How Not to Waste Catalogers' Time: Making the most of subject headings

John Mark Ockerbloom
Good morning. I’m thrilled to finally be at Code4lib. I’ve been seeing so many people in this room do great work these last 10 years, and what I’ve seen here the past couple of days at this conference has impressed me to no end.

Right now, I’d like to focus on the work that catalogers do, and how this community can make the best use of that work. I’m not sure what sort of representation catalogers have in Code4lib, though I know there are some of you here. How many of you consider yourself catalogers, in some capacity? (Raise hand.) [Good, glad to see you; thank you for being here.]

A lot of the work I’ve been hearing about here fundamentally depends on data. And a lot of the data has been created by catalogers. But when I look at the work I’ve seen catalogers produce, and the OPACs and discovery systems I’ve seen, I worry that many coders are overlooking important aspects of how catalogers describe resources, particularly when it comes to subjects. And I worry that this means we’re wasting their time....
..and in the process, we’re also wasting the time of our libraries’ users, who still depend on subject descriptions for resource discovery. We heard yesterday, for instance, about how archivists were finding it increasingly important to include descriptions of the subjects of archives into EADs to direct researchers to archival materials they might overlook, particularly when those materials aren’t online. And even when full text is searchable online, simple term frequency searching doesn’t turn up the most relevant texts in many cases. It certainly doesn’t when we’re just searching metadata, as we’ll see in a bit.

It’s also important to recognize that ordinary users also do a lot of subject cataloging informally, through mechanisms like tags, folders, and reading lists. Ideally we’d like that informal cataloging and the more formal cataloging done in libraries to reinforce each other. Let’s talk first about some of what goes on in formal cataloging.
What catalogers do with subjects

• Standardize the representation of subjects
  – Eases reuse across systems, analysis
• Note relative importance of subjects in works
  – Eases presenting relevant works for user interests
• Define rich set of subject relationships & alternative expressions
  – Eases finding and clustering appropriate works, even when users have different vocabularies

• For instance, in formal cataloging, we have STANDARD IDENTIFIERS FOR CONCEPTS! And Christina Harlow and Missy Elliott told us yesterday why standardizing conceptual identifiers is awesome. And challenging. But if it’s done right, we can bring together a wide variety of works across our library that deal with common concepts. Not only that, we can use those same identifiers to explore across many libraries, and even outside of libraries, as we’ll see later.

• But catalogers do more than just classify and standardize. They also note the relative importance of subjects in works. In a traditional catalog record, the first subject assigned is the most important: it’s the one that determines where a book will go on the shelf. The remaining subjects are usually listed in descending order of importance.

• And the subjects aren’t just an unconnected set of concepts. Catalogers have spent a fair bit of time specifying connections between related subjects, and also paid some attention to different ways you might refer to a subject. This is important because no two people classify the world in the exact same way, and it’s easier to navigate other people’s organization of the world when you can get a sense of how it all fits together, and what it might relate to in your own understanding of the world.
Unfortunately, too many library catalog and discovery systems don’t do much to help users take advantage of the rich subject organization I’ve just described. If you want to browse subjects in the Library of Congress’s catalog, for instance, your options are pretty limited. You can do a linear call number browse. Or you can do an alphabetic subject browse. So if you’ve visited the historic sites here and gotten interested in the American Revolution, and decide to browse that subject, you’ll be offered 39 books that were cataloged with FAST. But you’ll have to click on other links to find the more than 2000 books on the Revolution cataloged with LCSH.
And if you click through to the LCSH version of the subject, those 2000 books are simply presented in alphabetical order by title. You have a few options to change the sort, but it’s awfully hard to find the most relevant books on the subject with what you’ve got here.
What basic Solr does with subjects

Well, okay, you might say, but the catalogs we build nowadays are much more about search than about browse. We’ve got Solr! And faceting! And yes, those are good things, but you can’t just throw your subject headings into a term-weighted index and call it a day. We can do better than that. If I put the LCSH subject heading for the American Revolution into our Solr-based catalog at Penn, the first thing that comes up is a book from 1954. It is a survey of the revolution, but it’s far from the most important book on the Revolution for readers today. It’s not even the most important book on the Revolution by this author. But it does have a lot of the same terms in my subject heading, “Revolution”, “1775” and “1783” so it gets a great score from our basic Solr algorithm.

And yes, I can use faceting to narrow down the 11,000 hits I get in this catalog into something more manageable. But while faceting lets me slice and dice, it’s not so good at letting me explore related subjects that might use different terminology, like “The Boston Tea Party” or “The Battle of Lexington”. I can narrow and broaden the set of works I’m seeing, but it’s trickier for me to explore laterally.

But we can make our catalogs smarter, when we pay more attention to what catalogers do and how users might want to explore.
Here’s another way we can view and explore works on a particular subject. This is a catalog I’ve built of public domain and other freely readable texts available on the Internet. It organizes works based on an awareness of subjects and how subjects are cataloged. The works we see at the top of the list on the right, for instance, tend to be works where “United States -- History -- Revolution, 1775-1783” was the first subject assigned. Books where that subject was further down their subject list tend to appear further down in this list. I worry about whether I’ll still be able to do this when catalogs migrate to RDF. [You just heard in the last talk] that in RDF, unlike in MARC, you have to go out of your way to preserve property ordering. So here’s my plea to you who are developing RDF catalogs: PLEASE GO OUT OF YOUR WAY AND PRESERVE SUBJECT ORDERING!

