Metadata Analysis for Pre-Migration Cleanup

1. Metadata Analysis for Pre-Migration Cleanup

Hello everyone, and thank you for coming to my presentation on metadata analysis for pre-migration cleanup. I want to let you all know that these slides and a transcript of my talk are posted at http://tiny.cc/AmigosMigration.

If you've tuned in today, I'm guessing that there might be a migration in your future. So I thought I'd start with a quick poll to see…

2. Where Are You?

…where you are on a migration journey. Have you recently completed a migration? Maybe you're in the middle of one right now, or are getting ready to start a migration soon. Or maybe there's no planned migration on the horizon.

[review poll results]

If you've already migrated, you might be finding metadata in your new system that needs to be cleaned up. And even if you're not planning on a migration any time soon, it's never too early to start cleaning up your metadata.

3. Our Recent Migration

You've already heard a bit about me in the introduction, and I want to also give some background about the migration that my library recently went through. James Madison University Libraries migrated from Innovative Interfaces' Sierra system to Ex Libris' Alma and Primo VE in the first half of 2020. We had an 11-member group, which we called the Alma Implementation Team, that was responsible for all the tasks associated with the migration. To give you an idea of the size of our database, we migrated over 1.5 million bibliographic records, over 1.6 million item records, and about 5,600 [five thousand six hundred] checkin records for print serials. About two-thirds, or just under a million of the bibs represented electronic resources, or a combination of print and electronic, and those records undergo some additional conversions during an Alma migration. I'm going to focus today on just those types of records (bibs, items, checkins, and e-resources), because that was the scope of my work. In addition to serving on the Alma Implementation Team, due to my role as Metadata Analyst Librarian, I was also responsible for identifying issues in our metadata that would cause problems during the migration and then delegating those to the appropriate people to be cleaned up.

4. What Is Metadata Analysis?

What do I mean when I use the term "metadata analysis"? You have probably heard by now that metadata is "data that describes and gives information about other data."

When it comes to the term analysis, a definition that I think works particularly well when talking about metadata analysis is this one: a "detailed examination of the elements or structure of something, typically as a basis for discussion or interpretation," and also "the process of separating something into its constituent elements." So we're taking something, in this case metadata, and breaking it down into smaller parts and closely examining those parts. The reason this is so important when talking about

metadata cleanup is that you first need to perform analysis in order to accurately identify the problems that you'll be cleaning up. If you want your remediation efforts to be effective, you need to do metadata analysis to inform decision making and choose a course of action.

5. Overview

When undergoing a migration, there are a few tasks involved in data cleanup. You need to be able to assess your metadata to identify problems, prioritize which of those problems to work on, conduct analysis to identify errors in specific elements and values, and then perform the necessary actions to remediate the data. I'm not going to talk about the last step of doing the cleanup today, as the steps involved are specific to the particular system that you're using. We're going to cover just the first three. There should be time left at the end for questions, so feel free to type those into the chat box as you think of them and I'll address them at the end.

6. Metadata Assessment Criteria

The first step in the process is being able to identify various types of issues with metadata quality.

7. Criteria for Metadata Quality

The Digital Library Federation has put together a Metadata Assessment Framework that defines seven criteria for evaluating the quality of metadata in digital collections. These criteria also apply more broadly to metadata for other types of collections, including the electronic resources and physical materials that make up a large portion of many library collections. The criteria come from a chapter by Thomas Bruce and Diane Hillman called "The continuum of metadata quality." The definitions I'll be sharing today are from the DLF Metadata Assessment Framework.

8. Criteria for Metadata Quality [highlighted]

I'm going to highlight four of these criteria that I think are the most relevant to what you might be looking for when you're doing cleanup for a migration.

9. Completeness

The first is completeness, which refers to whether "the element, property, and/or attribute is present." You need to know which fields are required in the new system, and then check to make sure that all those required elements are present in your data.

10. Accuracy

The next element is accuracy, which means that the "information is correct both semantically and syntactically." Your metadata has to have the correct structure (syntax) and also mean what it's intended to mean (semantics).

11. Conformance to Expectations

The third criterion, conformance to expectations, refers to whether your metadata "values adhere to the expectations of your defined user communities (both internal and external)." The metadata should conform to standards like MARC as well as any local guidelines that may exist.

12. Consistency

And finally, consistency: "semantic and structural values and elements are represented in a consistent manner across records. Values are consistent within your domain." Any metadata value that's used in multiple records should be used in the same way across all those records.

