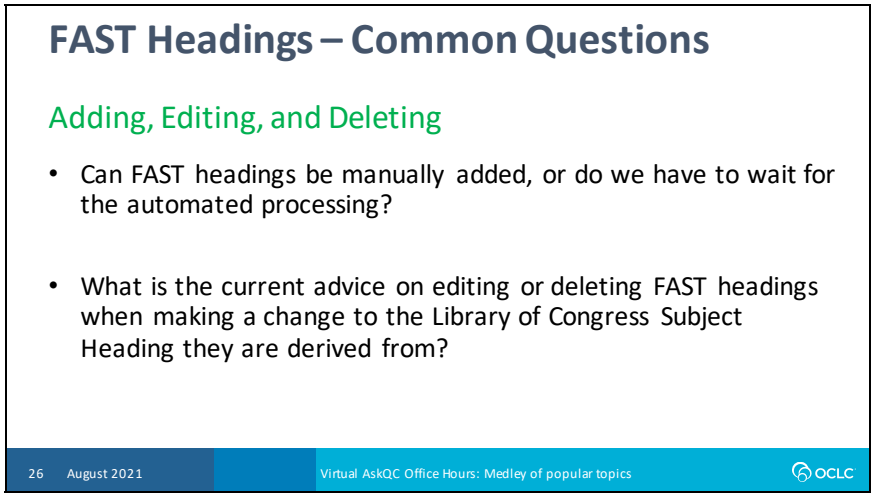
Faceted **A**pplication of **S**ubject **T**erminology

650 _0 Beagle (Dog breed)

650 _7 Beagle (Dog breed) ǂ2 fast ǂ0 (OCoLC)fst00829173

Miss. Marley Mae

Hello. Today I'll be answering a few common questions that we have received regarding FAST headings. In case you aren't familiar with FAST headings, FAST stands for Faceted Application of Subject Terminology, and are derived from the Library of Congress Subject Headings. On the slide I show an example of the derived FAST heading for the LCSH Beagle (Dog breed), and the picture on the slide is of my cute beagle, Miss. Marley Mae.



A couple of the most common questions we get regarding FAST headings involve advice around adding, editing, and deleting FAST headings.

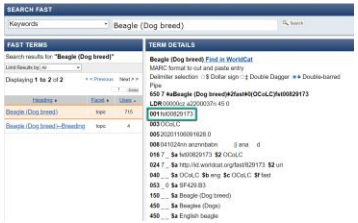
Adding, Editing, Deleting:

- Currently, FAST headings are automatically added by algorithm to WorldCat bibliographic records when there are no existing FAST headings in the record and there are Library of Congress Subject Headings present in the record. After the initial FAST headings are added, the headings are maintained field by field, regardless of whether the fields were entered by a user or by the FAST algorithm.
- FAST headings may be manually assigned to a new record for a resource being cataloged, or to a record found in the database for the resource
- If you are changing, deleting, and/or adding Library of Congress Subject Headings in a record, it is encouraged, but not required, to delete the entire set of FAST headings, and they will be completely regenerated within a month or so

FAST Headings – Common Questions

Searching and Controlling

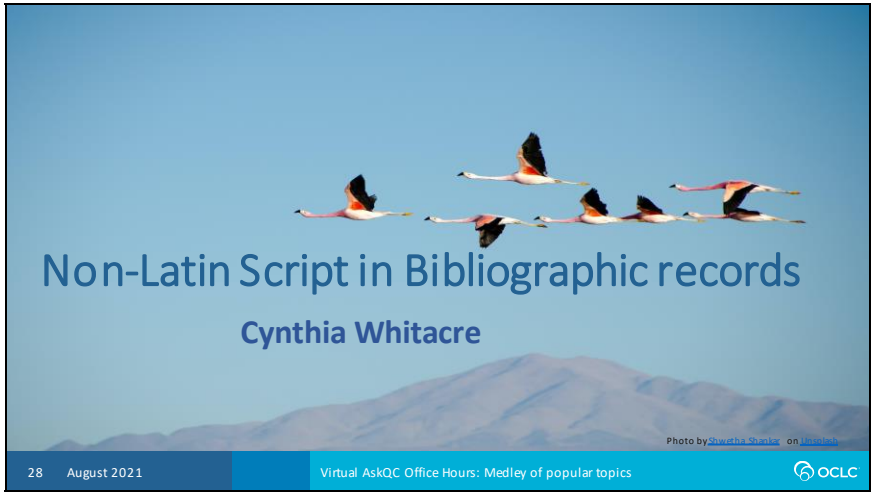
- Can FAST headings be searched?
  - ar:oclc-fst00829173
  - searchFAST - http://fast.oclc.org/searchfast
- Why aren't FAST headings controlled?

