The Metadata Challenge: Promoting Discovery, Access, and Usability for Online Books

John Mark Ockerbloom, University of Pennsylvania

Available at: https://works.bepress.com/john_mark_ockerbloom/13/
I’d like to thank Gabriella, Peter, and the other organizers for arranging this workshop and inviting me to speak. It’s a exciting time to talk about books online. We really live in a world of riches, even compared to just a few years ago. When I first got interested in the subject, in the early 1990s, there were a few hundred books known to be freely readable on the Internet, and the director of Project Gutenberg had attracted a number of volunteers and doubters when he said that there could be as many as ten thousand books online by the year 2000.

Here we are in 2010, and earlier this month, James Crawford, who you'll be hearing from later today, announced that Google had scanned over 15 million volumes. That’s about the size of the entire library system of Harvard University, the largest private library system in the world. And Google is far from the only book digitizer. Here in this city, the University of Toronto has digitized over a quarter million volumes for free online access, and we’ve also had Microsoft, the Internet Archive, the Europeana initiative, and a host of other large and small projects digitizing their books. Together, they provide millions more volumes, in all kinds of languages, genres, and areas of knowledge.

At the same time as we’ve seen an explosion in the supply of digital books, we’ve also seen a great increase in the demand. Nowadays, you don’t need to be sitting at a computer desk to read a book online. People are reading them on their Kindles, their Nooks, their IPads, and their mobile phones. They’re listening to them on their IPods and MP3 players. They’re taking them in on their fingers with refreshable Braille displays. The market for ebooks has also been growing rapidly; this summer Amazon announced that they were selling more Kindle books than hardcover books.
.And yet, while we now may have as many books online as a world-class research institution, we still have a ways to go before we really have a library online that provides all the services that we expect for world-class research.

Which is not to say that the online book digitizers haven't made some important new library services possible. The ability to search in fractions of a second the full text not just of a single book, but of a massive collection of books, is a huge step forward, and a service that I think will be revolutionizing scholarship for years to come. And it's also a big step forward to be able to access a book instantly, from wherever we are, instead of having to retrieve it from some inconveniently located bookshelf, or wait days for a book order to come in from Amazon or inter-library loan.
But at the same time, a lot of those millions of volumes are ones that I can’t actually read online. I may be able to search for words in them, but the full passage that contains the words I find may be blocked, due to concerns about copyright. Even if there’s a copyright holder somewhere who would be happy to make the book available for a little compensation, or just to get her work more widely read, or if the work is in fact in the public domain where I am, I might not have access to the books.

Then there’s the question of finding the book I need. Will I find it using full-text search, if it doesn’t use the same terminology, or the same language, as used in the book? Maybe, maybe not. Sometimes important work in a particular subject area uses different terminology than we use today. And if you’re using Google, it’s often a lot easier to find 10 of the most important web pages on a particular topic than 10 of the most important books on that topic.

And then, even if I find a volume from a work I need, will I be able to find all of the work? I’ve heard from more than one university professor who’s been frustrated at the scattered nature of multi-volume works in many online collections. When you’re in a field like Victorian literature, where your first editions often tend to be 3 or 4 volume novels, being able to find only 1 or 2 of those volumes just won’t cut it. We need to have libraries that are more than piles of books.

[notes continue on next page]
The knowledge that is encompassed by a library goes well beyond the books that are in it. It’s also the people who are part of it. It’s the librarians that have built up the collections, that catalog them, and that can help guide readers to the books they need, and teach them how to use them effectively. It’s the readers who use the books in the collections, who teach with them, who recommend books, and whose requests for resources and assistance help shape a library’s collections and services. And let’s not forget that, especially in academic libraries, lots of the books themselves are written by the people who use the library, or others like them.
So an effective library reflects the collective intelligence both of the authors in its collections, and also of the people who use and maintain it. Much of that intelligence is distilled and preserved in the form of metadata. And that metadata can do a lot to help libraries serve the needs of their users. In 1931 the great Indian library scientist S. R. Ranganathan expressed these principles for library user service.

