Planning Batch Cataloging Projects

1. Planning Batch Cataloging Projects

Hello everyone, and thank you for coming to today's webinar on planning batch cataloging projects.

As we get started, I want to let you know that the slides and speaker notes for this webinar can be found at https://tiny.cc/PlanBatch.

2. About Me

I'm the Head of Metadata Analysis & Operations at JMU Libraries. My background in academic libraries includes traditional MARC cataloging for a variety of special formats, metadata for special and digital collections, and e-resources. I also develop workflows for efficiently managing and providing access to our resources through automation and batch processing, which brings us to the topic of today's webinar, planning batch cataloging projects.

3. Overview

To give you an idea of where we're headed today: We'll begin with figuring out where you're starting from with your project, the materials you have and the tools and other resources that are available. I'll then break down how plan a batch cataloging workflow, and also cover how to select the appropriate tools and technologies for completing your batch cataloging project.

4. Batch Cataloging [literature]

To get us all on the same page, I'm going to start by sharing some definitions from the literature. Young defines batch cataloging as "obtaining (or creating), transferring, manipulating, and editing groups of MARC bibliographic records." Another definition comes from Turner, who considers batch cataloging to be "editing and adding large batches of MARC records to a catalog at once," as opposed to "individually cataloging each title." Both of these authors mention working with groups of MARC records, but the only action where the two definitions overlap is editing.

5. Batch Cataloging [this presentation]

For the purposes of this presentation, I'm defining batch cataloging as any of the following actions performed on metadata in bulk: collecting metadata, searching for existing metadata, transforming metadata from one schema or format into another, matching up metadata from multiple sources, editing metadata, and loading it into a system. This definition is broader in scope than any I've found in the literature.

6. Good Candidates for Batch Cataloging

What types of materials are well suited for batch cataloging?

Because you'll be working with the items in a group instead of individually, it helps if they're similar. This often means they're all the same format, but there could also be other characteristics they have in common. One such characteristic is whether items are commercially published and relatively common, which means they're likely to have existing records.

Another thing to consider is the need. Does the collection meet the needs of researchers or support a particular aspect of the curriculum at your institution? It could be a collection that already sees high use but is not very discoverable. Also consider whether the items can be discovered in other ways, like through shelf browsing. Finally, good candidates for batch cataloging are collections that might not otherwise get cataloged if the items had to be cataloged individually. This might be due to the size of the collection, for example.

We did a batch cataloging project for a collection of jazz LPs which had limited metadata in the catalog and were stored in closed stacks and later in off-site storage. This was a collection that was highly used by our jazz program, and there was clearly a need to provide more detailed metadata in the catalog to facilitate discovery. Since all items were the same format, the collection was well suited for batch cataloging.

7. [Poll 1] Do you have materials that would benefit from batch cataloging?

I'm curious to hear if you have materials at your library that would benefit from batch cataloging.

[review poll results]

If you're thinking about a particular collection that you'd like to batch catalog, I'd encourage you to keep that in mind as we walk through the steps for planning a batch cataloging project.

8. Evaluate Resources

The first thing to do in planning your project is to take stock of the resources you have by considering who might be working on this project. This involves thinking about staff workloads as well as the skills that particular staff members have. Think about what tools they know how to use and their level of expertise with those tools. A batch cataloging project can be a great opportunity to further develop skills that you or your staff members may have, as well as to learn new skills by putting them to use in a project. If you want to use your project as an opportunity for professional development, consider the career goals of your staff and whether a project like this could support them in their goals. It's important to consider what resources are available for learning, which could be books, courses, websites, or other people. It also takes time to develop competency in new areas, so factor that in to your ideas about how long the project might take, and keep in mind what other tasks it might be replacing.

9. Two Key Questions

In addition to evaluating your resources, there are two key questions that will guide you in planning a workflow for your project.

The first is "Are records available for the items, or will they need to be created?" If there aren't any existing records, you'll be doing original cataloging and creating records from scratch. If records are already available, for example in OCLC, you'll be doing copy cataloging (going out and finding those records). So the first question directs us to one of two possibilities, original or copy cataloging.