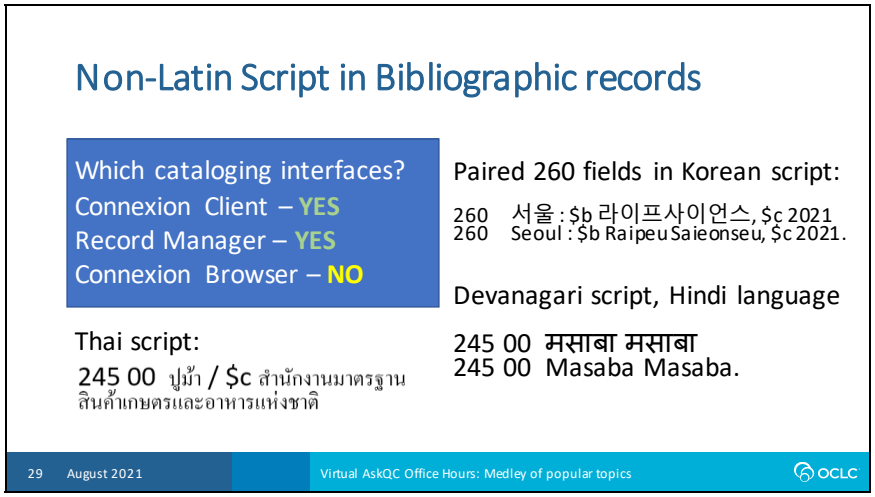


A couple more common questions we get involve searching and controlling FAST headings.

Searching and controlling:

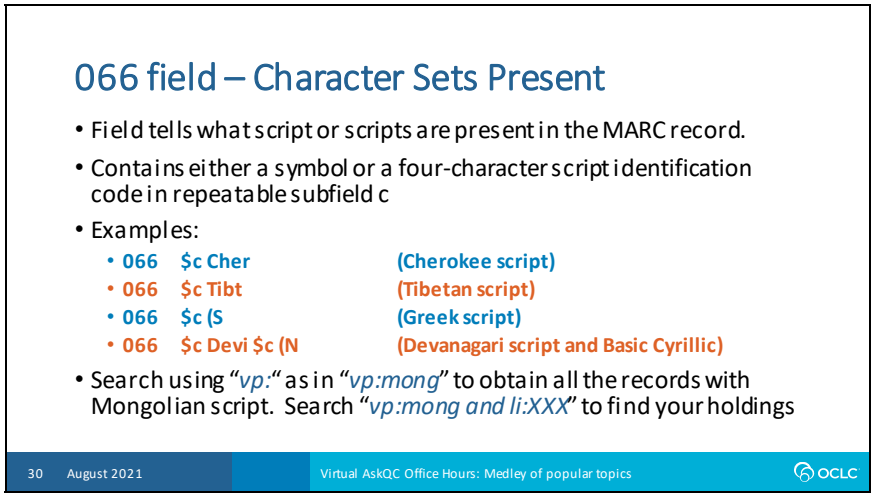
- The 'ar:' index works with all cataloging interfaces and allows you to search bibliographic records by the FAST authority number, FAST is also accessible via Auto-suggest in Record Manager.
- In addition, FAST headings can be searched in searchFAST and added manually to WorldCat bibliographic records
- Currently, FAST headings are not controllable. This is something that Metadata Quality would like to do but is further down the list of things to implement


Non-Latin Script in Bibliographic records
Cynthia Whitacre
Photo by Shwetha Shanker on Unsplash
28 August 2021
Virtual AskQC Office Hours: Medley of popular topics
OCLC

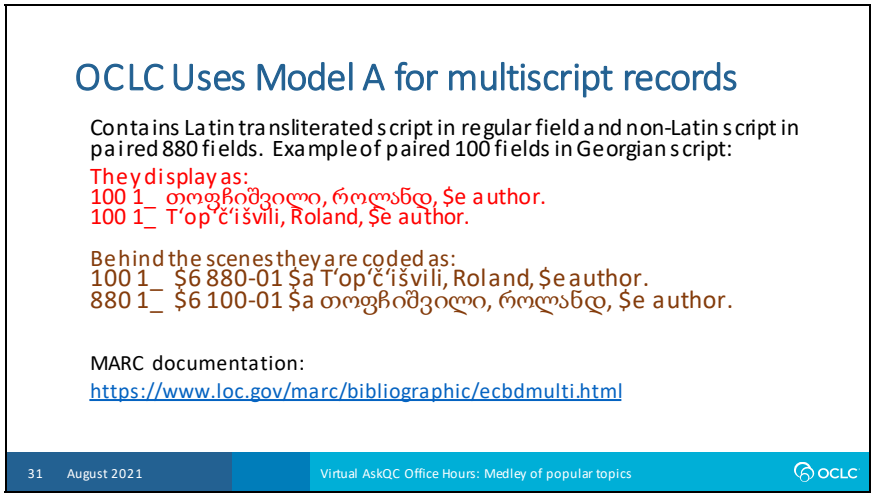


I'll now be providing some brief information on non-Latin script in bibliographic records. To illustrate the diversity of scripts, I've added three examples of different scripts on this slide, all drawn from bibliographic records in WorldCat.

