Batch Cataloging at JMU: A Framework and Four Projects

1. Batch Cataloging at JMU

Hello everyone! It's great to be here with you all today, and I'm thankful for the opportunity to share with you some of the batch cataloging work we've done at JMU.

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

My name is Rebecca French, my pronouns are she/her, and 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, including music, 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 this presentation, batch cataloging projects.

2. Overview

I'm going to briefly talk about what I mean when I use the phrase "batch cataloging." Then I'll share a framework I've developed for categorizing potential projects into one of four types, which also provides guidance on how to design a batch cataloging workflow. I'll also talk about four examples of projects we've done or ongoing batch workflows we use at JMU.

3. Batch Cataloging [literature]

When looking at the LIS literature, there are a couple definitions of batch cataloging. An article by Philip Young defines batch cataloging as "obtaining (or creating), transferring, manipulating, and editing groups of MARC bibliographic records." Another definition comes from Ariel 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 librarians are specifically talking about working with MARC records, and they cover a few different actions you could be performing in batch.

4. 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 records into a system. This is intentionally format-agnostic – it's not limited to MARC records – and it covers a broader range of activities than the other definitions.

So now we want to take this list of possible actions and use it to design a workflow for a particular project.

5. Two Key Questions

To do this, we'll start with two key questions that will lead to establishing a framework for batch cataloging.

The first question 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?

6. Batch Cataloging Matrix

Based on our answers to these two questions, we’re now able to categorize a batch cataloging project based on this framework. 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 determine which of four categories a project will fall into. We’ll go into more detail about each of these four categories in just a bit …

7. 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 one of the examples I’ll share later, we had a collection of LPs that were all commercial recordings, and the majority had records in OCLC. We determined that searching by issue numbers was the most reliable way to find those records. For other types of materials, you might use the ISBN or some combination of other fields like title, author, publisher, and date.

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.

8. Additional Considerations

In addition to the two questions that provide a framework for project planning, there are some additional things 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.

At the end of the project, you will most likely be adding a set of records to your ILS or IR 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. The four projects I'm going to talk about later in the presentation include examples of both one-time projects and ongoing workflows.

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.

9. Batch Cataloging Matrix
Returning to the batch cataloging matrix, which, as we saw before, is organized by the two key questions, …

10. Batch Cataloging Matrix [categories numbered]
… that gives us four possible categories a batch cataloging project could fall into. I'm going to go through each of these categories and describe the steps that are involved in the workflow, and I'll also give an example of a project that we've done at JMU for each category to illustrate each of the steps.

11. 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.

12. Example 1: ETDs
Our workflow for cataloging our electronic theses and dissertations at JMU is an example of this category. Students submit their theses to our Bepress institutional repository along with metadata such as the title, author and contributor names, the academic 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 XSLT scripts I wrote, which are 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 have created both bibliographic and authority records for these materials, using the same process for both types of records – transform the QDC metadata into MARC, edit as needed, and then load.

13. 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 this type of project 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.

14. Example 2: Jazz LPs
This was the approach we took to catalog a collection of jazz LPs that had only brief records in our ILS. These items were stored in closed stacks and later moved to off-site storage, and because the collection was highly used by our jazz program, there was a clear need to provide more detailed metadata in the catalog to facilitate discovery since shelf browsing wasn't an option. The limited metadata already in our catalog included publisher names and issue numbers, which we used 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 and links to those resources are in the slides. 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.

15. 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.

16. 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. Sometimes it's possible to streamline the process by using a single tool for multiple stages of a workflow.

17. 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 step 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.

18. 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.

19. 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 involve 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.

Identifying which category a project falls into and the stages needed in the workflow is just a starting point for project planning. Next you’d start thinking about how you’re going to do each of those steps (searching, editing, etc.). I don’t have time today to get into the various tools that can be used, but if that’s something you’re interested in …

20. Resources

… I’ll point you to a webinar where I went into that in more detail; it’s the first resource listed on this slide. I’ve also included information here on the other projects I mentioned, as well as my email.

That brings me to the end of what I had planned to share with you all today. I hope this presentation has given you some ideas and strategies for tackling your own batch cataloging projects. We have some time now for Q&A, and I’m happy to answer any questions.