2018-05-30 Virtual AskQC Office Hours

Topic presentation

Validation in Bibliographic Records in WorldCat, presented by Robert Bremer

Robert Bremer (Senior Consulting Specialist) explains how the use of Validation helps ensure that bibliographic records contain valid MARC coding in various applications such as Data Ingest, Connexion, and Record Manager. He explains with detail the process that uses a database of MARC data elements, such as fixed field elements, tags, indicators, and subfields.

Member questions

How do you turn on 006 mneumonics?

Answer: The only way you would do that in the Connexion client for example is by going through the drop-down menus for the guided entry and you can look at existing field to potentially edit it with the fixed field mnemonics in that case. If you have an 006 field in the record and you right click on it the guided entry box will come up as well as in the drop down from the top.

How is record validation different for batch loaded records?

Answer: It really is the same. It used to be that in our batchload processing that we had a separate set of validation rules, but over time we have come to use the very same set of validation rules so that we don't have to maintain multiple sets of rules. The difference comes in how we deal with the errors that are spotted in the records after the fact. In the batchload processing there is an error level that is assigned to various records and causes the records to go through different kinds of processing in those cases, it is not as if records have to be absolutely perfect to be added to the database but we do detect the very same set of errors.

Why do I see records that have 6xx fields with second indicator 7 and no subfield 2. I have to clear this before I can validate these records in the Client.

Answer: That is a case where the record has most likely come the batchload process, because it is impossible to do a record like that as an online input. We have a relationship in place between 6xx with second indicator coded 7 and a subfield 2 but relationships in batchload are considered a lower level error and those can find their way into the database.

Can you give us any tips on how to identify bad characters?

Answer: This is really tough thing to go look for and find. This may be different in Record Manager, but in Connexion client you could input the vertical bar charter at different positions in the field as a way to spot where you should be to find that invalid character. There may also be macros out there that can help find that spot.

Why is it that a record can fail validation, but I can still attach our holdings?

Answer: This is where you have options that you can set in your holdings so that you don't require full validation. A lot of libraries would prefer to be able set their holdings without necessarily fixing everything on a record. If you do an explicit validation command you will get back full validation with all of the errors listed, but that is not necessarily something that you have to fix in order to set a holding.

When exporting a record from OCLC and it shows a validation error and you fix the error then export the record with no errors, does that correction stay in the record in OCLC and the next time someone exports it the record will show the same error?

Answer: In a case like this you need to replace the record if you are able to do so. If the error is on a PCC record, and you only have a full level authorization, you can report so that we can fix it.

I have noticed that some records have 336-338 $b not compliant with $a (and with the rest of the metadata), when $2 = rda.

Answer: The question here is that the code in subfield $b is actually out of step with the term in subfield $a. At this point we validate the term in $a, we validate the code in $b but we don't actually have the two of them related together.

I also see records that show 650_1 headings underlined as controlled.

Answer: This is an error where we have transferred the controlling from an LC heading in the past, if you see ones like that report them so that we can investigate.

Our library used to use a single record for print and online resources, on the monograph form. We are a PCC library. We now create individual records. When we are adding a 776 for the online version, can/should we remove the electronic elements from the print version?

Answer: In part, there was a decision within CONSER in the past that considered the issues related to the print record carrying so much information about the electronic resource because the record should represent the print and note the existence of the electronic. We were including elements from the electronic version in the record that could then be confusing when processing the record in the future. So, a decision was made to remove certain things like 006, 007 relating to the electronic version, at least for serials initially, and then that conversation carried on over into monographs where there was a PCC decision to handle that exactly the same way. But you will still find records in the database that represent the print version and form is coded blank in the fixed field also indicating that it represents the print version but there might be a 006 or 007 there for electronic information that may come out.

Unless I am mistaken, the validation tool does not let me know when I forget to include the subfield b or e in the 040 field. Are there plans to include this in the validation process?