These principles are still core tenets for library service. There are some obvious updates: We manage a lot of different kinds of knowledge resources other than books, and we don’t assume that the reader will necessarily be called “he”. But they get across a lot of the things we want to accomplish with our metadata and other library services.

We want to make sure that users discover all the books that will help them learn what they need. We want to make sure the books can be easily accessed. We want to save the reader’s time. And we want to make sure the books can be effectively used, even as the size and variety of information available to readers grows at tremendous rates.
Our metadata challenge, then, is to effectively distill and record the collective intelligence of the people who work with books, and other knowledge resources, and then use this data in ways that help people get the most from these books. I’ll be spending some time in this talk discussing a few of the more interesting ways to do this to support discovery, access, and use.

One thing that’s clear, especially with the millions of books online, is that we have to be able to do this at scale. I started working with book metadata in the early 1990s when I created a web site that helped people find free online books, no matter who had digitized them. Over the past 17 years or so, working part-time, I created nice, high-quality metadata—titles, creators, subjects, publication information, and the like—and by the start of this summer, I’d assembled a nice listing of about 40,000 titles. Which isn’t a bad size for a library in a high school or a small town, but it’s only a tiny fraction of the books now freely available. Earlier this month, though, I incorporated data from a few other sources that instantly grew the collection twenty-fold, and my site now lists more than a million volumes. That’s still not everything freely available, but it’s a lot more of it, and I’ve got plans for further expansions. I’ll tell you more about how I scaled up in this talk.
I didn’t do it just by saying “Okay, let’s round up all the metadata we can”. The fact is, there’s lots of metadata out there, as I’m sure folks at Google can attest. Not only is there lots of metadata, there are lots of metadata standards. This is a recent poster produced by the Indiana University Libraries that shows more than 100 relevant standards for metadata. It’s well worth browsing in detail; you can find metadata standards there for all kinds of needs and uses.

You don’t need all the metadata that’s out there to build a good library. Some of the metadata isn’t all that great anyway. But there are ways you can use both well-managed and not-so-well-managed metadata to help your users, if you know how to work with it intelligently.
Let’s consider discovery, for instance.

It’s not that difficult to find a book online if you already have a specific one in mind, especially if you already know the title and the author from a citation. In most cases, you can just type that information into Google, or Amazon, or your local library’s online catalog, and find the work you’re looking for if it’s available.
If you're of a certain age, you may also remember looking up books in an actual card catalog that was well-set up for this sort of “known item” lookup. And what’s on these cards was metadata, served up in a massively parallel, human-powered, zero-energy interface.

Now, there are other times when you don’t have a particular book in mind in advance. Instead, you have a question you want to answer, or a concept you have in mind that you want to find out more about. Now, you could use card catalogs like this one to find books by subject, but it often wasn't particularly easy to do.
As we’ve moved online, we’ve found other ways to identify and index knowledge resources by subject. At Penn, for instance, we developed a social tagging system where some of our scholars bookmark things they find online that they’re interested in. And then they can go back and find them later, and so can anyone else. Here’s a view of some recent things that people have tagged with the keyword “discovery”.

This is all metadata too, and it’s not just coming from librarians or publishers (though some of the details are). It’s coming from our users, and we don’t even have to pay them to give it to us. They tag things because they’re useful, and because they like to be able to remind themselves about them, and tell others as well.
We can analyze these individual bits of metadata, to provide some interesting overviews. This is a recent high-level view of the “folksonomy” of PennTags: the most common keywords that PennTags community members have assigned to resources. If I’m interested in any of the concepts expressed by these keywords, I’ll find a lot of stuff of interest to me.

And this sort of social metadata creation and sharing is happening all over the Net. It happens in Zotero, and Mendeley, and LibraryThing, and lots of other communities. Right now, my favorite current awareness service is Twitter. I follow about 100 colleagues in information technology and library science, and they tell me in their tweets about things I might want to read. And I return the favor when I find an interesting article or news story online. I can even follow the conferences my colleagues are attending, when they include keywords in their tweets known as hashtags that identify where they are.

