2020-08 VAOH Session

Presentation summary

Unraveling the mysteries of a merge and DDR improvements.
Cynthia Whitacre introduced the session, informing participants that today's topic, and the topics for the remainder of this year's sessions, were selected from topics suggested from past user survey responses. Shanna Griffith presented a background view of merging and Metadata Quality staff's use of Journal History (an internal tool) to investigate possible incorrect merges. Laura Ramsey presented on improvements to DDR (Duplicate Detections and Resolution). Joining them during the Question & Answer section were Robert Bremer and Jay Weitz.

URLs mentioned during the presentation:

"Cataloging Defensively" presentations
https://help.oclc.org/WorldCat/Cataloging_documentation/Cataloging_defensively

Bibliographic Formats and Standards (BFAS), Chapter 4
https://www.oclc.org/bibformats/en/input.html

BFAS Chapter 5: Requesting Changes to Records: Reporting errors, etc.
https://www.oclc.org/bibformats/en/quality.html#requestingchangestorecords

Detailed information for using LC authority history begins on p. 27/56 here: Connexion Client documentation: Authorities: Search authority files:
https://help.oclc.org/@api/deki/files/5151/connexion-client-search-authority-file.pdf?revision=2

Member questions

2020-08-04

Does the recovery process include correcting the incorrect coding that caused the merge in the first place?

No, it does not. Usually if we can discover what caused the incorrect merge, then we make corrections to the record manually after the recovery process.

We try to learn something from every incorrectly merged set of records. Obviously, if the incorrect merge was caused by incorrect coding, that's one thing. But, if it's something else, that suggests something more systemic or something that we overlooked or not treated better, we try to learn something from that and go back and do our best to build into DDR ways to avoid making the same mistake in the future, if possible.

While recognizing you all are super busy, would it be possible to let a requester know when an incorrect merge is fixed? I made a request on June 29, the record was corrected on July 23, but I didn't think to check the status until today.

Yes, we do try to let you know when the records have been pulled apart so that you're able to make any corrections on your end, or add your holdings to the appropriate record once they've been pulled apart, etc. Unfortunately, that doesn't always happen, but we do try to get back with you and let you know.

As records are submitted (manual) to QC to merge, does QC staff keep notes to help improve DDR?

We watch the process as we merge manually, and we have a match tool where we can input record numbers tool to see what DDR would do with the set of duplicates. We provide that kind of feedback to the DDR team. It depends on what you notice when you're merging and the time that you have stop and take a look then follow through with the investigation. As was mentioned earlier, we are continuously working to improve DDR.

Do you ever call in an item to see exactly what is being cataloged or do you just review the records in question based on what is in both records?

We're not a holding library, so we do not have access to materials. Sometimes we are able to view item information, including full text, from various websites, but that's not always the case. Consequently, if you report an error such an incorrect title, wrong paging, or bad publishing information, we do ask that you provide scanned copies of the item as proof. That way we're able to make those corrections based on what is on the actual item. But no, we do have to take what is in the records unless we are able to find enough information for the item on the internet.

Would you merge an AACR2 record with an RDA record?

Yes.

Is the list of what constitutes a significant change (that would trigger the DDR) available online?

I don't believe that that's ever been online anywhere. It doesn't really change because we’ve used the same comparison points in DDR for many years. So, it seems like it's something that we could consider adding to the documentation in the future.

If you have a record that's online and there's a change to it, if the change happens in a field that we never look at, in terms of a comparison, then it doesn't make sense to necessarily put it through DDR because it's already been through that process in the past. We concentrate on those kinds of changes that are made in fields that we look at. So, if the wording in the 245 is corrected or the coding in the 245 is corrected, we compare the title fields, of course, so that's something that would then go into the DDR stream for processing. And then, it would be looked at seven days later, but it would be in that processing stream. Place of publication, publisher, changes to date. Whether it's the date in the fixed field, or the date in 260 or 264 subfield $c, changes to the extent, changes to the size. All of those are the kinds of things that that would trigger a record going back through DDR.

If we know there are a category of records that have a high number of incorrect merges (CIA maps where the only difference is the presence of relief), is there a way to get them unmerged en mass?

