Open records, open possibilities

John Mark Ockerbloom, University of Pennsylvania
I’d like to thank Charles Wilt and Karen Calhoun for organizing this forum, and inviting me to it. I’d like to thank both Karen and OCLC for their willingness to stop and listen to the concerns that many have expressed about OCLC’s proposed sharing policy. I hope this session can help clarify and address some of those concerns, and that the review board OCLC is convening will continue that process.

We in libraries often lament, sometimes unfairly, the costs of online information. When we consider buying a subscription or license to a new online journal or database, I sometimes hear the complaint that we’re just buying back from publishers the same scholarship that we pay our scholars to produce and evaluate. Whether or not that complaint is justified, there’s another cost of limited access to scholarship: the cost of missed opportunities to build on that work as broadly as would be possible with open access.
Not long ago, Peter Murray-Rust of the University of Cambridge lamented that so many of the basic facts about chemical compounds were locked away in proprietary databases. He wrote “We are more than 10 years behind other data-rich sciences because we protect data in archaic silos.”

He’s not just complaining here about the state of affairs, but saying that it can be different. Other fields of science have flourished with more open information sharing.

For example, the Human Genome Project’s open access genome data makes it easier for scientists around the world to discover how our genes work, and treat genetic disorders.
Physics, likewise, has longstanding norms of sharing that help populate huge open access resources like ArXiv.org, and even gave us the World Wide Web itself. Not only did people in the physics community invent the Web, but they made the deliberate choice not to seek out intellectual property constraints to restrict the use of what they had created.

It’s worth remembering that many early discussions of the future “information highway” were pretty pedestrian. They imagined things like bigger and better cable TV, in 500 high definition channels, with just enough interactivity to let you buy things shown on-screen inventing and opening up the web showed us how much more was possible and today’s information highway is much richer and more varied than most people imagined a generation ago.

Today the library world is thinking long and hard about its own future. Many of you here will remember, for instance, the report that Karen Calhoun published in 2006 about the changing nature of the catalog. One of its more controversial recommendations was to dismantle the Library of Congress subject headings system in favor of simpler keyword-based subject descriptions. One of the more compelling arguments in favor of that recommendation was that, with the tools we had for working with subject headings, LCSH classification was costly and error-prone, and the search and browse features of our OPACs made subject-based discovery tedious and confusing, and of little benefit to most users.
But when the report came out I wondered “What if we had better tools for working with subject headings?” And I created a browser based on subject maps that some of you may have seen.
This browser uses LC subject headings to cluster together similar subjects and books with those subjects. Here, for instance, we see items on cooperative cataloging right next to other items on library cooperation, and a variety of other related terms. This lets users explore subjects much more easily than traditional OPACs allow, and it uses a richer network of relationships than purely facet or keyword based searching gives you.

I’ve had this in production now on The Online Books Page for a couple of years now, and based on the usage logs and feedback I’ve seen, it’s a popular way of exploring my index of free books online.
But can it be scaled up? What if, instead of just mapping the 30,000-odd books I list on the online books page, we’re trying to map a catalog that’s a hundred times bigger: our university’s library catalog, with more than 3 million items?
Well, it works there too. I don’t know if you can see it on the slide, but not only are there more books shown under Cooperative cataloging, but there are more relationships between that subject and other similar ones. That’s because my program looks not only at subject headings themselves, but how they’re used in our catalog’s bibliographic records, and infers new relationships from that data. We don’t have this in production in our university library catalog, but it’s a promising technology I hope we will support eventually.
Subject maps: 125M scale

But why stop there? Our users today have access to a lot more than what we list in our local catalog. What if we scaled up our subject maps again, not even 100-fold this time, just 50-fold, to include all the content they can borrow from our peers, or view in our online databases, or get via ILL? What, in other words, happens when we have subject maps the size of WorldCat?
Well... I don't know. I don't have the data to make these maps, and under policies like the one proposed for WorldCat, my use of its data would be considered unreasonable, since it would have not only similar size, but also similar function, to WorldCat.
Let’s look at another example. The Google Books settlement agreement has a lot of people excited about all the books that will be readable online, but it’s also raised concerns about monopolizing control over much of our cultural heritage. One of the things that tends to concentrate control is the scarcity of information about copyrights. More ubiquitous information about copyright would allow a wider range of organizations to make more material available. For instance, a few different organizations have digitized back issues of the celebrated children’s literature magazine Saint Nicholas. Many of these issues were put up by Google. But Google isn’t showing issues of this magazine, or virtually any other past 1922, and it’s not showing any issues at all outside the United States; basically, because they don’t know whether or not copyright applies to them.