13. Prioritization

As I've described these four criteria, you've probably been thinking of a number of things that would fit into each category. It seems like there's always more cleanup to be done than there is time to do it in. So it's important to prioritize the time and resources you have. But if you have all these things that you could potentially clean up, how do you know what to tackle first?

14. Keep the Goal in Mind

I think the most important thing you can do is to always keep the goal in mind. And the fact that you're thinking about this cleanup in the context of a migration is a great help. The goal is to make sure that your data is in a format that will migrate the way you want it to, that the meaning of the data isn't changed in the process, and to minimize data loss. If your migration involves converting metadata into a different metadata schema, you'll never be completely free of data loss; instead, you want to minimize the loss and be able to intentionally choose what will be lost.

To help with prioritization, I'm going to share a series of questions to think about that will help you determine what the impact of various cleanup activities might be on the migration outcomes, and some examples of each that we found in our own migration. As I go through them, you'll see that many of them line up with the four criteria I just covered.

15. What Records Are Migrating/Not Migrating?

The first question is "What records are migrating, and what records are not migrating?" Maybe you're planning to migrate every single record in your current system, in which case you have an easy, straightforward answer to this question. But if there are some records that you don't want to migrate, do you know how you will be identifying them? This is one of the first things to figure out, because there's no point in spending time cleaning up records that won't be part of the migration.

For example, maybe you have records for missing, lost, or damaged items, and depending on how long they've been unavailable, you might not want to bring those along to the new system. In our old catalog, we were required to keep records for withdrawn items for two years, so we didn't need to migrate the older year which would have been purged from the system at the time of migration anyway. Another area to consider is brief records. You may have brief records representing items that were ordered but then got cataloged on a different bib that don't need to be migrated. So a good place to start is thinking about what is migrating and not migrating and how you will identify those items.

16. What Data Is Missing?

The next question, which comes from thinking about the criterion of completeness, is "What data is missing?" You'll need to figure out what fields are required in the new system, and then check whether they are present in all records that will be migrating.

This will probably be fields like the 245 field, record identifiers, call numbers, and barcodes for physical items.

17. What Data Will Be Lost?

When thinking about conformance to expectations, that prompts the question "What data will be lost?" This is most likely going to be data that can't be migrated in its current format or field. As you saw with the previous question about missing data, you'll need to know the schema used in the new system, and

then consider any data that doesn't fit that target. This is also a good reminder to look at any locally defined fields or locally defined values in standard fields and ensure that those can be mapped to the new system if necessary.

In our Sierra system, we used non-MARC fields in certain types of bib records, and so we needed to check that they would be mapped to the correct MARC field when extracted. We also had some locally defined material type values in a Sierra field that were used instead of the material type from the MARC Leader, which wouldn't be retained in the migration. So those are two examples of data that could be lost.

18. Where Will Data's Meaning Be Incorrect or Unclear?

The next question is "Where will the data's meaning be incorrect or unclear?" This gets at the accuracy of your metadata – does it mean what it's intended to mean, and will that meaning still be correct in the new system? One example of this that I came across in our metadata was that the majority of our bib records lacked an 003 field, which is the control number identifier. We had a variety of types of control numbers in our 001s, and without the 003 we had no way to know which type of identifier the 001 contained.

It's also important to check for field scope creep. We had used the location codes in our old system for many things that weren't actually physical locations. Some of those were indicating whether e-resources had been purchased by our institution or by our consortium, or whether materials were oversize. In these cases, the data could have been lost on migration, as I mentioned in the previous slide, or it could have caused some unnecessarily confusing or inaccurate results if not modified.

19. What Data Inconsistencies Will Affect Migration?

When working with metadata, it's not uncommon to find inconsistencies that could be corrected. The important thing is to ask "Which of those inconsistencies will affect the migration?" They could have an impact on which records migrate, or whether records migrate as intended. One key place to check for consistency is between parent and child records, meaning a bib and its attached item records, or a bib and its attached checkin records. You may also have sibling records, such as item and checkin records that are attached to the same bib. If there's a shared code that appears in both types of records, you'll want to check that they match.