If you want to send a list of the OCNs that have been incorrectly merge to the bibchange email address, we would be happy to look into them. If they have not been merged too long ago, we'd be happy to have them recovered. A lot of times with these when the presence of relief is the only difference once they're unmerged then supplying an edition statement in brackets to both of the records will help prevent them from being merged in the future. There have also been improvements made to DDR so the numbering in quoted notes in these types of records is taken into account. Sometimes when the only difference between some of these maps is the presence of unique numbering, if those are added as quoted notes in 500 fields, that is taken into consideration.

We were in close consultation with the maps and cartographic materials community in making a whole bunch of improvements to map matching, including things like looking for various constructions of dates in notes, especially in quoted notes, looking for unique numbering in quoted notes, or in non-quoted notes, various other things as well. It was also in consultation with the cartographic community that we changed the date, that is to say, we automatically do not merge records for maps that were published previous to 1900.

You may be familiar with the cataloguing defensively series that we've been putting together for basically the past decade. There is a cataloging defensively presentation specifically about maps and it gives all sorts of hints about how to make a record unique so that it will not merge to another record that is very similar, but distinct.

What's the time limit on recovering merged records? That is, are merged records not recoverable after a certain amount of time has passed?

That is April of 2012 is the limit. Journal History keeps a record of transactions that have happened since April of 2012 and so anything that happened prior to that we would not be able to view or to recover.

There are large numbers of UKM records which were apparently batch loaded three or four times with variations which likely resist DDR, e.g., run-on 245 $a, edition in 245 $a, 260 with place and publisher transposed, garbled 300 fields. Could DDR be customized to target and merge these so that the best record in the set becomes a better bet for getting all of the records merged into a record from a different source? Or am I making wrong

assumptions about these? (Why? These records triple or quadruple the work of manual heading maintenance.)

Yes, we have been talking about these lately. DDR is not easily customized. It's designed to deal with all sorts of situations that exist in bibliographic records, in terms of correct coding to incorrect coding and various kinds of issues with the way the data is formulated. But when it comes to these particular records, and what you're really talking about are the ones that have an 040 field that will say UKMGB, is that they are a result of a retrospective conversion process by the British Library. And the data is mixed around in fields in a way that DDR cannot handle. So, one of the more typical problems is that you see the paging in 260 subfield $a, rather than in field 300. And then the indication that the book is an octavo is also in subfield $a in 300, rather than in subfield $c. Those kinds of issues really do get in the way of DDR and have to be dealt with in a whole different way. We had a similar issue with records from the Bavarian State Library in the past, what we did was use a macro to look at different pieces of information and basically sidestep DDR and do a sort of a quicker evaluation of whether the records were duplicates and merged the lower quality record out of existence. Something like that could happen here, but we've also been thinking about possibilities of getting replacements for these records that we would perhaps match on the number that's supplied by the British Library in 015 or 016. That kind of thing could maybe take care of the problem once the data is cleaned up then possibly these records could be processed by DDR and merged. In a lot of these cases, we go looking at these records, look at the messed up record, go looking in the database for the same resource, and find that there is a duplicate - it's just that DDR couldn't match it to that record. So, yes, we are aware of these records and trying to do something to take care of them, but they'll probably be around for a little while longer.

If the same type of resource (e.g. HeinOnline cataloging) which are all batch loaded from the same institution, and have the same error (e.g. 260/264 with $a indicating no place of publication, but $b has a location and the publisher), if several of these have been reported, does OCLC note that this may be a repeated error and look for more of these from the batch? Or does OCLC require new error reporting for every item found.

I would recommend calling that out when you report it, so that we're aware. We don't necessarily go out looking for more duplicate records when we see a particular cataloging issue. Something to note is a limit to how many records will match. That limit is twenty records. Which means, if there are more than twenty records with this problem, even if we tried to send them through DDR, even if they are seemingly identical, it's possible that the twenty record limit is getting in the way of those getting merged. So that's always a good thing for us to know as well as maybe there's something that we can do to take care of these duplicates and fix the records in a different manner.

Is DDR flow automatic? Can anyone suggest a merge and how?