The second key question is "What metadata is already recorded about the items?" I'm talking here about metadata that you have in your own systems. This could be brief records in your catalog, or a spreadsheet with an inventory. What metadata fields do you already have recorded?

10. Batch Cataloging Matrix

These two key questions give us a framework for categorizing the project based on the batch cataloging matrix. Across the top we have the first question, "Are records available, or will they need to be created?" which divides things into original or copy cataloging. The second question, "What metadata is already recorded?" is on the left, and that divides into having some metadata already recorded or not having any metadata. The answers to both of these questions will determine which of four categories your project will fall into. We'll go into more detail about each of these four categories a bit later…

11. Two Key Questions [expanded]

…but first I'm going to return to the two key questions. The first question determines whether you'll be doing original or copy cataloging. If records are available (if you'll be doing copy cataloging), you'll also want to think about what metadata you would need to have in order to search for those records. In the example I mentioned earlier with the jazz LPs, these were commercial recordings and the majority had records in OCLC, and we determined that searching by issue numbers was the most reliable way to find those records.

The second question involves taking stock of what metadata you already have for your items. If you already have some metadata, you'll want to also consider whether what you have is unique enough to search on (if you will be doing copy cataloging), and if it's not unique enough, what additional metadata would need to be collected to facilitate searching.

12. Additional Questions

In addition to the two questions that provide a framework for project planning, there are some additional questions that it can be helpful to think through. If you'll be working with copy cataloged records or metadata that has already been collected, consider what editing needs to be done and how you might accomplish that. I'll be sharing more about some tools for this later in the webinar.

At the end of the project, you will most likely be adding a set of records to your ILS or another system, so how will that be done? Will you need a special import profile or load profile? And if you're planning to overlay records that are already in your ILS, how will those records be matched with the incoming records?

Another thing to consider is whether this is a one-time project or ongoing process. You might make some different choices in planning out the workflow for something that's intended to be repeated multiple times.

Finally, when working in batch, it's good to build in checkpoints to make sure you're able to maintain the desired level of quality even though you're not looking at records one-by-one. This might mean verifying the presence of particular fields, or checking the format of data, or validating against a schema.

13. Batch Cataloging Matrix

Returning to the batch cataloging matrix, which, as we saw before, is organized by the two key questions, …

14. Batch Cataloging Matrix [categories numbered]

… that gives us four possible categories a batch cataloging project could fall into.

15. [Poll 2] Which category does your batch cataloging project fall into?

If you had a batch cataloging project in mind earlier, which category does it fall into?

[review poll results]

The reason these four categories are helpful is that they determine the stages that the project will move through. I'm going to walk through each category and outline the stages, and also give an example of a batch cataloging project I've done for each category.

16. Some Existing Metadata, Original Cataloging

Category 1 are projects where you have some existing metadata already recorded and will be doing original cataloging.

The first step is to transform the metadata you already have into your target format or schema. Then you edit the records, adding additional fields or modifying other fields. The final step is to load the records into your ILS. Transform, Edit, Load.

17. Example 1: ETDs

An example of this type of project is how we catalog our electronic theses and dissertations at JMU. Students submit their theses to our institutional repository along with metadata such as the title, author and contributor names, department name, and an abstract. We retrieve this student-submitted metadata in XML via OAI-PMH. The first step is to transform that metadata from Qualified Dublin Core into MARCXML; we do this with an XSLT script I wrote, which is available on GitHub. Then our cataloger converts the MARCXML records into MARC binary. So there are two transformations that happen in this example. Our cataloger then edits the records, correcting inconsistencies and adding subject headings and classification, before loading them into OCLC and our ILS.

We create both bibliographic and authority records for these materials, and the same process is used for both bib and authority records – transform the QDC metadata into MARC, edit as needed, and then load.

18. Some Existing Metadata, Copy Cataloging

Category 2 projects are ones where you have some existing metadata recorded and will be using it for copy cataloging.

The first stage for these types of projects is to use the metadata you already have to search for copy records. Then you will match those records up with your existing metadata, if that's necessary. You'll edit the records as needed and load them into your system. So the steps are Search, Match, Edit, and Load.