Answer: For 040 subfield $b which is the language of the cataloging, we have discussed before making that a mandatory element so that you would be required to input subfield $b when you were creating a new record. Subfield $e is a harder thing to require in that there is not another element to link that to because if you created an AACR2 bibliographic record you would not have a subfield $e in the 040. Desc in the fixed field would be just coded as just "a" and there. But, Desc i without 040 subfield $e is a valid combination so subfield $e may be

something you will always just have to remember, subfield $b may be something we input a relationship for in the future. .

So even though the Client is not being updated anymore as a tool, you can update the validation rules it uses?

Answer: Yes, that's correct, since it is the same validation rules that is used for various services. When the MARC update is applied we update the validation database. It is automatically updated for Connexion and Record Manager.

I have a found a number of RDA records that have a 264 #1 $c 2017 when the resource only has a copyright date. It's my understanding that the date should be in brackets in this case. These are recently published books so unlikely that there is a different publication which includes the date as a publication date. Am I misunderstanding how the date should be input in the 264 #1 field? If not, should I assume it's an error and correct it or assume it's not an error and create a new record?

Answer: In most cases a copyright date can be used in RDA to infer a date of publication if there isn't an explicit date of publication. So, in a 264 1 that inferred date of publication would be bracketed. A subsequent 264 4 could be input with only the $c identified as a copyright date. So, if you find items without the brackets that is an error and it should be fixed or reported to OCLC. Do not input a new record.

Why can't we enter both an ISSN in tag 022 and ISBNs in tags 020 in one and the same record? E.g. yearbook comprehensive record.

Answer: This question has come up before when our validation was previously based on AACR2. In the context of MARC records we took this issue to CONSER and had a discussion with them about what would this mean? Because I could potentially have a serial that has all the ISBN's that are assigned to all the individual volumes. So, this is really the constraints of MARC at this point that is a consideration in the decision to continue how we have done it in the past, which to omit the ISBN's for the individual parts.

Is there a reason 020 $z requires a valid check digit for the ISBN number? It makes it difficult to record information for books which represent a number as an ISBN, but that number is not actually a valid ISBN.

Answer: 020 subfield $z should not require a valid check digit. In fact if you had a number that you were going to include in 020 subfield $a, if the check digit was incorrect that is a case where you would put it in $z in addition to the cases where it is not the number that is appropriate for the item being described in the record.

Since the new data ingest software was implemented, I have had to do way more cleanup in the 6xx fields than I did previously. Could the software be a little more fussy about bringing in duplicated headings and headings that have been controlled in error?

Answer: I agree, it could be more fussy about bringing in those kind of headings. There are some issues that are being resolved about the number of headings that transfer and the state they are in when they do transfer. We are also attempting to clean up problems that we know about e.g. certain combinations of headings. We will try to go after them and get them out of the way.

Omitting the ISBNs for individual parts means you may miss the record if you search by ISBN. This has happened to me and I almost made a new record because of the non-existent of appropriate ISBNs in the record. I hope

this decision will be changed. Making a decision like this because a few records might have 500 ISBNs seems unhelpful.

Answer: The flip side would be the person that wants to catalog the individual volume in a series who would then complain to us that "I searched this ISBN expecting to get the one monograph record but I am also always retrieving the serial and I don't want that" it is a really difficult position and it can be extremely useful but in other cases not so much.

Will it be there validation for other languages in $b in the 040?

Answer: We have validation in place for all languages of cataloging.

With some Arabic records, we see some problems with "dot below" diacritic combined with s, t, d, z. (precomposed character). This cause the "transliterate" macro to fail as well as the ability to control the field in 1XX or 7XX? they pass the validation though!

Answer: They pass validation because all of those characters are now valid with OCLC's implementation of Unicode. The pre-composed characters that have the diacritic combined with the letter are somewhat problematic when using macros because the macro language is not Unicode compliant. There are issues with any macro that are written, including the transliterate macro. We are thinking about possible solutions but do not have any definite plans yet.

I sometimes come across English-language records which have non-English subject headings. Should those subject headings be deleted from the master (English) record, or should they be left in place?