My catalog also uses other clues in the metadata to try to bring more useful results to the surface. All else being equal, for instance, recent books will appear higher in the list than older books, since they tend to reflect more up-to-date knowledge and scholarship. [But the importance of recency can vary by subject. You wouldn’t want to learn Web development from a 20-year-old book, but a book about math or natural history could still be highly relevant decades after it was published.]

Now here, this collection is mostly public domain books, and we don’t have recent books to show at the top for this subject. Instead, we’re seeing contemporary books, books that were being published as the Revolution was underway, or shortly thereafter. We know that works published close to an event can be of special importance in primary source research. And because catalogers have conveniently placed the time period of the event right there in the subject heading, it’s easy for us to extract that information and use it to boost the relevance of works close in time to the event.

On the left side of the screen, there are various links you can use to explore a cluster of concepts around this subject. Some of them involve the same kinds of subdivision addition and subtraction you get with faceted search. But you can also find related concepts that are expressed in different language. From here, you’re only a couple of clicks away from
...the Battle of Germantown, the Revolutionary battle fought about 10 miles north of here, near my house. I don’t list a lot of books that are just about that battle, but thanks to the network of relationships that are in LCSH, or that I can derive from LCSH, I can cluster those books with books about other things that happened in the area during the Revolution. Some of those books might have more about Germantown; or they might have other things I could be interested in if I’m interested in Germantown.

But I don’t have to stop just with what’s in this collection. I can click on that link marked “your library” up there...
...and find a bunch of other resources I can use if I head back to Penn, or log into our proxy.

Or I can check out another library’s collection. The Historical Society of Pennsylvania is just a few blocks from here...
...so if I want to see what they have, I click on another link to choose a different library...
...and I find all kinds of items I can look at over there, including some unique ones, like the papers of the family whose house was right in the center of the battle!

But that’s not all. We can use the subject heading that catalogers have assigned along with some crosswalk data to find resources *outside* of libraries.
For instance, you can click straight from my catalog’s subject view of the Battle of Germantown to Wikipedia’s page on the Battle of Germantown. Here you get a picture of the Chew family house I was just mentioning. You also get maps of the battle, and cites and links to a few dozen online and offline resources with more information.
...including links back to those hundreds of library collections you got a glimpse of earlier.

And this is where we can start linking the LCSH-based cataloging that professionals are doing in libraries with the cataloging that the wider public is doing out on the Web.
One very simple way of connecting, for instance, is taking the terms that non-librarians use and linking them to terms in library subject vocabularies. LCSH has no heading labeled “Battle of Germantown”, either as a preferred or as an alternate form, but we can add the terms that get used in Wikipedia and in FAST to the Library of Congress headings, so that users are more likely to find subjects expressed in their own language.

The community can help us out in other ways as well. The Open Syllabus Project has collected over a million lists of readings assigned by faculty for classes on various subjects. I’d love to be able to mine that data so if someone’s browsing a historical subject, the relevance of resources that are being actively used in history classrooms could show up high in the results list.

And the broader community also does a lot of organization and cataloging themselves. In some ways they go farther than what library catalogers have been able to do on their own.
Our keynote speaker yesterday talked about the Black Lives Matter movement. That’s been going on for 3 years now, and there still is no Library of Congress Heading for it. But there’s been a Wikipedia article for a long time. I wouldn’t say it’s the *best* reference source on the movement, but it does have links to over 150 articles in the media, as well as a prominent link to the activist-run website, where you can see what people in the movement are saying. And there are many other ways in which people who care about the issues raised by Black Lives Matter have organized information about it, often on the periphery of librarianship. I’m thinking of things like the Ferguson Syllabus, or the archive of hashtag tweets that Ed Summers has made to document the information sharing that’s happened over social media.

I’ve taken some stabs at trying to integrate this sort of community cataloging with the more traditional subject clusters you saw earlier. I don’t really have time to cover the details in a 15-minute talk, and personally I’ve had to step back from it to avoid the sort of overworked burnout that I think Becky Yoose is going to be talking about next. But if this sort of linkage interests you, I’d love to talk with you after this session and share ideas.
Because really, sharing is what makes possible everything I’ve talked about here. I couldn’t have built the catalog I showed you without a lot of bibliographic data that catalogers have shared through projects like HathiTrust, or authority data that LC has shared, crosswalking data that OCLC has shared, or article content and data shared from Wikipedia. I’ve added a bit of my own data to launch those searches into other library systems, and to link to topical subjects in Wikipedia. And I’m sharing that data on Github, and am happy to collaborate on it.

(For instance, we can add more library systems to that list of searchable library collections if you contact me or submit a pull request.)
Basically, I’m standing on the shoulders of a lot of giants. And a lot of those giants are catalogers.

I hope I’ve given some sense of some of the things catalogers do, whether they’re professionals or amateurs, and how coders can use what they do to better support the information needs and practices of our users. If this sounds appealing to you, you should check out mashcat, which is both a hashtag and a series of online and face-to-face meetings for discussing how catalogers and coders can work together better. And if you’ve found it useful to listen to me these last 15 minutes, my slide here has places you can find me later on.

Thank you.