19. Example 2: Jazz LPs

This was the approach we took to catalog the collection of jazz LPs that I've already mentioned. We had some brief metadata already in our catalog, which included publisher names and issue numbers. We used those to batch search for OCLC records. We matched up the OCLC records to the bib numbers from our ILS by using the search queries stored in the OCLC save file database. I'm not going to go into detail about that process here, but I have published and presented on it elsewhere and the citations for those resources are available on the slide. After adding our bib numbers to the full OCLC records, we did some editing to clean up a few fields and then loaded the full records into our ILS, overlaying the brief records.

20. No Existing Metadata, Original Cataloging

Moving on to the bottom row of the matrix, Category 3 are the projects where you're not starting with any existing metadata and will be doing original cataloging.

Because we have no metadata to begin with, the first step is to collect metadata. Then, as with Category 1, you'll transform that metadata into records, edit them, and load them. So the full process is Collect, Transform, Edit, and Load.

21. Example 3: Comic Books

We've been doing this with a collection of comic books featuring Black characters and creators. The collect step involves staff and student assistants recording metadata in a spreadsheet. If you're thinking that this still sounds like cataloging items one by one, you're right! We still need to review each item to record the relevant metadata, but we're doing it in a way that will facilitate batch processing throughout the remaining steps of the process. Next, the spreadsheet is converted into MARC records, which are edited to add some boilerplate fields and to reformat some of the data. After that, the records are ready to be loaded into our ILS.

The transform and edit steps here are done with a single Python script, which is available on GitHub. When we talk about tools in a bit, you'll see that many of the tools used for editing can also be used in other stages of a project, which provides an opportunity to streamline things.

22. No Existing Metadata, Copy Cataloging

In the final category, Category 4, projects have no existing metadata to start with and will result in copy cataloged records.

You'll begin by collecting some metadata that will then be used to search for existing records. The next stage in the process is to match those records to the metadata you've collected, if necessary, and then edit and load. So the full process is Collect, Search, Match, Edit, and Load.

23. Example 4: CD Backlog

We cataloged a backlog of CDs in this way. Like with the jazz LPs, we had some brief metadata in our catalog, but in this case it wasn't unique enough for accurate searching. So we started by having a student collect UPC barcodes from the items by adding them to the existing brief records in the ILS. We then used those numbers in a WorldCat Search API lookup in OpenRefine to retrieve the OCLC number of the matching record. We merged our bib numbers into the full OCLC records, made a few edits, and then loaded the records into our ILS, overlaying the brief records.

24. Batch Cataloging Matrix with Project Stages

To recap, here's the batch cataloging matrix again with all the project stages listed out for each category. You can see that the "original cataloging" categories in the first column both require transforming, editing, and loading, while the "copy cataloging" categories in the second column use searching, matching, editing, and loading. The two "no metadata" categories on the bottom row both require collecting metadata as the first step.

Once you've identified which category your project falls into and know the stages that the workflow will follow, it's a good time to start thinking about how you're going to do each of those steps. Most of these functions are probably things you're familiar with doing for individual records, but sometimes it can

require a bit of a mindset shift to start thinking about how to do those same actions in bulk. I have a few questions to help you start to evaluate your own skills in these areas from that batch cataloging mindset.

25. [Poll 3] Do you have experience editing records in batch?
Do you have experience editing records in batch?

[review poll results]

26. [Poll 4] Do you have experience searching for records in batch?
Do you have experience searching for records in batch?

[review poll results]

27. [Poll 5] Do you have experience transforming metadata in batch?
Do you have experience transforming metadata from one schema or format into another in batch? This could be turning Dublin Core into MARC, converting a spreadsheet or CSV into MARC or XML, turning MARCXML into MARC binary, or many other possibilities.

[review poll results]

28. [Poll 6] Do you have experience matching or merging metadata records in batch?
Do you have experience matching or merging metadata records in batch? One example of this could be if you have a file of MARC records and a spreadsheet with a unique field that needs to be added to each record, merging the data from the spreadsheet into the MARC file.