But there’s only so far you can go with home-grown folksonomies. A tag-cloud like the one you see here can only include a couple of hundred items before it starts getting unwieldly. And while it’s not hard for me to find and follow the conferences my colleagues are at now, it’s not so easy for me to go back and find where that great paper was on social book annotation.

[notes continue on next page]
When you’re dealing with more than a small set of familiar subjects, it’s useful to have a controlled set of concepts to explore. Lots of libraries use a set known as the Library of Congress Subject Headings, which defines millions of subjects of study that are linked to knowledge resources. Having a large ontology of concepts that have already been written about by somebody or other can make it a lot easier to learn more about a subject area, particularly if it’s an area that you’re not very familiar with.

But you may be thinking: this sounds like a step backwards. Didn’t people already try to catalog the Internet by subject, with things like the Yahoo subject menus back in the 1990s, and the Open Internet Directory? And didn’t we pretty much abandon those, in favor of the wide open spaces of keyword searching through Google or Bing?

Well, in fact, concept catalogs are still used a lot online, but they look a little different now. The most popular one should be familiar to you--
…it’s Wikipedia.

Now, you might not think of Wikipedia as a catalog, but that’s how a lot of people use it. If you do a keyword search for Google on a concept, and there’s a Wikipedia page on it, chances are that page will appear high in the search results. And this happens pretty often, because Wikipedia already has more than 3 million articles in English alone, and lots more in other languages.

And lots of people who wouldn’t necessarily trust or cite a Wikipedia article in their research— and there are lots of good reasons not to— still say they use Wikipedia a lot as a “starting point” for information seeking. What do they mean when they say it’s a starting point? Well, assuming they’re telling the truth and not also using Wikipedia as an ending point, they find an article on a particular subject, get a bit of orientation on it in the article text, maybe follow an internal link to an even more relevant subject, and then they typically find a bunch of external links pointing to more detailed, and often more authoritative resources on the subject.

Which is basically what a library catalog does: you start with a concept in your head of something you want to know, and the catalog gets you from that concept to some good stuff to read about that concept.
Which is why I often wince when I think of how badly our traditional online library catalogs do at this. I was recently watching a DVD of Doctor Who, the British science fiction series that got revived about 5 years ago, and in the first episode I saw, one of the characters says to the Doctor, "You talk like you’re from the North", and I had no idea what she meant by that, or how she could tell. Well, it turns out the Penn library has lots of books about dialects and accents in England that could answer my question, but good luck finding the best ones in this catalog. Subject searches in our catalog are sort of like digging through cards in a drawer, only worse. Even if I know where to start looking "English language – Dialects – England -- (place name)" I’m stuck with clicking through lots of screens that show me a bunch of headings in alphabetical order, and no books. To see what books are here, I have to click on the headings one at a time, going back and forth. Maybe I should look in "Lake District" – that is, after all, in Northern England. Or in some other heading that I will only see after several more screenfuls. It’s no wonder that our logs tell us that hardly anyone browses by subject in our catalog.

At least in the old card drawers, I could riffle through the cards quickly to see what books were there, and what the major headings were.
We have other ways of organizing information in libraries, but they often don’t work much better by themselves. There’s hierarchy, for instance. Historically we’ve used that a lot when we’ve had to arrange books on a shelf, or documents in folders. When we’re online, though, we don’t have to settle for putting everything in exactly one place. We know from our experience with Yahoo directories (or, for some of us, from Gopher that people want to spend their time finding stuff and reading it, not learning to navigating through menu trees. We can do better online.
One kind of catalog that's gotten a fair bit of attention lately is the faceted catalog. Computers and relational databases are great at slicing and dicing, and that's what facets are all about: choose something from column A, something else from column B, and pretty soon you've narrowed down your range of search to something tractable.
A growing number of library catalogs, as well as sites like Google Books, let you browse using facets, and they often work quite well. You can see facets here down the left side: they’re great for narrowing down searches to specific types of content, time periods, availability, or language. They don’t work so well for subjects, though, because subjects don’t slice and dice so easily, particularly when you have a lot of them. They exist in a more organic network of relationships.

