May 8, 2019

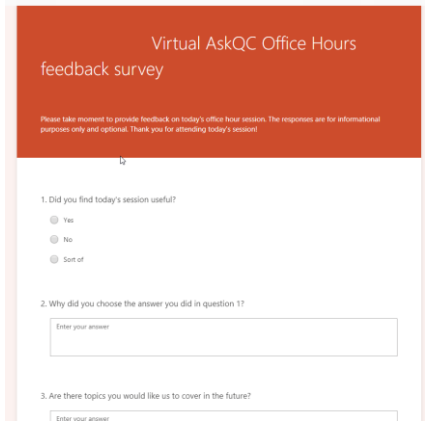
Virtual AskQC Office Hours

Small errors with big consequences

OCLC Metadata Quality

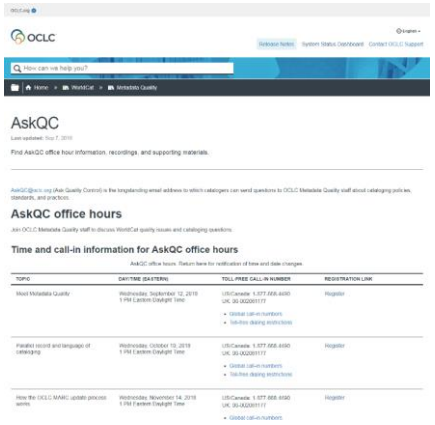
Housekeeping

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



Housekeeping

- After the session you will be directed to a quick, optional survey
- All session recordings, slides, and notes are available at oc.lc/askqc



On the call today

Robin Six
Database Specialist II

Bryan Baldus
Consulting Database Specialist

Robert Bremer
Senior Consulting Database Specialist

Hayley Moreno
Database Specialist II

Charlene Morrison
Database Specialist II

Cynthia Whitacre
Manager, Metadata Policy

SMALL ERRORS WITH BIG CONSEQUENCES

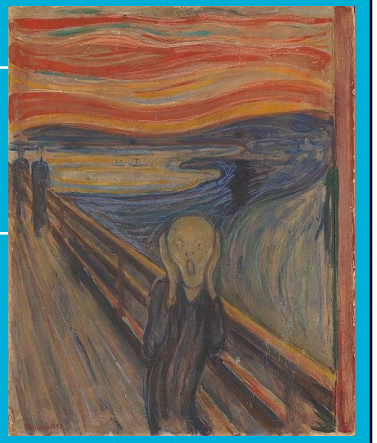


Image from Wikimedia Commons, in public domain

OCLC

Cataloging process

Catalogers normally focus on key elements such as:

- Accurate transcription of bibliographic data
- Addition of authorized access points
- Assignment of correct call numbers
- Assignment of appropriate subject headings

However, all aspects of the resulting bibliographic record are important in varying degrees

Small errors

Some small errors can have a significant impact on:

- Indexing and retrievability – if catalog users cannot find the resource, they cannot use the resource
- Identification of the resource – may have an impact on user recognition of a needed resource and often gets in the way of automatic processing of duplicate records
- Impact on copy cataloging workflows – particularly if you need identify and fix these kinds of errors
- Impact on local system processing – varies by system

Single digit incorrectly coded

Issue: Incorrectly coded nonfiling indicators

Impact: Incorrect indexing of titles – inability to search title phrases (and title keywords in some cases) – impact on DDR and the ability to automatically eliminate duplicate records

Common cases:

245 0**0** The sacred books of the Hindus

245 0**0** Der Jacobi'sche Garten zu Pempelfort bei Düsseldorf

245 0**0** La cabeza del diablo

Single digit incorrectly coded

Cases without nonfiling indicator where article was retained:

246 1 ǂi First line of text: ǂa The shades of night were falling fast

Cases that are more difficult to find where no article is present and the nonfiling indicator is coded with a value other than zero:

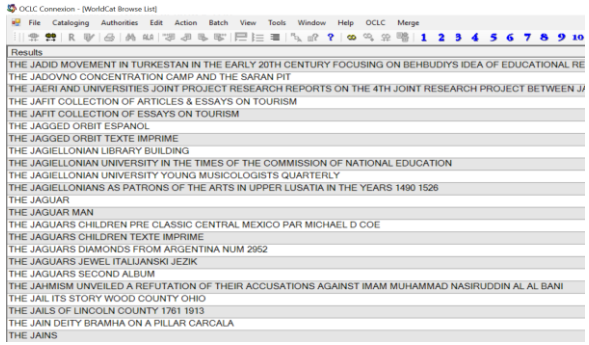
245 12 Study of heat transfer to the end wall of a shock tube

245 13 Administración para pequeños empresarios

740 41 Story of the collection and the museum

This last case affects keyword indexing of the title

Single digit incorrectly coded



Personal name access points

Which one is incorrect?