[review poll results]

29. Project Stages
Again, here are all the stages that we may be going through in our project.

One important part of planning a batch cataloging project is figuring out which tools and technologies you're going to use. Some tools are better suited for certain stages than others, so when thinking about choosing tools I find it's helpful to talk about which tools are appropriate for a particular stage. So we'll go through each of the project stages and I'll give a few examples of tools that work well.

30. Tools for Collecting
First up is the stage of collecting. A pretty common tool here is spreadsheets. Whether you're using Microsoft Excel, Google Sheets, or another spreadsheet software, you can enter metadata directly into the spreadsheet. You can also create forms or surveys that provide a nice, user-friendly interface for collecting metadata. The data collected through tools like Google Forms, Qualtrics, Airtable, and other survey software is often then available in a spreadsheet. There's also the option of creating a custom app to collect data, which gives you more freedom in determining the format of the output. You could design an app that collects data through a form and then outputs the metadata directly in XML.

As I mentioned earlier, this is the least batch-like stage, because the metadata is often still being recorded by looking at items one by one. But the idea is to collect it in a more efficient way than directly creating a full record from scratch for each item. It's more efficient to enter a new row into a

spreadsheet for each item than it is to create a new record for each item. It then facilitates batch processes in the following stages, and also allows you to only collect the metadata that's unique to each item, saving any boilerplate fields to be applied in batch later.

31. Tools for Transforming

Next is transforming, and the tools that can be used will vary depending on the format or schema you're putting in and what the output will be. One example is MarcEdit's Delimited Text Translator, which converts a delimited text file or spreadsheet into MARC records. OpenRefine has an export templating feature that allows you to customize the transformation of tabular data into a variety of formats, including JSON, MARC, MODS, and XML. Scripting languages are another tool that can be used to transform metadata. XSLT is a good option when working with data in XML, and Python is another common programming language.

32. Tools for Searching

When it comes to searching, you can use the built-in batch searching functionality in OCLC Connexion if you're looking for OCLC records. You can use the Z39.50 protocol to search a number of library catalogs; MarcEdit is one tool that provides this functionality. There are also APIs for OCLC and other systems, and those can be accessed through URL lookups in OpenRefine as well as custom scripts. If you're working with a system or database where you have SQL access, you can also use SQL queries to search for records.

33. Tools for Matching

For matching up metadata from multiple sources, MarcEdit's Merge Records function can be used when both sets of data are already in MARC format or can be put into MARC. Matching up records can be done in Excel with lookups, or for larger datasets, in a database such as Microsoft Access. Matching is also something that ILS import profiles are designed to do, although the available options will vary depending on your particular system.

34. Tools for Editing

Editing a batch of records is a core function of MarcEdit, and one that I think is probably the most familiar to many people when thinking about batch cataloging. Another tool for editing is batch update functionality found in your ILS, like Sierra's Global Update or Alma's normalization rules. Scripting languages like XSLT and Python, which are great for transforming, can also be used for batch editing. If you're working in category 1 or 3 (the original cataloging categories), consider whether you can use the same tool for both the transforming and editing steps to streamline your workflow.

35. Tools for Loading

For loading records, the tools used will generally be pre-determined by the system you're loading the records into, whether that's an ILS, institutional repository, or a bibliographic utility like OCLC. If your goal is to add records to your ILS, consider whether you can use an existing import profile or whether you need to customize one for this particular project. If you plan to overlay existing records, you'll need to consider what field will be used as the match point and whether you need to protect any existing fields so they're not overlaid.

If you've got a one-time project, manually loading records is probably fine. You may also have the ability to automate record loads from an FTP server or to add records using an API, both of which can be useful for projects that require ongoing record loads.

36. Mapping Pathways between Tools

So I've just gone through a bunch of tool options for the different stages, and you might already have an idea of what tools you'd like to use in your project, based on the skills you know people have and the systems you have access to. But how do you connect all those pieces together into a complete workflow? To do this, you'll need to walk through the entire workflow and ensure that the output of one stage matches the input for the following stage. The available output formats are determined by the particular tool you're using for that step. I'll walk through two examples to illustrate.

