Published in Technicalities, Jan./Feb. 2008, v. 28, no. 1

Metadata Wants to be Free (and Muddy)

Diane I. Hillmann

With the recent merger of OCLC and RLG, the near monopoly of OCLC over library metadata distribution is complete. This is, to some extent, a good news/bad news situation. The good news is that OCLC is as aware as any organization in libraryland (perhaps more aware, given its dependence on libraries for its future) of the need for new models and strong innovation in the very near term. They've taken leadership in surveying the landscape and their membership (which includes all but the smallest libraries) to determine where they should spend their resources to meet library needs.

The bad news is that with the exception of the growing open source movement, OCLC has no effective competition, and the long-noted tendency for the organization to operate as a virtual monopoly will likely proceed unchecked. Although the association with RLG has encouraged OCLC to begin to store the record versions created by members and cluster them as RLG did, the OCLC model of metadata distribution is still, for the most part, based on a "master record" concept. In other words, the services that OCLC sells to its user/members are based on the notion that OCLC will manage its database to ensure that member holdings will be attached to a one "correct" record for a particular item. This one record will be the basis for determining the particular libraries that hold an item for the purposes of interlibrary loan or resource sharing. This model requires that "duplicate" records be reconciled, clustered and/or expunged to ensure that there is one, and only one, record per resource.

It should be noted that one of OCLC's new services--WorldCat Local--is also to a great extent based on the selfsame concept with the difference that attached holdings information for a particular library can enable a virtual view of that library's holdings, a consortial or state view, or any other conceivable aggregated view. WorldCat Local is attracting a great deal of interest, particularly from large academic libraries attempting to shift resources from the cataloging of redundant trade books and mainstream serials to focus more on materials that they can uniquely claim to own and manage.

Advantages

One big advantage of the current distribution model is predictability. Because the "master record" supported by OCLC and imported into local systems as the basis

for member's integrated OPAC/ILS is based on forty years of so of standardization, the product is generally very predictable. This is a similar predictability to that exploited by successful franchise fast food operations--sort of a MacDonald's effect--reassuring the consumers of the product that what they buy or use will be what they expect it to be.

Another advantage is that the one remaining bibliographic utility takes responsibility for testing changes in the MARC format before they're unleashed on an unsuspecting public. Prior to acceptance of changes, a utility will generally analyze the impact on their system, thus protecting the distribution system that most current library systems still rely on.

Limitations

Despite the admitted efficiencies of our current data and its distribution system, there are also limitations embedded in that efficiency. One significant limitation is that our current data is really designed to support discovery in a closed system--our current ILS system(s). We are now in an environment where there are a variety of discovery tools available to our users, and for the most part they are choosing other options.

The irony of this is that we spent many years asking our vendors to integrate our back end and discovery systems even closer together, and of course, we eventually got what we asked for. What we didn't expect as we sought the efficiencies of one system built around one data format was that the expectations around resource discovery would change so much all around us, limiting the usefulness of what we built with our vendors.

Perhaps the most worrisome aspect of the current model is that we're finding it enormously difficult to extend our model to an increasingly digital environment where the old "granularity consensus" is close to completely unraveled. By "granularity consensus" I refer to the accepted view that libraries catalog things at book level and serial title level, commercial services deal with article level, and archivists cope with collections. By optimizing our distribution system for de-duping and rigid standardization--and firmly supporting the OCLC master record-based view of the world--we limit our ability to experiment, to merge data streams, to accept data from vendors not using MARC, and to increase our ability to provide relevant results to our users by drawing on their expertise in new ways, much as our 'competitors' in the web world are doing so well.

Change is difficult: while we continue to maintain the old model in our heads as fixed as the planets, it's difficult to consider other models, much less the steps needed to accommodate the old and new for a rational transition.

Another distribution model

One of the most interesting recommendations from the Library of Congress Working Group on the Future of Bibliographic Control is that we "Re-examine current economic model for data sharing in the networked environment." [1] Oddly enough, this recommendation comes under the rubric of "eliminating redundancies," which I think is wrong headed--in my view redundancies of various kinds will be a necessary feature of what passes for a transition from the current model to any future models. Particularly because we have been somewhat slow to recognize that other models exist and have experimented only fitfully with the ones available at present, we're not necessarily in a good position to insist that redundancy is unacceptable.