If you’re trying to find out about at say, “computer network security” there’s no really easy way to use facets to show books in other subjects that might have something to say on the matter, like “network protocols”, “public key cryptography” and “cyberwarfare”. These are all related topics that might also be of interest. In catalogs that do use subject terms as limiting facets, it’s very easy to get caught in cul-de-sacs as you’re searching.
We have other models for exploring rich information networks, though, and one very powerful model is the map. You can see how it works in this old map of Sydney. You see in a glance not only the different areas of the city and how they relate to each other, but also a lot of specific places you might want to visit, and how to get to them. Online, if you see a map like this with an area that looks interesting, you can zoom in on it to get more detail, or pan the map one way or another to see nearby areas. It's very informative, and very intuitive.
You can use similar kinds of maps for exploring subjects. They don't actually have to look like maps you'd find in an atlas, they just have to display information in a similar way. At this point I’m going to try leaving this slide show and showing you how they look on The Online Books Page…

[show subject descriptions, relationships, examples (Information Oragnization). Show related subjects and books under them (1812). Explain how most of this metadata comes from elsewhere. Explain how you enhance it, improve subject metadata in Hathi data; how there’s interest in sharing improvements with others]
United States -- History -- War of 1812 -- Campaigns

Broader term:
- United States -- History -- War of 1812

Narrower terms:
- Baltimore, Battle of, Baltimore, Md., 1814
- Beaver Dams, Battle of, Beaver Dams, Ont., 1813
- Chateauguay, Battle of, Quebec, 1813
- Delaware, Battle of, 1813
- Fort Erie (Ont.), History -- Siege, 1814
- Lake Erie, Battle of, 1813
- Lambs Lane, Battle of, Ont., 1814
- New Orleans, Battle of, New Orleans, La., 1815
- North Point, Battle of, Md., 1814

Filed under: United States -- History -- War of 1812 -- Campaigns

[3] American naval battles: being a complete history of the battles fought by the navy of the United States from its establishment in 1794 to the present time, including the wars with France, and with Tripoli, the late war with Great Britain and with Algiers, with an account of the attack on Baltimore, and of the Battle of New Orleans. With twenty-one ... engravings. (Concord, NH: Luther Reby, 1848) (page images at Hathi Trust)
[4] Official history of the military and naval operations of the United States during the war with Great Britain in the years 1812, 13, 14, & 15: with some additional letters and documents elucidating the history of that period / collected and arranged by John Bonner. (Washington City: Printed by Way & Gideon, for the editor, 1823)
Demo (snapshot 4)
The Online Books Page

Browsing subject area: Toronto (Ont.) (Exclude extended shelves)
You can also browse an alphabetical list from this subject or from:

Toronto (Ont.)

Broader term:
- Ontario

Narrower terms:
- Toronto (Ont.) -- Description and travel
- Toronto (Ont.) -- Fiction
- Toronto (Ont.) -- History
- Toronto (Ont.) -- Poor
- Canadian Aeronautics Limited (Toronto, Ont.)
- Trinity College (Toronto, Ont.)
- Books and banking -- Ontario -- Toronto
- Charities -- Ontario -- Toronto
- Commerce -- Ontario -- Toronto
- Education -- Ontario -- Toronto
- Natural history -- Ontario -- Toronto
- Streets -- Ontario -- Toronto

Filed under: Toronto (Ont.)