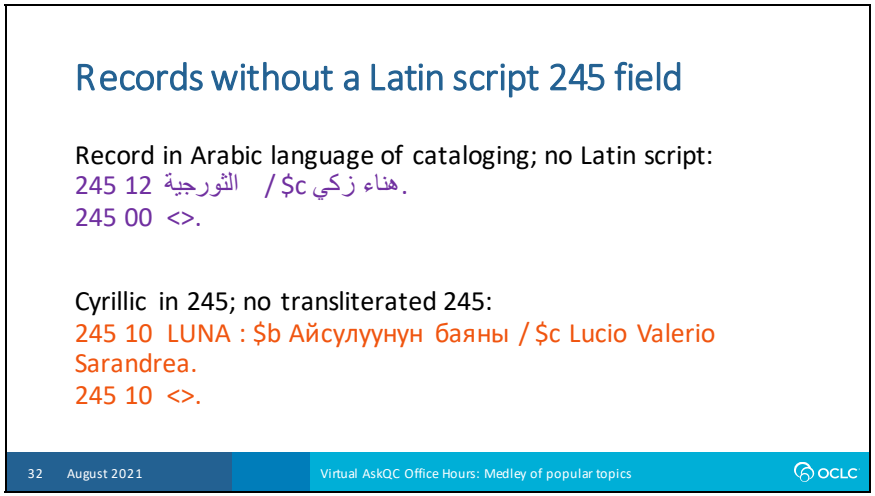
Non-Latin script is also referred to as non-Roman script or as vernacular. One may enter, edit, and/or view non-Latin scripts in both the Connexion Client and Record Manager, but not in the Connexion Browser. We've received questions sometimes about why a record cannot be edited in the Connexion browser. Usually this is because it is a non-Latin script record. The browser does not display the non-Latin fields or allow editing of records with any non-Latin script.



The 066 field contains a symbol or a code that tells what non-Latin scripts are present in the record. This field is system supplied, so when cataloging you do not need to look up or enter this data. The subfield c is repeated, as in the fourth example, if there are multiple non-Latin scripts in a record. It can be a very useful clue to tell you what you are seeing in the record. And, the field is indexed, so you can search for the script or use this to qualify a search. For example, if you want to find all the records with Mongolian script held by your library, you could do a command line search in Connexion of: li:XXX where XXX is your OCLC symbol combined with vp:mong.



MARC standards, available at the link at the bottom of the slide, provide 2 models for encoding non-Latin script. OCLC uses Model A. That puts all non-Latin script in 880 fields that may be linked to Latin fields. In cataloging interfaces, OCLC displays these as paired fields, so that you do not see the 880 field coding. The example shows a Georgian script field paired with its transliterated Latin script field. Behind the scenes, the non-Latin field is coded as 880 with a subfield 6 pairing it to the correct Latin field. Likewise, the 100 field contains a subfield 6 pairing it to the correct 880 field. When doing non-Latin script cataloging, you do not need to enter the subfield 6 coding or the 880 field tag. This happens automatically behind the scenes when you pair fields.

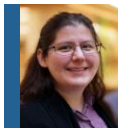


OCLC does accept records that contain only non-Latin script, since many libraries around the world do not use Latin script. We are still using Model A, so all the non-Latin fields are actually 880 fields behind the scenes. However, our OCLC system requires a Latin script 245 in bibliographic records. So, when no Latin script 245 is present, a 245 containing only angle brackets is system supplied to provide that required Latin script field. The first example is, in fact, from a record that is entirely cataloged in the Arabic language in Arabic script from a university in Lebanon.

If you have non-Latin script mixed into an otherwise Latin script 245, such as Greek letters, the system may supply those angle brackets and your 245 is really an 880 field. The second example shows a mixed field, which is system coded as an 880 field, since it contains some non-Latin characters. These 245 fields with angle brackets display in Connexion but do not display in the Record Manager interface.

On the call today

Robin Six
Database Specialist



Charlene Morrison
Database Specialist

Bryan Baldus
Consulting Database Specialist

Jay Weitz
Senior Consulting Database Specialist

Shanna Griffith
Database Specialist

Cynthia Whitacre
Senior Metadata Operations Manager

Thank you!

September Virtual AskQC Office Hours
The complexities of field transfer

Tuesday, 14 Sept. at 10:00 AM Eastern
Thursday, 23 Sept. at 4:00 PM Eastern

Registration and session links available at oc.lc/askqc

Send cataloging policy questions at any time to askqc@oclc.org

Photo by Eric Rothermel on Unsplash