We used a code to indicate withdrawn items, and that code needed to appear on both the item record and on the bib that the item was attached to. Because the bibs and items would be extracted separately at the time of migration, if one record had the withdraw code but the other didn't, the one without the code would have migrated unnecessarily. Checking for consistency was also important in identifying our electronic resources. We used an "Internet" location code to identify those records, and thus needed to check that anything with an Internet location had a URL in the 856 field, and vice versa. A final example that highlights the problems that can arise from inconsistencies between sibling records (checkins and items) is that location codes were used to match item records to the correct holdings record during the migration. So if the location codes didn't match between the two record types, we would end up with extra holdings records. I'll talk more about this example later.

20. What Will You Need after Migration?

And finally, it's helpful to think about what you'll need after migration. Think about your current workflows and what aspects of those processes will be different in the new system. For example, the options for match routines for loading records were more limited in our new system than they were in the system we were coming from. I verified that the fields we used as match points for all of our regular record loads would still be available as match points after migration, and that they would function in the same way.

Another thing to consider is additional cleanup that has to be performed after the migration, and whether there's anything you could do now to facilitate that cleanup. After our migration, we needed to be able to pull together all the records from each of our electronic resource collections in order to group them as collections in Alma. To facilitate this process, we moved our collection identifiers to a different field and requested that that field be indexed in the new system so we could easily pull those sets together.

21. Prioritization

So to recap, here are the questions you can ask yourself to help identify the areas that would be most important to work on cleaning up before a migration. Take a few moments and let me know in the chat which of these questions have pointed you towards areas to investigate in your own metadata, or cleanup that you'd need to do for your migration.

[review chat responses]

The variety of answers here illustrates a final point I want to make about prioritization, which is that there's no universal number one priority. Everyone's metadata and context is different, so something that might be the top priority at one institution might not matter at another, either because they don't have that particular issue in their metadata or it may not have the same consequences in their environment. These questions give you the tools to help identify what's important in your unique situation.

22. Metadata Analysis Techniques

So now that you know what issues to be looking for in your metadata and have some questions that can guide you in narrowing down which ones are important to tackle before a migration, I'm going to talk about three techniques you can use to dig into your metadata and find those problems.

I'm going to cover these techniques from least complex to most complex, and I would encourage you to choose the simplest technique that can efficiently solve your problem. For example, I could write a Python script to identify a particular problem, but it might also be possible to find it with a simple search, so take that route instead. Don't do the more complicated thing if you don't need to. Given the constraints of a migration timeline, look for the simplest solution that allows you to analyze your metadata accurately and efficiently.

23. Search Your System

The first technique you can use is whatever advanced search or querying capabilities you have in your current system, anything that would let you create searches using multiple criteria. In Sierra, this is the Create Lists function, for example. If you're using another system, feel free to share in the chat what this is called so others can benefit. You can use these advanced searches to pull together a set of

records that have a particular problem. This works especially well for finding required fields that are missing, and for identifying brief or incomplete records.

24. Search Your System – Examples

In our own migration, I used this technique to find bib records that didn't have a 245 field, because that was one of the fields that was required for the system we were migrating to. I did the same thing for the 001 field, which we used for OCLC numbers.

For identifying brief records, you'll need to have a field or combination of fields that would typically be present in any fully cataloged bib, and then search for records that are missing those fields. For example, maybe the lack of subject headings would indicate a brief record. I was able to search for records that had a receive date but were missing a cataloged date. Many of them turned out to have been fully cataloged on another bib, so we were able to exclude those short records from migration. That gets back to that first prioritization guide question about whether you can identify what will and won't be migrating.

In doing any of these searches, remember to always limit your search to just the records that will be migrating, because we're not interested in spending our time cleaning up things that won't be part of the migration.

25. Scrutinize Spreadsheets

Using the built-in searching capabilities of your system works well for simple things, but in many cases it's useful or even necessary to have a bit more context around the particular issues you're looking for. So you can use that ILS search and then take the next step of exporting metadata from the ILS in order to review it in a spreadsheet. This allows you to more easily look across multiple fields while not having to open and address each record individually.

This strategy is useful for a number of things. It can help you find inconsistent values and multiple occurrences of fields that shouldn't be repeated. It can also be used to identify coding mismatches within a record or between parent and child records. That's an issue that could also be identified by searching the system, but you might have to do a separate search for each possible combination of codes, which could be time consuming, or there might be other fields that you'd want to check to confirm what the correct data should be. So having the ability to view multiple fields from many records at once is helpful.