[2] Toronto "called back" from 1862 to 1867. (Toronto : W. Briggs, 1912), by Coningham Crawford Taylor (page images at Hathi Trust; US access only)

Filed under: Toronto (Ont.) -- Description and travel

[2] The Queen's jubilee and Toronto "called back" from 1887 to 1887. In wonderful growth and progress... and reminiscences extending over the four deсяtennial periods, from 1847 to 1887... This revised ed. contains the progress of the city from 1858 to 1857..., with a full account of the celebration of the Queen's jubilee in London, Toronto and other places throughout the world... By Commissioner President Taylor. Toronto : W. Briggs, 1887, by

John Mark Ockertboom  © UNIVERSITY OF PENNSYLVANIA LIBRARIES  Oct, 28, 2010

Page 22
Building effective subject maps

• Start with the best metadata, ontologies you can
  – Subject assignments in existing metadata records
  – Subject concepts, relationships from LCSH authorities
  – Bibliographic and authority data can be downloaded in XML!

• Automate analyses for more connections, relevance
  – Subdivision, geographic, lexical, co-location analysis
  – Analysis can also auto-correct, localize subject headings

• Specialize and refine where it gives the greatest benefit
  – Logs can tell you what people are looking for, finding
  – You know what your special collections and communities are

• Design your displays and navigation well
  – give users birds-eye view, and teach them the terrain

• Share your improvements
Some other ways to build connections between books

• Citation analysis
  – What does this work cite? What cites it? How important are the citers?
  – THE main distinguishing feature of Google over earlier search engines
  – Has a long history in bibliometrics (Web of Science; now Google Scholar)

• Quotation analysis
  – What does this work quote from? What quotes it, or has common text?
  – (now used in Google Book Search)

• Clustering
  – What other resources share features (metadata or content) with this one?
  – (e.g. latent semantic indexing)
I could spend all my time just talking about discovery, but there are some other important metadata challenges. One big one is access. Remember those 15 million volumes that Google’s scanned? I can’t actually read most of them right now. And I’d like to be able to read them.
I recall hearing from Google’s Librarian Central website that of all the books they’re digitizing, maybe about 20 percent of the volumes are ones they consider to be in the public domain. So if this map of Treasure Island represented the books that Google has digitized, the clearly public domain titles would all be on this peninsula at the top of the island. I can read all those books now as they go online. Librarian Central also said that about 5 percent of the digitized books are currently in print, or otherwise actively licensed. So it’s conceivable under copyright law as it stands that Google could sell me, or my library, access to these books, so I could read them too.

That leaves about 3 quarters of the island in limbo— a land of digitized treasures buried under the soil that I can’t read. We could call it “Greater Orphanlandia”. I may have the proportions slightly off— James can correct me if he likes— but that’s more or less how things stand now.
What lives in Greater Orphanlandia?

- Books up to 140 years old
  - (Lacking other evidence, many projects assume books may be copyrighted up to that long somewhere)
- Books whose rightsholders cannot be reached
- Books whose rightsholders cannot be determined
- Books where the overhead cost of clearance is prohibitive
  - (CMU library study: Over $77 per book on average just to clear, before royalties)

[summarize above]…. It now costs a lot more to clear the copyrights for most books than it does to put them online in the first place.
When you get down to it, copyright clearance is really an optimization problem, especially when you’re doing it at large scale…. [summarize]
The job of copyright clearance can be made a lot easier if you’ve got good metadata to draw upon. Metadata’s especially important in a global context because you don’t just want to say “this book’s okay to provide online” or “this book isn’t”. You need to know why. The rules for copyright status vary a lot depending on where you are, who you are, and what changes have been made in your local copyright law recently. The book 1984 by George Orwell is in the public domain here in Canada, because this is a country where copyrights expire 50 years after an author’s death, and Orwell died in 1950. In most of Europe, and in the United States, it’s still under copyright, because copyright terms there are life plus seventy years, or in the case of the US, can run up to 95 years from publication for works from that era. That’s why Amazon suddenly recalled a bunch of copies of 1984 from Kindles in the United States; they’d gotten them from a publisher in a life+50 years country, and thought that the same rules applied elsewhere, and they didn’t.