Yes, the DDR flow is triggered by significant changes, or if it's a new record added to WorldCat, We do have a process here that we can use to feed records into the DDR flow, but that's not something that's available externally. But you could also call that out, if you're reporting something and you feel that those records are identical, so they should be merged, it's something that we could get into the flow. If anyone does spot duplicate records, they can report those to bibchange@oclc.org.

This may be an issue for holding library to do their record maintenance. Since DDR is very dynamic, it is possible that the holding library may have obsolete Connexion records. Is a recommendation from OCLC to the individual library to update their records with the latest Connexion records. I understand it could be a local issue for the individual library, but I just wonder about the OCLC's recommendation.

I think this is talking about when you're getting WorldCat updates through Collection Manager. And so, whenever we are merging records here, then you're getting an updated record through that feed for your library. It is totally up to you, whether you change that in your local catalog or not but certainly, if you want to keep things up to date, using that Collection Manager, WorldCat, updates feed is a good practice.

Even if your holding was on a particular record that got merged to another, the control number is still indexed even though it's not the main record in WorldCat. The control numbers of the merged records are retained in Field 019 for the very reason.

How do you handle records of the same item with different languages of cataloging?

These are actually not considered duplicates. They’re what we call allowable duplicates and are not merged. Duplicates that use the same language of cataloging can be merged. But if you have an English language of cataloging record and a German language of cataloging record that are duplicates for the same resource, those are not considered to be duplicates.

We use WMS. Do we have to update our records when they affect Collection Manager? I have reported duplicate records for journals we own in Collection Manager when the records were merged, were our E holdings transferred to the merge record?

Yes, when records are merged, all of the holdings are then merged into the retained record. If it's an incorrect merge we send them to be recovered, then once the records are reinstated the records are entered back into WorldCat as separate records with their respective holdings intact prior to the merge.

Hopefully, this isn’t too specific a question for this platform. If so, just pass on my question. I am looking for ways to protect records against being identified for DDR when the difference is not necessarily represented anywhere but the 300. Ex: same title in two braille languages such Unified English Braille vs American Braille English edition. Since I am creating both records myself and have the opportunity, is there any advice on how to protect these records?

If the resources themselves do not have edition statements to the effect that you've indicated here, you can legitimately under both AACR2 and RDA add a cataloger-supplied edition statement that will differentiate the two records. If it's not stated on the item, you could bracket unified English, Braille edition as a 250 on the appropriate record and bracket American Braille English edition on the other record. And the 250s will be compared against each other and DDR will not merge them again. This is an example of the kind of thing that's dealt with in the Cataloging Defensively series. So, you may want to take a look at that.

What do you need when we manually report duplicate records? I usually use Connexion Client >>Action >> Report error. I usually put the message "Please merge with #12345678" (The "report error" command automatically sends the record number that you are viewing to QC.) Is this an OK way to report duplicates?

Absolutely. We will take reports for duplicates anyway you want to send them even if it's just a plain email to the bibchange address stating what the record numbers are and that they may be duplicates. The method you've been using is great. The window that opens when you choose Report Error does send us an image of the record as it appeared when you filled in the report.

Regarding e'holdings, are the merged OCLC numbers changed in the Knowledge Base?

The process that pulls merged OCNs runs nightly.

Non-DDR question: For bib records that have a 776 pointing to another version of the resource, if we change the form of the main entry in the 1xx, are we also expected to change that heading in the 776 $a to make them match?

It's not a requirement, but it would be good if that change were also made in 776 field. Also, if your workflow permits, it is best to call up the record that is cited in the 776 and make the change to that record directly as well, but it's not required.

A real quick way to update the 776 is to use “Insert from cited record” under the Edit menu in Connexion. You could just pop the OCLC control number in that field and update it that way.

When a heading in 1XX in a name authority record is changed, how long should it take for controlled headings that are linked to it to be updated. I've seen cases that take weeks.