37. Mapping Pathways between Tools – Example 1: ETDs [a]

The first is the ETD project we looked at earlier as an example of Category 1. In the first stage, Qualified Dublin Core metadata is transformed with an XSLT script into MARCXML. Ultimately, we want our MARC records to be in MARC binary…

38. Mapping Pathways between Tools – Example 1: ETDs [b]

… which is why there's a second transformation to turn the MARCXML from the first step into MARC binary. The output of the first transformation is the same as the input for the second.

39. Mapping Pathways between Tools – Example 1: ETDs [c]

When we move on to editing, we again take the output of the previous step, MARC binary, and use that as the input for this one. In this case, the input and output formats of the editing stage are the same, but the records themselves are modified to add and edit fields, so the records that we end up with as output are more complete than what we started with in the editing stage.

40. Mapping Pathways between Tools – Example 1: ETDs [d]

The final step is to load the records into OCLC, which we do by importing them into Connexion.

When looking at all of the stages together, you can see that once the records are in MARC binary then the format doesn't change for the remainder of the steps. That's something to keep in mind when designing your workflow – it can streamline things if you're able to do multiple steps with the data in the same format or using the same tool.

41. Mapping Pathways between Tools – Example 4: CD Backlog [a]

As another example, I'll show the workflow for our Category 4 project, cataloging a backlog of CDs. We started with some brief records, added UPC barcodes directly in the ILS, and then exported that data as a spreadsheet.

42. Mapping Pathways between Tools – Example 4: CD Backlog [b]

Then we loaded that spreadsheet into OpenRefine, and used a URL lookup to search via OCLC's WorldCat Search API. The result of that process was the same spreadsheet with a column of OCLC numbers added.

43. Mapping Pathways between Tools – Example 4: CD Backlog [c]

After exporting the spreadsheet from OpenRefine, we made a text file from the OCLC number column. That list of OCLC numbers was the input for a batch search in Connexion. It's possible that you would only need to use part of the output from one step as the input of the next, as was the case here. The results of the batch search were exported in MARC format.

44. Mapping Pathways between Tools – Example 4: CD Backlog [d]

The next step is matching. For any match, you're going to have two inputs, the two sets of data that you are matching up. We needed to match up the full MARC records just exported from OCLC with some additional data that was in the spreadsheet initially exported from our ILS. We turned the spreadsheet into brief MARC records using MarcEdit and were then able to merge those together with the full records to end up with OCLC records that had our local bib numbers added. After that, we edited those records in MarcEdit (as you saw in the previous example) and loaded them into our ILS.

45. Testing

The process of connecting all the stages of the workflow will probably involve some testing, especially if you're experimenting with new tools or features you haven't used before. Once you have a pathway mapped out, you'll also want to test it completely. I recommend walking through the entire process once or twice with a single record, and then with a small group of records, to make sure that everything works the way you expect. If you have access to a sandbox or testing environment for any of the systems you're using, that's a good option to take advantage of.

Testing with a small number of records to begin with can help you find errors in the data more easily, because you're able to examine the records individually. Use any errors you find as an opportunity to incorporate some quality control checks at appropriate points in the workflow so you can catch those types of errors even when working with larger batches.

As you're doing testing, this is also a good time to make sure that you have the process documented.

46. Documentation and Training

Write down all the steps in detail, even the ones you think you'll remember later. This is especially important when you are the one designing the workflow but someone else will be performing the work. As you train that person on the process, you'll find areas where you need to be more specific in your instructions and can continue to refine the documentation. With our jazz LP project, I did the first third of the project myself before it had to be set aside for a few years, and then I later had to train a colleague who finished the project, so I was very thankful I had kept detailed notes for myself.

Documentation is also great if you want to share about your project later. If you've developed a great workflow, other organizations could benefit from your work and would love to hear about it.

47. Questions – Planning Batch Cataloging Projects

That brings me to the end of what I had planned to share with you all today. I hope this webinar has given you some ideas and strategies for tackling your own batch cataloging projects. We have some time now for Q&A, so if you have questions please put them in the Question pane. If you think of additional questions after today, you can contact me through my website, https://www.rebeccabfrench.com, or by email.