But whatever the rules are where you happen to be, the facts that underlie a determination of copyright status hopefully don’t change…
Now, people have been doing this sort of fact-finding in the United States, and have been able to free a lot of books as a result. In the United States, copyrights can still be in force as far back as 1923, and because of this a lot of online book projects will only provide public access to books up to 1922. [Explain about Hathi Trust’s clearance; newspapers; low artwork renewal rate]

US Copyright renewal rates

- Hathi Trust has determined 52K of 95K 1923-1963 books it investigated were public domain
- Renewal rates for periodicals lower
  - No daily newspaper outside New York renewed issues dating before the end of World War II
  - Only a few dailies from 1923-1950 renewed at all
- Copyright for artwork often not registered or renewed
  - 1923: 23,030 copyright registrations for visual works
  - 28 years later: 198 renewals (less than 1% renewal rate)
  - Somewhat higher in later years, but not by much. Commercial art renewal rate very low
- Other genres:
  - Drama: Higher than artwork; radio scripts notable
  - Music: Registration and renewal rates relatively high
We recently observed Open Access week, where a lot of attention was paid to making current scholarly literature freely available. It’s nice to know that a lot of the scholarly literature of the past can also go freely online. When I analyzed the copyrights of journals in the JSTOR collection that were publishing before 1950, for instance, I discovered that most of them didn’t renew copyrights on any of their issues, and only a handful renewed copyrights on all of their issues.
Now, I should really talk a little bit about the Google Books settlement, because that could potentially simplify a number of the challenges I’ve mentioned. It’s essentially a workaround to some of these problems of copyright clearance, using the legal system. In particular, it makes it easier to create a large collection of readable online books, with very quick throughput. If there’s a book that’s covered by the settlement class, Google will consult a Book Rights Registry, created through the settlement process, and check if any rightsholder has made a claim or a request concerning the book. If they have, Google would follow the request they made about making the book available. If no rightsholder has come forward, Google can just go ahead and make the book available, using a standard pricing and revenue-sharing scheme. So size and throughput are enhanced, and costs can be carefully controlled as well.

Google’s risks also become very low; the settlement lays out a clear road-map for managing the rights to a copyrighted book and not getting sued. It even has a safe harbor for making public domain determinations in a manner that’s streamlined and simplified, if not quite 100% accurate.

[notes continue on next page]
In principle, I very much like the idea of cutting a straight path through what's now a very tangled thicket of copyright restrictions and rules. I'd even like it better if we can find a way to do this through legislation that lets everyone do what Google would be allowed to do in this class action settlement. But I can't pretend that this sort of legal workaround is going to solve all our access problems.

Remember, the settlement is not yet approved. If it is approved, it would only apply to what Google does in the US; everyone in other countries would still face the same access problems as before. And I don't want us to be dependent on a single Book Rights Registry, particularly not the kind the settlement proposes.
For one thing, we already have some copyright registries established that don’t provide everything we need. The US Copyright Office, for instance, keeps records of registrations made when a work is first copyrighted, but it doesn’t maintain current contact addresses of rightsholders. And the Book Rights Registry proposed by the Google Books settlement is both an information registry and an advocate for its clients, who are rightsholders. So its incentives are to make sure their clients get paid well. And that’s fair enough, but if we want help establishing that certain works are in the public domain, meaning we DON’T need to pay their clients, I can’t imagine they’ll be the most enthusiastic collaborators.