700 1 Beethoven, Ludwig van, ǂd 1770-1827.
700 1 Custer, George A. ǂq (George Armstrong), ǂd 1839-1876.
700 1 Lincoln, Abraham, ǂd 1809-1865.
700 1 Patton, George S. ǂq (George Smith), ǂd 1885-1945.
700 1 Rasputin, Grigoriĭ Efimovich, ǂd 1869-1916.
700 1 Robespierre, Maximilien, ǂd 1758-1794.
700 1 Scriabin, Aleksandr Nikolayevich, ǂd 1872-1915.
700 1 Twain. Mark, ǂd 1835-1910.
700 1 Washington, George, ǂd 1732-1799.

Personal name access points

It was just a comma incorrectly keyed as a period:

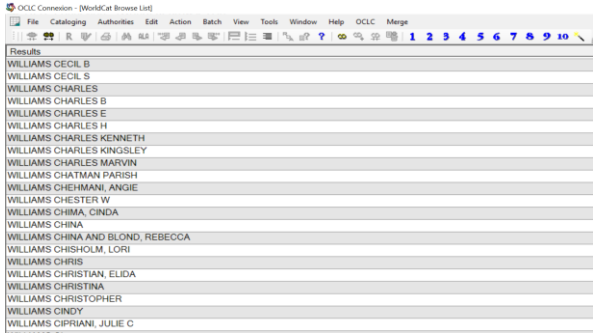
700 1 Twain. Mark, ǂd 1835-1910.

But, that is enough to cause the access point to be indexed incorrectly:

TWAIN MARK 1835 1910
vs.
TWAIN, MARK 1835 1910

Fortunately, this does not affect the ability to control a heading.

Personal name access points



Character issues

Correct use of characters does matter— it's not just how it appears to the catalog user, it's as much how it appears to the system

245 00 Bake at 375° [that is a degree symbol]
245 00 Bake at 375⁰ [that is a superscript zero]

Per NACO normalization used in WorldCat indexing
245 00 BAKE AT 375
245 00 BAKE AT 3750

Character issues

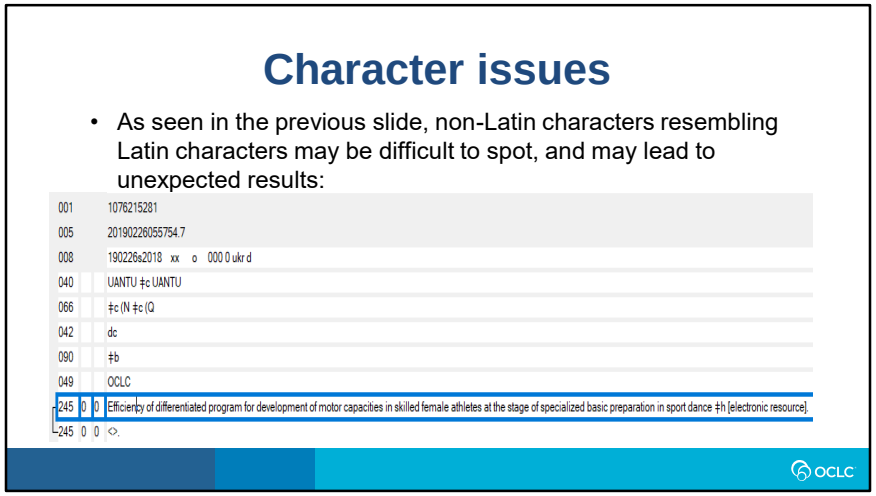
Are these the same or different: **Москва** vs. **Москва**?