26. Scrutinize Spreadsheets – Example 1[a]

As one example of finding inconsistent values, I exported our OCLC numbers, which were in the 001 field, and then sorted the list in Excel. What you see here is the end of that sorted list, and a few things stand out.

27. Scrutinize Spreadsheets – Example 1[b]

The first three are OCLC numbers, but then there are a number of things that stand out. There's one number that's too large to be an OCLC number, and a few fields that aren't OCLC numbers because they're not strictly numeric. There's a blank cell that just contains spaces and some others where the field is not present at all. So you're able to pretty quickly find these inconsistent values from sorting the list and glancing through it.

28. Scrutinize Spreadsheets – Example 2[a]
Another thing you can do with spreadsheets is look for multiple occurrences of non-repeatable fields. In this example, I've exported a list of barcodes. By double-clicking on the edge of column B, Excel will resize the column to fit the width of the data it contains, which lets you see that there must be some barcodes that are longer than the standard barcode.

29. Scrutinize Spreadsheets – Example 2[b]
Then you can filter to show only the rows that have a semicolon in the barcode column, and that gives a list of the items that have more than one barcode that you'd need to clean up.

I'll also point out here that this is a great example of the sort of thing you'd want to clean up at any time, not just when you're going through a migration.

30. Scrutinize Spreadsheets – Example 3[a]
As a third example using spreadsheets, you can check whether codes are consistent across parent and child records (bibs and items). Column B shows the withdraw code on the bib, column C should have the same code as column B, and columns D and E should both have a note indicating "withdrawn." So while I could have just done a search within our ILS to find records with z withdraw code in the bib and no z withdraw code on the item, being able to look at these other fields at the same time is helpful for context and for identifying which part or parts of the record need to be fixed.

31. Scrutinize Spreadsheets – Example 3[b]
You can also use formulas to help you identify problems. In column D here I've added the EXACT() formula, which indicates whether the values in columns B and C are an exact match. In column F I've used a different formula which will return TRUE if "WITHDRAWN" appears in column E, and FALSE if it doesn't. You could then filter on either of those formula columns to see only the set of records that need corrections.

As you can see from the three examples here, spreadsheets are a really versatile tool. Most of your analysis work will probably use this technique. But it's also possible to come across some things you'll want to investigate where spreadsheets won't be sufficient.

32. Delve into Databases
Which brings me to the third and final technique, delving into databases. All library systems are built on a database, and some systems will give you the ability to directly access that underlying database. The Sierra system has the Sierra Database Navigator Application (or SierraDNA), for example. If that's not available, you can also export data from the system and construct your own database in order to be able to do the queries you need.

So when might you need to use this technique? I would treat this as a last resort, as it can be time intensive. But if the other tools and techniques you have are unable to perform the types of queries you need, this can be a viable solution. It's just not the most efficient way of answering simpler questions. You can do lookups in Excel, for example, but that's not as efficient as using a database, and often not even possible if you have very large amounts of data.

For us, this technique became necessary when checking that codes were consistent between sibling item and checkin records. Built-in ILS searches were unable to connect those two types of records

through the bib they shared, and matching up that data in Excel was too much for the program to handle.

33. Delve into Databases – Example [a]
To compare the location codes in our checkin and item records that were on the same bib, I first exported location code data and record numbers from the system, doing that separately for items and checkins. I then did some manipulation to remove duplicate codes and combine them all into a single field and used Microsoft Access to set them up in a database, which enabled me to match them up based on bib number.

Here's part of what the resulting report looked like. In the first three lines, the same location codes are used in both the checkin and item records, which is what we wanted. But the rest of the rows show various combinations of codes that don't match.

34. Delve into Databases – Example [b]
Here's one of those rows, where the checkin had a location code "clp," and the item record had a location code "clpe." During the migration, those mismatched codes…

35. Delve into Databases – Example [c]
… created a second holdings record that didn't have any associated item records.

36. Delve into Databases – Example [d]
And as a result, the way this record displays to patrons is misleading and confusing. Unfortunately we weren't able to get around to cleaning these up before the migration, but we at least have the analysis so that our periodicals folks know which records need to be looked at now.

These are just three general techniques for analyzing metadata, and there are of course other techniques I didn't talk about today. For instance, you can use regular expressions to check the format of values like barcodes or identifiers, or use APIs for a variety of things, both identifying and fixing problem metadata.

37. Metadata Analysis for Pre-Migration Cleanup
To close, I want to thank you all for coming today and wish you the best with your pre-migration cleanup.