Fortunately we’ve seen that once some people start contributing information on copyright, others can add onto it, and we can start building up independent, collaborating registries. For instance, back about 10 years ago when I was at Carnegie Mellon, we started scanning book renewal records from the Copyright Office’s Catalog of Copyright Entries, and putting them online. From those scanned records, volunteers at Project Gutenberg made full-text transcriptions, and making those available. Stanford University then took that text, and other data from the Copyright Office, and created a searchable database for people to find out if a given book had had its copyright renewed. This database is one of the sources of information that Hathi Trust uses for its own copyright determinations, which it then makes available as part of its book metadata.
So you see how one group of people sharing metadata in an accessible fashion encourages other interested people to use and improve on it, in all kinds of interesting ways. And I hope we’ll continue to see this information spreading throughout the world. It would be a great help in encouraging the spread of all kinds of public domain material in other countries, like it has in the United States.
Okay. We’ve discovered a book that would be useful for us to read. We’ve determined that we should be able to access it. Now: Can we actually get to a copy online that we can use? This is a trickier question that it might appear at first.
You need to know: How can I get to a copy that I can use? And how can I make sure I can get to all of it? These seem like trivial questions, but they’re very important, both for people who want to read books online, but also to libraries and other organizations that want to preserve them. In the context of libraries, preservation is essentially MAINTAINING USABILITY OVER TIME. Nothing more, nothing less. So if we want to preserve something, we need to make sure we can get a copy that’s under our control, and not someone else’s. Online books can go away. Companies can change their business priorities, and stop maintaining an online library that we used to take for granted. Amazon can yank Kindle copies of George Orwell’s works down the memory hole, as I mentioned earlier. Libraries have traditionally been caretakers of our cultural heritage, and I want us to keep playing that role as books move online.
So, how hard is it really to find all of a work online? Well, I know a professor of English literature, who often works with somewhat obscure Victorian novels. One of the ones she’s blogged about recently is a three-decker novel called Oldcourt. By “three-decker” I mean a book that was published in three volumes. She was only able to find two of those volumes in Google Books, even though she’s used Google Books quite heavily over the years. In fact Google has digitized all three volumes, but there’s no obvious way to find all of them in Google’s catalog.
Fortunately, the library that lent the books to Google to digitize is a member of Hathi Trust, and their metadata includes a record of all three volumes used in this work. Moreover, when they got the digitized files back from Google, they added references to the digital files to each volume record. So if I look up the book in Hathi Trust, I can easily see all three volumes, in the expected order. There are various standard formats one can use to express this structure, such as OAI-ORE, so that other institutions can make and preserve their own copies of all three volumes if appropriate.

So we have a complete copy of this book online. But it’s not necessarily a usable copy. Hathi books are displayed as page images. That works fine for me, when I’m sitting in front of a nice big monitor. But it doesn’t work so well for someone who’s visually impaired, or someone trying to read books on a small Ipod screen. They may well want a different format, one with nice OCR’d text that can be reflowed or automatically read out loud. Someone 50 years in the future might need a completely different image format, our present-day ones having become obsolete. Someone living in Beijing might like a translation into Chinese, if one is available.

This sort of information about formats, languages and the like, are all things that we know how to encode in metadata, and they can help us keep this book usable for a variety of present-day and future audiences.
It isn’t always as simple as just finding the right format or language, though. Certain works, like the plays of Shakespeare or the Bible exist in so many different editions that finding a suitable one isn’t always easy. And it’s not something that library catalogs give you much help for, though librarians will, if you ask them.

Say you’re looking to read Hamlet. Well, there’s no one Hamlet that works for everyone. If you’re a high school student, and Elizabethan speech sounds like a foreign language to you, you’re probably going to want an edition that explains all those strange words the characters use, and it’d be nice if you could get summaries of the scenes too. If you’re a scholar, you might want a critical edition that faithfully reproduces the best source editions, highlights variations between them, and has an extensive critical bibliography. If you’re an actor, you want an edition that makes it easy for you to see what your lines and stage directions are, and gives you space to write in the notes the director’s going to give you.

[notes continue on next page]
So when someone simply finds or looks up Hamlet, why not give them an organized view of the different editions available to them, and they can easily see which ones are good for what they want to do? We're experimenting with doing that on The Online Books Page, as you can see here. Now, there's a fairly complex data model that this is loosely based on, called FRBR, where you can model things like Hamlet as works that have expressions, and manifestations, and aggregate editions, and we could go to town with all kinds of data and automated reasoning and the like. But I wanted to keep things simple for readers. So here we just have a page for the work Hamlet, and we give some background information that tells you what kinds of editions to look for, and then a bunch of editions that I just went ahead and categorized myself. [I could live demo this as well, but I think this screen-shot gives you the general idea.]