Москва	Москва
U+041C : CYRILLIC CAPITAL LETTER EM	U+041C : CYRILLIC CAPITAL LETTER EM
U+043E : CYRILLIC SMALL LETTER O	U+006F : LATIN SMALL LETTER O
U+0441 : CYRILLIC SMALL LETTER ES	U+0063 : LATIN SMALL LETTER C
U+043A : CYRILLIC SMALL LETTER KA	U+043A : CYRILLIC SMALL LETTER KA
U+0432 : CYRILLIC SMALL LETTER VE	U+0432 : CYRILLIC SMALL LETTER VE
U+0430 : CYRILLIC SMALL LETTER A	U+0430 : CYRILLIC SMALL LETTER A

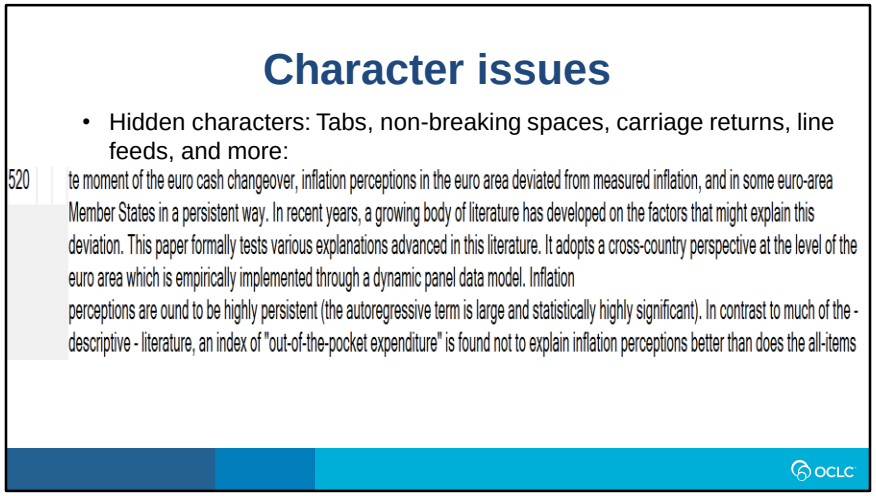
Character issues

- Expansion to use of all Unicode characters has made correct coding of characters more important that ever before
- Incorrect characters can cause bibliographic records to be incorrectly indexed and to fail to match
- You can check suspicious or problematic text by copying and pasting text into this website:

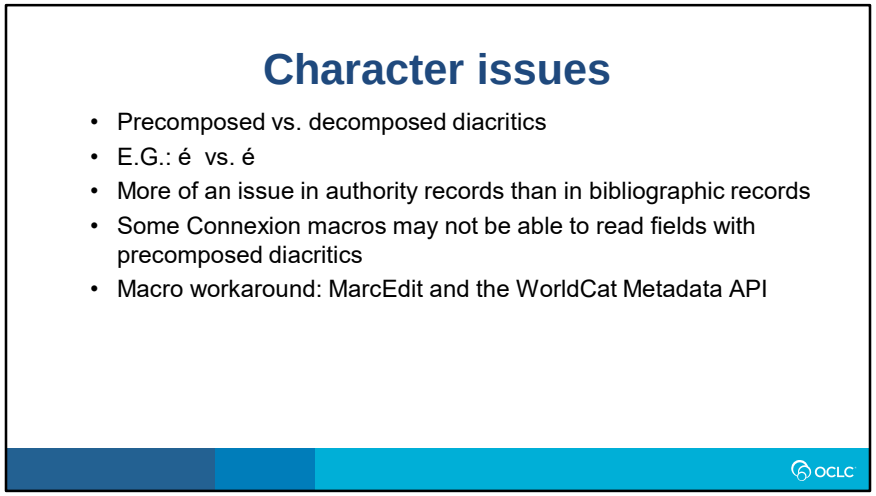
http://www.babelstone.co.uk/Unicode/whatisit.html