Normally it should not take weeks. If there is some kind of issue within the NACO nodes and LC in terms of getting records distributed, there could be some delay in processing. On our side if we receive an authority record that's been updated by the Library of Congress, and the heading is changed once we load the record it will normally take 48-72 hours to get the change made across the database where headings have been controlled to that authority record. That's not to say that it's a perfect system, sometimes there are various kinds of issues. It may be worthwhile, if you've noticed that we have loaded a record or that if you're working in NACO you've made a change to a record and that change has not been propagated across the database, let's say after a week or so that you might send us an email to say, is there a delay or something like that? So that we can investigate what's going on because it is unusual for it take more than 72 hours for the process to complete.

A library which shall remain nameless has been adding local 856 fields to EEBO microfilm records. Could I send one example and you can correct the rest?

Absolutely, we do that kind of work all the time. So please send it to us, we'll take care of it.

Could OCLC look for 856 fields that have URLs that start with "proxy" and are clearly local URLs, could these be batch removed?

If the URL is one of these proxy URLs where the real URL is embedded in a longer URL, we have some coding in a macro to transform those into what should be the real URL, which then is oftentimes a duplicate of a URL that's already in a record that causes them to be collapsed into a single field. Requests like this for us, would mean, perhaps doing a database scan and running our macro through a set of records to try to clean them up.

If you have a set of records that you wish us to look at, send them to bibchange@oclc.org and we'll take a look and see if we can do some sort of batch processing on them to fix them.

2020-08-13

Do you want catalogers to report duplicate Dublin Core records? Are they being deduped by DDR? Sometimes I see duplicates. Some are character-by-character duplicates and others have the same data, but fields are in different order. When I reported some Dublin Core duplicates earlier this year, I was told no action is taken because OCLC policy is that Dublin Core records can co-exist with regular MARC records, but that is not what I meant.

DDR does not process the Dublin Core records that you see come through as part of the Digital Gateway. So, they are not being de-duped, and we don’t merge those manually either. Instead, at this point, we essentially warehouse those records in the database because libraries will go ahead and harvest data. They get added to the database and then later they can be taken out and reinserted again. So, they’re sort of a different category of record than the traditional bibliographic records that you find in WorldCat.

We do not manually merge them because of the re-harvesting. If we were to merge them then they could possibly just re-harvest and another record be added to WorldCat

How about DDR records in other cataloging languages? I deal with a lot of those that have $b spa and they are clearly dups in the same language of cataloging.

Yes, all records get considered for DDR regardless of the language of cataloging, other than the exceptions listed. We merge all different languages of cataloging, the rules do not really change how records are considered for merging it’s all the same, independent of the language of cataloging.

However, records will not be merged across languages of cataloging, so a Spanish language of cataloging record will not be merged to an English language of cataloging record. But within the language of cataloging they will be merged.

Is the matching algorithm used for DDR the same as used in matching for Data Sync? I'm wondering if there are discrepancies between these two systems. For example, could I send a record via Data Sync, resulting in the creation of a new record in OCLC (i.e., no match); then, days later, this original record is merged to a pre-existing OCLC record as part of DDR?

The algorithms used for Data Sync are different, there aren’t many similarities, but they are different, they have different purposes. So, yes, it’s very possible for a record to be loaded via Data Sync and then a week later be merged by DDR.

When you say that newly added records are added to the queue in DDR within 7 days, does that mean that they are actually evaluated by DDR within that time, or does it take longer for that to happen and records to potentially get merged?

It may take longer; it really depends on how much is in the flow already for DDR to work on. So, if we happen to have a higher amount of records that were added within a certain timeframe that may slow it down a bit. But it’s generally within seven days.

How does DDR work in tandem or separately from the Member Merge project?

With the Member Merge Project, the participants are actually merging the records in real time. So, the merging is happening instantaneously. They’re comparing the records and then going through the process of actually getting them merged. DDR is the automated process. The two are completely separate and different.

How many bibs does DDR evaluate on average per day?

Just looking at our stats, for example, in July they were seven point two million records that went through the DDR queue.

It's usually between five and eight million records each month going through the DDR queue. So, we do examine a lot. Based on what was said, that means if a new record is added and a new OCLC number is created, then merged seven days later it shows which OCLC number is kept.