It sounds like this sort of special-purpose edition metadata wouldn't scale up very well. But most books only have one edition, and most others have only a handful of editions. There are a fairly small number of works like this one where there are a lot of editions, or where there are significant differences between editions that experts care about, and those are the ones where readers really could use a bit of expert help.
So if you’re maintaining a large set of metadata, you need to know where it makes the most sense to expend your efforts to meet your audience’s needs. Not that long ago, Penn started a film studies program, and we were told that it was very important for people in that program, which grew to be rather popular, to take full advantage of our video offerings. So we took a bunch of metadata records we already had in our catalog for videos. We spent a fair bit of time cleaning up the metadata, making sure we had people like the director and the principal actors clearly identified with their role. We also brought in details like film synopses, posters, and DVD cover art. We got a lot of this data from external sources, and when we put it all together, we were able build a very compelling special-purpose catalog. The year after we put up our video catalog at Penn, videos made up more than 10% of our total circulation, even though videos make up a much smaller share of our total collection.
Inviting collaboration

- Let others contribute information about your resources
  - If authorities, great; other users can be useful as well
  - Design incentives to contribute (either automatically, or for benefits)
  - With linked data, provenance, information can be automatically evaluated and segregated as appropriate

- Other important principles
  - Evaluate needs of your users; range, quality of sources
  - Share what you can (not just for altruism, but because it encourages contributions, improvement)
  - Accept and adapt messiness

[summarize; may want to skip ahead to big takeaways if short on time]
What about messiness?

- **Mess is inevitable, tolerable:**
  - Catalogs already have a lot of messy metadata
  - New techniques, tools help (auto-correction, fuzzy matching...)

- **Mess can tell us something useful:**
  - User input shows how “real people” classify, find, value things
  - We can adapt, augment controlled metadata accordingly

- **Mess lets us scale up:**
  - Wikipedia lets lots of people build an encyclopedia
  - LibraryThing lets lots of people build book catalogs

- **Mess can be progressively improved:**
  - OBP: Automated subject assignment, imported subjects -> appropriate, controlled subjects
  - Penn videos: Hastily cataloged entries -> detailed, high quality descriptions
  - Target improvement based on community needs

[Wikipedia is a useful case study for just about all the points on this slide]
The big takeaways

- **Metadata turns piles of volumes into libraries**
  - Greatly assists discovery, accessibility, and usability of books online

- **Metadata, like books, are for use**
  - Concentrate on the metadata that will do the most to meet user needs
  - Find the best sources; upgrade or infer from other sources
  - Use the networked community to multiply your capacity

- **We can make metadata scale**
  - Promote open data, open format standards, open APIs
  - Design incentives for getting needed metadata
  - Cultivate multiple sources, multiple stores for metadata, content
  - Automate wherever possible; plan for expert human intervention at critical points

- **“Automatically processable, humanly comprehensible”**

[summarize the big points] …Your potential constituency is bigger, smarter, and can contribute more than you alone. If you give it the right incentives and tools, they can do a lot for you....
For more information

- Slides for this presentation (posted soon):
  - http://works.bepress.com/john_mark_ockerbloom/
- Subject maps
  - http://labs.library.upenn.edu/subjectmaps/
- Copyright metadata
  - http://onlinebooks.library.upenn.edu/cce/
- Online Books Page (subject and work demos)
  - http://onlinebooks.library.upenn.edu/
- Me (and more writings on these topics):
  - Blog: http://everybodyslibraries.com/
  - Email: ockerblo@pobox.upenn.edu

These slides (except where noted) are CC-licensed: BY-SA

So– I hope I’ve given you some things to think for building up and maybe designing your own libraries of books online. If you’d like to know more about anything I’ve talked about on this whirlwind tour, here are some URLs you might find handy. And I’ll very happy to answer any questions you have now. Thank you.