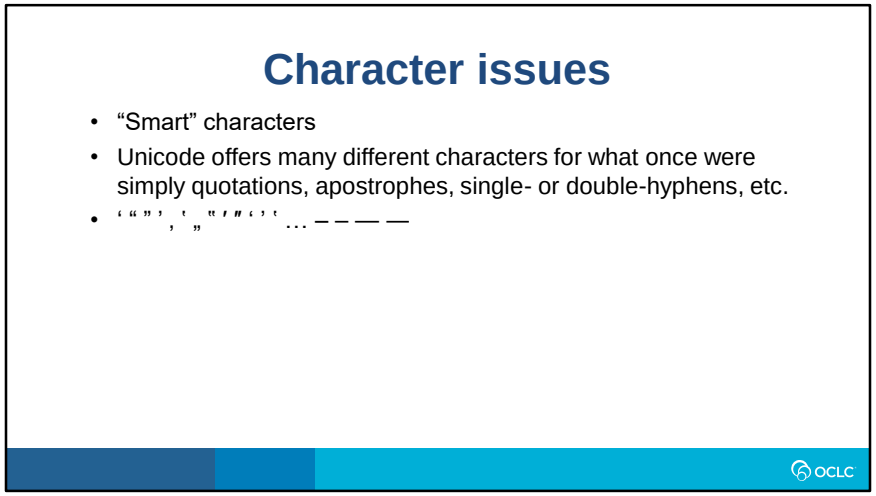
In this case, the first E in the title is really ye (a.k.a. U+0415 : CYRILLIC CAPITAL LETTER IE). This produces no Validation errors, but the 2nd 245 with less-than, greater-than, period, should cause you to wonder if something might be out of place somewhere within the 245 above (which behind the scenes is really coded as an 880). After thoroughly reviewing the field and not immediately identifying the character or characters causing the 245 to flip into this linked pair, my next stop would be http://www.babelstone.co.uk/Unicode/whatisit.html?
As for why the 2nd 245 is given as "<>.“, that would be because all records within Connexion client must have an actual 245. If one is not found, a place-holder field with <>. Is added, linked to the 880 coded as corresponding to the 245.



When characters are invisible, it's difficult to come up with a screen shot of them. However, in iths case, you might be able to notice a significant break between the 4th and 5th lines. A record containing a carriage return, and/or line feed was imported into Connexion, resulting in the display seen here. Not seen here is a non-breaking space character between "term" and "is" within the parentheses on the 5th line. For this specific example, I created the record and then imported it into Connexion in a local save file. I was unable to successfully import tab characters, but that doesn't mean that tabs, in addition to these and other essentially invisible characters, aren't lurking within WorldCat database records. In the case of carriage returns and line feeds, Validation will produce an error message, "520 occurrence 1, $a occurrence 1, position 502 - invalid character - data must be ALA characters“. Other hidden characters may or may not produce a Validation error message--NBSP does not.
Why are the other hidden characters a problem? One reason is that when conducting a phrase search within Connexion client on “the autoregressive term is large” (with spaces between each term), that phrase will not be found.

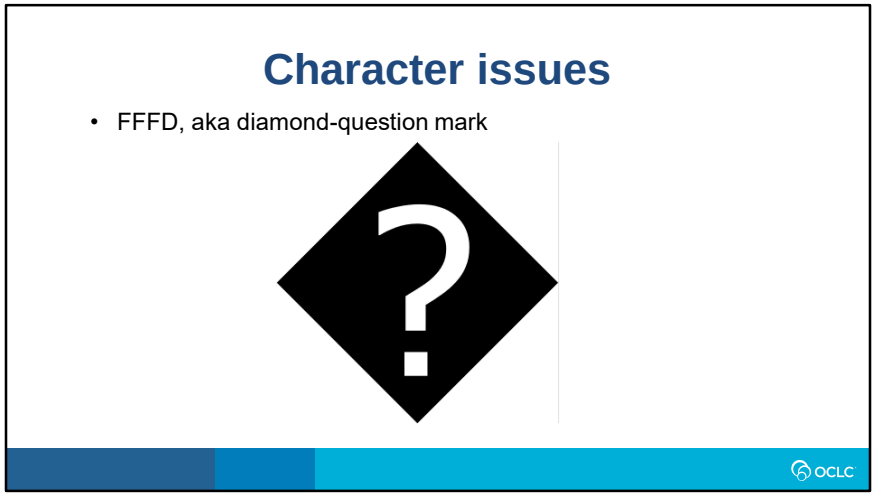


Characters containing diacritics are generally stored in WorldCat in the form in which they enter the database, either decomposed, in which a combining diacritic is applied to a letter, or precomposed, in which a there is a single character representing a letter with all of its associated diacritics built-in. Within the LC/NACO Authority File, all characters must be decomposed, or they risk being rejected by a NACO node upon distribution, or being improperly indexed within OCLC. For bibliographic records, the form of diacritics is less important, but some Connexion macros may have more difficulty reading precomposed characters, mangling a field as they attempt to act upon it. Asa workaround to this issue, users with access to the WorldCat Metadata API could make use of MarcEdit's built-in functionality for handling precomposed characters, to convert them to their decomposed equivalents, and then replace the WorldCat record. Once that is done, Connexion macros which require decomposed characters should be able to work with the freshly updated records.