The OCLC number that's kept is the one that belongs to the record that ends up being retained. Whichever record makes it through the criteria in terms of the record retention hierarchy, is the one that's kept, and that number remains in the 001, the records that end up being deleted are in the zero one nine field. Most likely if everything is otherwise the same, you have a member record versus another member record, that was just added through Data Sync, probably the existing record in the database is going to be the one that's kept because it's been there longer and it had more opportunity to pick up holdings and possibly have been enhanced, but it could go the other way around. It’s not necessarily the case, I mean, if the incoming record is a far better record, more complete, and the existing database record doesn't have very many holdings and is really sort of skimpy in terms of the description, the number of fields, etc. then the new record could be kept, but the number that ends up being retained is based on which record is kept.

Is there a table for automatic/manual merge? What it is the dup merged? i.e. is there a table for what kind of records we keep?

There is a hierarchy of records that is used in automated merging to, for instance, keep a Conser record over an ordinary serial record. When it is member to member record, we look at the number of fields that are present and the number of holdings, and then decide on which record to keep on that basis.

On the other side of the aisle, am I correct in assuming that you are still detecting duplicate authorities and reporting those to LC?

Yes, we report any duplicate authority records that are reported to us that are from users that are not NACO participants. We report them to LC on behalf of the library because only LC staff can merge/delete authority records as the LC/NACO Authority File is the Library of Congress’ local authority file.

There's also a report that OCLC generates and sends to LC monthly for duplicate name authority records that are exactly the same.

We've found records with dozens of OCLC numbers in the 019 field are there that many duplicates?

Yes, there are unfortunately. We do have cases where records that have just slight variances that would not get triggered or caught by our process and then, therefore, get merged at a later time either manually. Or we have cases where they are identical. They get added via Batch and they are identical, and they do end up getting picked up with DDR and merged at that time.

It should also be mentioned that in a single DDR transaction there's a limit of twenty records being merged into a retained record and if it goes above that, we set it aside and the merge does not happen.

We are allowed to modify records to suit our local needs by including data that should not be added to the master record, such as coding subject headings as LCSH or LCGFT when they are not valid headings in either. Then the record goes through data sync and that "bad" data ends up in the master record anyway. Why?

They're actually what we call field transfer rules, but subject headings may transfer to the WorldCat record if there's a scheme of subject heading on the incoming record that isn't present in the WorldCat record. So if, for example, the incoming record had a Medical Subject Heading (MeSH) and the existing WorldCat record had only Library of Congress Subject Headings (LCSH) then that MeSH heading would transfer over to the WorldCat record. If the existing heading already had MeSH headings on it, then the MeSH on the incoming record would not transfer.

We have worked very hard to improve what transfers and what does not transfer. So hopefully for the newer records that are being added, or loaded, through Data Sync that's getting better and we aren't transferring as much as we used to.

User comment: An example of that: Piano music coded as lcgft, which it isn't.

… suggesting that people ought not to be coding those things, lcgft, in their local system either. Probably better, if your local system will permit that, to code those things as local, when they go into your local system, if that works in your display.

It doesn't seem like a lot of people will take Library of Congress subject headings that they then intend to use as form/genre terms and automatically add subfield $2 lcgft, when, in fact, they really should be consulting that particular terminology, to make sure that it's there. It is sort of a problem for us, because our software just looks at that subfield $2 code, and looks at the record that is being retained in the database, either in a merge or if it's a case of a record that's coming in through Data Sync where we might transfer the lcgft heading, and if the retained record doesn't already have any lcgft headings, then it will transfer. So, you can take a record that was actually okay and sort of mess it up by transferring a heading that is not Okay. But yes, careful coding is always needed.

We just found records that had incorrect editions statements. What should be in the 538 field in records for DVDs. Is that because edition statements are a field that's transferred?

Is that because edition statements are a field that is, I'm assuming, that the local system wanted that more visible?

We were once told that the 250 is important field for DDR and that catalogers could "supply" 250 fields to prevent merger, defensive cataloging. I can imagine this on DVDs

The edition statement, field 250, is not a field that will automatically transfer, but 538 field is one that is that will automatically transfer, if it's not already present in the retained record.

More conversation about the DVDs and the edition statements: the person who asked the questions states that she was referring to two fifty fields that had things like widescreen and full screen.