It turns out that St. Nicholas is still under US copyright from 1923 onward, but that’s very much the exception rather than the rule.
A couple of years ago, I managed to compile a comprehensive set of copyright renewal data on periodicals published prior to 1950. I found some interesting things, like the fact that most scholarly journals of the time did not renew any of their copyrights, and that only a handful of newspapers did. Those can be very useful findings for many projects, and are possible when we can analyze the entire relevant data set.

Now, the settlement is going to create a Book Rights Registry that will include a lot of copyright information. But that registry will be controlled by rights holders groups, and will manage information for their interests. If you want to use a work that may be under copyright, they’ll be happy to collect royalties, but they won’t necessarily tell you if there are public domain or freely licensed versions you can use.

Thankfully, there are groups building independent copyright knowledge bases as well. One of them, in fact, is OCLC, which has been piloting a promising looking Copyright Evidence Registry that hopes to acquire facts about copyrights from a wide variety of sources. Unfortunately, the current terms of use for the registry are at least as restrictive as those OCLC is proposing for WorldCat, and I can’t download the degree of data I’d need to do the sorts of analyses I just showed you.

Now, OCLC has come out with various APIs for querying WorldCat and some of the other knowledge bases they maintain, and making some of them free to use. That’s great, but you can only do so much through queries alone. There’s a lot more you can do when you can crunch through the whole data set.
Back in 1987, when the current WorldCat guidelines came about, it was hard to imagine anyone other than a major enterprise being able to handle a data set the size of WorldCat. But the last time I was in Target I was seeing portable half-terabyte drives even more compact than this one on the shelves for under $100. If you allocate 4000 bytes per MARC record—and most MARC records are quite a bit smaller—you could fit all 125 million bibliographic records in WorldCat on one of these, pop it in your purse or briefcase, and take it to ALA or wherever you want to go. At the same time, computing power, and free open source database and indexing software have progressed to the point that you could analyze all this data on quite inexpensive computers. If there were more time, I could also talk about what you could do for things like shared collection management, quality improvement, and preservation planning.
But there are a few things standing in the way of all that. And some of those things come straight from the proposed OCLC policy.

It claims a right to control distribution and use of all contributed WorldCat records, including those to which OCLC has made no original contribution. Its terms of use and distribution are much more limiting than the Creative Commons licenses that OCLC has sometimes compared them to. Okay, some Creative Commons licenses do restrict commercial use, or derivatives, but other than that, they let people use works however they want, and distribute them to whoever they want, with no need to get permissions.

And with that “sole discretion” clause, the proposed WorldCat restrictions would effectively give OCLC the right to decide whether any application using WorldCat-derived data in any way should be allowed to exist.

Now, I know that there are a lot of smart, reasonable folks working for OCLC, and for all I know they may well be quite happy to give permission for the sorts of applications I’ve been describing. But just the simple act of requiring permission to innovate can stifle innovation substantially. And it doesn’t help matters when OCLC’s proposed policy also claims a right to unilaterally make further changes to WorldCat policy at any time, without prior notice or consultation.
There is another way we could go. And we see it in the world of open source. Open source software, software that’s given away along with rights to copy, use, modify and redistribute it, has been with us for many years now, through economic boom and bust. Open source products like Linux, Apache, and Firefox underlie a huge number of servers, laptops, smart phones, and web browsers in use today. Open source software does have terms and conditions for sharing, but their licenses all allow anyone to use and adapt the work for any purpose, including commercial purposes. Which is a good thing, since it turns out that commercial use helps grow and sustain many open source projects. And open-source-style licenses are often used for content as well; Creative Commons license terms like “Attribution” and “Share Alike” reflect open source sharing principles.