Looking through the Insert Symbol tool within PowerPoint, I found this sample of curly apostrophes or quotations, including a few that are spacing modifier characters, along with a prime and double prime; plus the triple dot ellipsis character, figure dash, en dash, em dash, and horizontal bar. While these might be useful in a word processing document, they cause troubles within bibliographic records--there is no reason to be using any of these when a simple apostrophe, quotation mark, 3 dots, or one or two hyphens would work more universally across systems.

While many modern systems may be able to handle these, when MARC data is transferred across various systems, use of these smart characters will frequently lead to one of the more infamous characters…



Based on discussions on OCLC-CAT and elsewhere in the cataloging world, it seems some of you are familiar character. When systems corrupt characters beyond recognition, this character, the Replacement Character, takes their place. OCLC Metadata Quality has been hunting these down throughout the database, and have been doing what we can to prevent them from being introduced into WorldCat records in the first place. They will cause a field to turn into an 880, with 066 Zsym, so if you see that in 066, your first thought should be to look for any FFFD characters that don't belong. This is particularly true when you notice duplicated fields, as seen earlier with the Cyrillic "ye" character.

No place, no publisher, no date, etc.

It should be *publisher not identified* rather than:

Publisher not indentified · publisher not identifiable · publisher not identifed · publisher not idenified · name of publisher not identified · name of publisher has not been provided · publisher's name is not stated · publisher name is unknown · publisher name was unrecorded · publisher was not established · name of publisher not indicated · publisher's name was not given · publisher name not certain · publisher is not provided · name of publisher was not found · publisher's name cannot be determined · publisher name is unavailable · publisher was not cited · name of publisher not listed · name of publisher not designated · publisher's name was not given · publisher name not defined · no publisher listed · name of publisher has not been determined · publisher's name is not noted · publisher name is not presented · publisher name was not on label · publisher was not published · name of publisher not shown · publisher's name not recognized · publisher name not mentioned · publisher is not specified · name of publisher was not verified · publisher's name cannot be supplied · publisher name is lacking · publisher was omitted · name of publisher not given · name of publisher not recorded · publisher's name was not stated · publisher name not defined · no publisher stated · name of publisher has not been designated · publisher's name is not found · publisher name is missing · publisher name was not located · publisher was not mentioned · name of publisher not stated · publisher's name not certain · publisher name not cited · publisher is not presented · name of publisher was not named · publisher's name cannot be given · no publisher determinable · name of publisher has not been located

No place, no publisher, no date, etc.

Why does any of this matter? It has an impact on record matching and deduping, e.g.,

John Wiley & Sons = Wiley

s.n. = publisher not identified

but

Wiley ≠ Macmillan

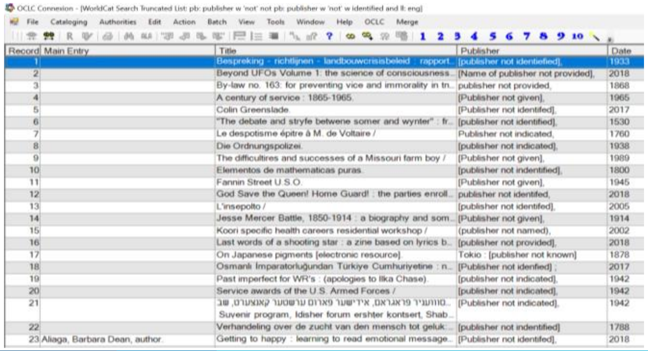
and

publisher not identified ≠ name of publisher omitted

No place, no publisher, no date, etc.

publisher publisher name publishers name publisher's name name of publisher no publisher no publishers name no publisher's name no name of publisher	*not* cannot be is not was not has not been un-	*identified*, available, certain, cited, defined, described, designated, determinable, determined, displayed, established, evaluated, fixed, found, given, identifiable, indicated, known, listed, located, mentioned, named, noted, on label, present, presented, provided, published, recognised, recognized, recorded, shown, specified, stated, submitted, supplied, transcribed, verified

No place, no publisher, no date, etc.