And then someone else added: We were once told that the 250 is important field for DDR and that catalogers could "supply" 250 fields to prevent merger, defensive cataloging. I can imagine this on DVDs

Yes, that is true. You can add a cataloger-supplied edition statement in Field 250 to help prevent records that are otherwise exact in their descriptions from being merged.

The Cataloging Defensively series can be very helpful in giving you hints about creating records or editing existing records in a way that they will be distinguishable by DDR from similar records to which they should not be merged. So, you may want to take a look at the whole series of Cataloging Defensively webinars that are available from the OCLC website.

Why it is that reports of dups for merge takes so long?

Unfortunately, duplicates are a problem and we do realize that, and we do have a substantial backlog of duplicates. We are working as best we can to get through them, but they do take time to go through as we have to analyze each one and make sure that they are duplicates and merge them accordingly.

I manually merged two bibliographic records (as part of OCLC Member Merge) and an incorrect 650 field got saved in the retained record that "validation" would have flagged. I then realized that the merge process does not perform a "validation" check before merging. Am I correct about this?

Yes, you would have to do a manual validation on the records or record after you've merged it, there's no validation built into the merge process.

But if you are a participant in the OCLC Member Merge Project, you should always check, as we do when we do manual mergers, always check to make sure that everything that's transferred is something that should've transferred. You can clean up the record afterward the merge, and we encourage you to do so.

What about libraries use 830 field for their local collection title and add a subfield 5, the code of their library? The system simply accepts it. But we should not use 830 for a local collection title, right?

No, they shouldn't. Those can be reported to us if you're not able to remove the local series and you can report those to us if you're not sure you want to remove it and we will take care of it.

Is there an error that is a common trigger for DDR? Just wondering if there is a particular mistake catalogers make that we should pay more attention to.

One example of that is when catalogers enter a record for the online version and forget to code Form of item: o in the fixed field, or in the 008 field. That can trigger DDR because the lack of that code makes DDR think that both records are for print.

DDR can get confused by contradictions within a particular bibliographic record. A contradiction between a 260 or 264 subfield $c and the Fixed Field date, for instance, or a contradiction between the place of publication in a 260 or 264 subfield $a, and the country code in the Fixed Field, things like that. So those are particularly important to pay attention to: contradictions within the record.

We have a comment saying Bravo for Bibliographic Formats and Standards Chapter 4, which has the section on When to input a new record. It helps to determine whether there are duplicate records.

We're glad that's helpful. Chapter 4 in Bib Formats and Standards is written to reflect what DDR does, and DDR is programmed to reflect what Chapter 4 states. They really are supposed to be mirror images of each other. They should both be doing the same thing.

Does DDR treat records coded dcrm_ differently from other records, or are they put through the same system?

Actually, DDR tries not to deal with records for rare and archival materials. There are 25 different 040 subfield $e descriptive cataloging codes including DCRM. If DDR finds one of those, it will set that record aside and not deal with it at all. DDR won't merge those records. We leave the merging of duplicate rare material records to actual human catalogers.

Comment: Multiple editions (often gov docs) published in same year with different transmittal dates often get merged if a 250 isn't supplied to keep them from merging

Actually, we do look at field 500 to pick up on those kinds of date differences, particularly in the case of government documents, you might have a date like that that's in quotes. So, if you had something that said April 15, 2020, and something else that said, May 22, 2020, we should be alert to that kind of thing, and be able to differentiate on that basis. Although it probably is a good idea to have field 250 in that case.

I've had multiple dcrmb records merged incorrectly multiple times until a 250 was added...is this setting-aside a new practice or must that have been done manually?

That could've been done manually. We have to look at the records in Journal History to see how they were merged if it was a DDR process or if it was a manual merge. But DDR does not merge rare materials records.

What is your preferred way for us to report dups? Using the error report feature in Connexion? Or?

Any way you want to get them to us, BFAS Chapter 5, Reporting errors will show you the different ways that you can submit them. But if you just want to put the record numbers in an email message and shoot it to the bib change email address (bibchange@oclc.org), we'd be happy to take those too.