Now, OCLC has said that it’s not economically feasible for them to open up WorldCat like this. I’m not an expert in finances or OCLC’s operations. It may well be that their current business model wouldn’t support a fully open WorldCat database. There is a significant cost to coordinating a high quality knowledge base that has thousands of contributors. That coordination is work that we value, and it has to be paid for somehow. But I hope I’ve also demonstrated that there is also a substantial cost in keeping our metadata closed off from full access and use. That cost is largely in lost opportunities, so it’s hard to quantify in dollars and cents. But it could make the difference in determining how central or marginal libraries are in the new online information society.
I will point out that many open source projects find ways not only to sustain themselves but to flourish, and many of the ways in which they do so could also be done with a union catalog like WorldCat.

Some common themes involve keeping overhead low, and that’s sometimes accomplished by shifting work from central control to community control. Wikipedia and Craigslist content, for instance, is largely created and monitored by its users rather than by its staff. Sometimes you can change what you charge for. Some well-respected open access journals have found that they can get authors to pay for publishing their articles, instead of readers paying for viewing them. Sometimes you can profitably sell additional services and products that are tied into what you give away; that’s been successful for big computer companies like IBM, Novell, and Red Hat, and also fuels open source upstarts like Word press or Evergreen.

Some organizations invite people to pay for memberships that give them a say in what the organization does. For example, the LOCKSS digital preservation project lets anyone download its basic software for free and preserve content on their own. But if you pay to join the LOCKSS alliance, among other things you can help decide what scholarly content the LOCKSS network focuses on preserving. And there are enough libraries who’ve bought into the alliance that its director tells me that LOCKSS is already financially self-sufficient.

I think all of these basic approaches can help sustain a shared catalog like WorldCat. For instance, I think of all the services that OCLC provides, like ILL and WorldCat local, that benefit from their close ties to WorldCat and can help sustain it. And I think it should also be possible for libraries that value the metadata coordination of WorldCat to help support that coordination, in cash or in kind, even if they don’t pay for the data itself. The more a shared resource like WorldCat is a point of concentration for a community, the easier it can be to sustain a service-based or community-supported business model.
I’m not saying there will or should be just one point of concentration. Readers and librarians have used a number of concentration hubs, to find out about books. Some popular hubs like Amazon and Google, are commercial sites where readers may miss many relevant resources that libraries provide. Some of these hubs have not made it: RLG’s RLIN is no longer with us, for example. Some hubs focus on new audiences and applications, like LibraryThing or OpenLibrary. And some have taken some bold steps towards openness, such as biblios.net with the 30 million bibliographic records it says it freely licenses.
Some partial steps towards openness

- Recognize public domain status of citation and other factual data in contributed records
  - and don’t forbid their use or dissemination
- Honor individual contributors annotating their records with open access provisions
  - (e.g. use their 996 notation instead)
- Let members collectively ratify or reject basic changes to their cooperative charter
  - (such as data use and transfer policy changes)

Even if the WorldCat cooperative isn’t prepared to share all its bibliographic records with the world, we can at least stop putting roadblocks in the way of those who are ready to share. We could acknowledge that substantial portions of a catalog record are actually in the public domain, since copyright does not apply to unoriginal facts like a book’s citation or rights data, no matter how much sweat of the brow went into compiling these facts. We could clear the way to unfettered use of that public domain data, and that would do a lot to promote public goods like openly shared copyright knowledge bases.

We could let members who want to make their records more open annotate their WorldCat contributions with open licenses, which would not be overridden by more restrictive WorldCat rights statements.

And, whether or not you agree that the WorldCat cooperative should make its records more open, I think the membership should get the chance to decide for itself how to proceed. Every cooperative or commonwealth I’ve been involved in gives its members the last word on any major changes to its principles. Constitutional amendments have to be ratified by the states or the voters after the legislature passes them. And my natural foods co-op back home puts major rule changes up for a vote in open membership meetings. While I can’t speak for the WorldCat cooperative myself, I would hope that its members merit having the same kind of say in their collective future.
Because I really believe that we now have great opportunities to do more with our library resources than many of us thought possible, if we don’t close ourselves off. I’ve shown how open access principles can enable new inventions with unforeseen benefits: improved access and discovery to knowledge online, powerful and secure open source software, even the Web itself. And I’ve touched on some of the ways in which open access models can be sustained.

In the end, it’s our libraries and our collective records at stake. And we can, and should, choose the kind of future we want for them.

Thank you.