OCLC Connexion - [WorldCat Search Truncated List: pb: publisher w 'not' not pb: publisher w 'not' w identified and ll: eng]

Record	Main Entry	Title	Publisher	Date
1		Bespreking - richtlijnen - landbouwcrisisbeleid : rapport...	[publisher not identiefied],	1933
2		Beyond UFOs Volume 1: the science of consciousness...	[Name of publisher not provided],	2018
3		By-law no. 163: for preventing vice and immorality in tn...	publisher not provided,	1868
4		A century of service : 1865-1965.	[Publisher not given],	1965
5		Colin Greenslade.	[Publisher not identifed],	2017
6		"The debate and stryfe betwene somer and wynter" : fr...	[publisher not identified],	1530
7		Le despotisme épitre à M. de Voltaire /	Publisher not indicated,	1760
8		Die Ordnungspolizei.	[publisher not indicated],	1938
9		The difficultires and successes of a Missouri farm boy /	[Publisher not given],	1989
10		Elementos de mathematicas puras.	[publisher not indentified],	1800
11		Fannin Street U.S.O.	[Publisher not given],	1945
12		God Save the Queen! Home Guard! : the parties enroll...	publisher not identifed,	2018
13		L'insepolto /	[publisher not identifed],	2005
14		Jesse Mercer Battle, 1850-1914 : a biography and som...	[Publisher not given],	1914
15		Koori specific health careers residential workshop /	(publisher not named),	2002
16		Last words of a shooting star : a zine based on lyrics b...	[publisher not provided],	2018
17		On Japanese pigments [electronic resource].	Tokio : [publisher not known]	1878
18		Osmanlı İmparatorluğundan Türkiye Cumhuriyetine : n...	[Publisher not idenfied] ;	2017
19		Past imperfect for WR's : (apologies to Ilka Chase).	[publisher not indicated],	1942
20		Service awards of the U.S. Armed Forces /	[publisher not indicated],	1942
21		סוּווועניר פּראָגראַם, אידישער פאָרום ערשטער קאָנצערט, שב... Suvenir program, Idisher forum ershter kontsert, Shab...	[Publisher not indicated],	1942
22		Verhandeling over de zucht van den mensch tot geluk:...	[publisher not indentified]	1788
23	Aliaga, Barbara Dean, author.	Getting to happy : learning to read emotional message...	[Publisher not identifed],	2018

OCLC

Relator terms and codes

Which one is correct?

100 1 Mozart, Wolfgang Amadeus, ǂd 1756-1791. ǂ4 cmp
100 1 Mozart, Wolfgang Amadeus, ǂd 1756-1791. ǂ4 com

ǂe author of introduction
ǂe writer of introduction
ǂe introduction author
ǂe introduction writer

Relator terms and codes

- Errors in relator terms and codes are easy to make
- Short term impact on copy cataloging if catalogers feel a need to review and correct these elements
- Longer term impact the use of the cataloging data for linked data when attempting to identify the nature of the relationship

Other MARC codes

- MARC codes may be concise
- MARC codes are often based on the corresponding English-language term
- Guessing at a code is never a good idea
- Codes in the fixed-field such as *Type*, *BLvl*, *Form*, and *TMat* are essential for correct indexing and identification

Other MARC codes

Consider the fixed-field coding for Ctry. Which of the following combinations are correct?

1) Ctry: sw – 264 1 Geneva, Switzerland
2) Ctry: sz – 264 1 Geneva, Switzerland
3) Ctry: sw – 264 1 Manzini, Swaziland
4) Ctry: sz – 264 1 Manzini, Swaziland

Codes in *Ctry* are used in DDR processing when identifying and eliminating duplicate records from the database.

Small errors with big consequences

"A foolish consistency is the hobgoblin of little minds …"

—Ralph Waldo Emerson

Obviously, Emerson never worked with cataloging data in MARC 21 where consistency, foolish as some of it may seem, accomplishes what is needed to make library resources available.

Questions?

Robin Six
Database Specialist II

Bryan Baldus
Consulting Database Specialist

Robert Bremer
Senior Consulting Database Specialist

Hayley Moreno
Database Specialist II

Charlene Morrison
Database Specialist II

Cynthia Whitacre
Manager, Metadata Policy

Please submit questions through chat

Thank you!

Send cataloging policy questions at anytime to:
askqc@oclc.org

Session links available at:
oc.lc/askqc

Next Virtual AskQC Office Hours:

Member Merge Project

Wednesday, June 12, 2019
1:00 PM Eastern

Register at oc.lc/askqc

Because what is known must be shared.®

OCLC