Answer: The language of cataloging code in the 040 subfield $b applies to the descriptive cataloging not to the subject cataloging. So subject headings 6xx can be any languages as long as they are coded correctly.

Speaking about subject headings. It seems like OCLC is now doing some kind of validation process that when run puts the subject headings in order by number 600s then 610s, 650s ... When that is done then the most relevant headings are no longer necessarily among the first headings. Can this process be changed to not change the order of the headings?

Answer: Yes, we would like that process to change. It is the result of records being built in the data ingest process, batchloading, there was also an issue at one point in record manager where these things were sorted into a tag order rather than keeping the most relevant heading in tact.

Does OCLC have guidance for print-on-demand publications? Specifically, printers who print HathiTrust materials and bind the pages and sell them as reprints. These items have no dates other than the original and also often do not mention the printer/publisher. Should I just put the data about the printer in the 037? Thanks.

Answer:Treat the item according to the Print-on-demand and photocopy provider neutral guidelines, which you can find on the PCC website. We will be including information about provider neutral cataloging for print on demand publications in an update in Bibliographic Format and Standards, but otherwise you can search for that on the PCC website and find the guidelines. There will be one record for the reproduction. You would have a 533 field that would indicate that it is a reproduction in print. It would not include any details on its publication.

Sometimes validation finds an error that could simply be corrected by the system itself; for example, in the authority format, "name" may be coded "a" when the 1XX field is coded 110 (and therefore the code should be "n"). Couldn't the system simply fix the problem rather than giving a validation error message?

Answer: Validation was designed solely to report back errors. In a case like this when you have a mismatch between two elements the question is “which one is really wrong?” It may be that the heading that is coded 110 shouldn’t be changed to a 100. It may be that the 110 is correct and name should be coded “n” or it could be the reverse.

Where are we with expanding the number of institutions allowed to do bib record merges? I know a a 2nd cohort has at least been identified.

Answer: Yes, there is a second cohort that started last year and there going great guns. We are planning on starting a third group sometime this summer, at least later this year. That group is still being formed we are very excited about moving forward with that.

It used to be that if one subject heading on a record didn't have a match in the LCNAF or LCSH, no FAST headings at all were added to the record. We're now noticing that headings that match LCNAF/LCSH get FAST headings, while those that don't do not. Was this a deliberate change? If so, we like it.

Answer: Yes, that was a change on how those FAST headings are generated and applied to existing records. It is no longer a requirement that all the heading have to be convertible to FAST. We will do the ones that we can do.

The PCC guidelines for provider neutral records seem to refer only to e-resource items. Is there one for print?

Answer: Yes there is. The thing to look for is Print-on-demand.

How do you now choose the topics for these sessions?

Answer: A couple of the topics that we have done so far have been suggested by our members and we have also picked a few topics that we thought were important. A survey will be coming out soon and we are hoping that you will suggest lots of great topics. Feel free to let us know what you want in the survey, through AskQC@oclc.org or write to any of us individually.

Is OCLC aware of the great number of pairs of UKMGB records, in which one has: 260 $a Place : $b Publisher, the other has 260 $a Publisher, $b Place?

Answer: Yes, we are aware. We have attempted to put together a macro to try and fix them, but it is a very tricky thing to do. In many cases the ones that are in the combination of publisher/place are not necessarily subfielded. We have got commas and things that we have to rely on. We don’t get reliable results all the time. We are also aware that because of some of these having been corrected in the past while we were still batchloading them, that we have 260 fields that did not compare and match correctly so that we have lots of duplicates as well. We are aware of this and working to clean up as much as we can.

Could there be validation on incongruence of 337-338 and 008/23 (Form of Item) = 'online'?

Answer: This is a case where since you could include multiple 33x fields for different aspects of an item. It might be difficult to include this kind of thing in validation and it may be better for us to look for these things in the database. This one occurs frequently for Hathi Trust and Google Books because existing records are simply cloned and the 337 and 338 are not removed and replaced with their online counter parts.