It seems clear to me that the low hanging fruit available to us in the realm of new distribution models is the one alternate distribution model already in place: the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). [2] This is a distribution model that is well correlated with open source software already under development. Karen Schneider, well known as the *Free Range Librarian* talks about this in a post on October 11 of this year, enticingly titled "Our Exhilarating New Mix and Match, Slice and Dice World." [3]

She speaks about the necessity to try these new models out:

> *We, as a profession, could even build out own "silo" — a kind of librarian-built "Free OCLC." (The Free-C?) Both Evergreen and eXtensible Catalog are designed in part on the premise that the only good data is Web-readable data which can be harvested by protocols such as OAI, placed in a central catalog, and made available for all to use. (Since I'm hungry, I'll describe OAI as a humongous straw that can suck in all kinds of data, like those straws used with pearl tea.)*

This notion that "metadata wants to be free" is a really important one, and flies in the face of another of the LC Bib Control WG recommendations, which suggests that what we really need is to think more about the financial disincentives of the current model, and create more incentives, particularly for the Library of Congress. This is a bit disingenuous, it seems to me, given that another thrust of that report is that those of us outside LC should participate more in data sharing, and not depend so much on LC.

Perhaps what we need more in the short term, certainly, are incentives that depend less on return on investment and more on demonstration of usefulness. It might well be too early to build financial incentives for a new model (while a bit too late for the old one), since nobody is quite sure what we're selling, or buying, yet.

But Karen also includes some very realistic cautions:

> *None of these initiatives are "free" as in "free beer"; they are all free kittens, in need of care and feeding. (Companies such as Equinox and LibLime have sprung up to provide third-party maintenance.) All of them exist because some organization or person (or both) put up cash or sweat equity. Some, like eXtensible Catalog — which just got second-year grant funding — are still a gleam in someone's eye; some, like LibraryFind, are still very, very new, though undergoing rapid development. All need strong sustainability models to keep going.*
>
> *But they're all much different than the model we've been using for over a hundred years, where the cards — and then the records — were stuffed into local catalogs, and they're also different from the model of the last several decades, where the brass ring was the super-secret code, and the support and development were where everyone economized.*
>
> *Everything is different. That may even mean we're different. And if so — then vive la difference!*

One of the big differences is that lack of a natural "center" in this new distribution system. OCLC and LC are really the "centers" of the current system, where LC is still thought of (by some, and not others, it must be said) as the central source of the "best" cataloging that we all should prefer, with OCLC is now the only center piece of the distribution system. We all buy our records (including our LC records) from them, and they perform services like normalization and de-duping on our behalf.

In this alternative world, there is not a center, unless you count something like the OAI Registry at the University of Illinois/Champaign Urbana as a logical center--since it's there that one discovers where the OAI servers are and what they have available. [4]

Clearly the new model, without a large, centralized source of records, requires that some libraries will need to build and support expertise that was formerly outsourced to OCLC. In the old days, the thought of supporting such expertise was what pushed many libraries into the arms of ILS vendors and their supposedly off-the-shelf products, but clearly that experience has taught us that there's a high price to be paid for outsourcing everything to commercial entities.

So where's the MUD?

Firstly, I have to admit that I didn't make up the term "mudball"—I borrowed it

from Sally McCallum, who used it many years ago as a somewhat disparaging term for authority data that included sourcing at a statement level. The mudballs that I like to talk about are similar things, but in this case aggregations of metadata statements with full provenance. Mudballs are not necessarily a feature of a new distribution system based on OAI-PMH, but I happen to think that if we're to go beyond the information we currently see in records, and include things like usage data and iterative data that started somewhere in the outer galaxies of publishers and others, moving to statements with provenance will be a very important step to take—a point made more extensively (with illustration) in "Improving Metadata Quality: Augmentation and Recombination," published a few years ago. [5]

In our current record-based model, we know vaguely who was the last library to "touch" a record (if they followed the standard and added their code to the record). This doesn't give us a clue what that library changed. In a statement-based "mudball," where data is used and reused in various systems, gathering new information at every station, questions like "Who made this statement?," "How did they make it?," "Was there an algorithm, a human, a combination?" will be essential to know as we determine what information we would like to serve out to our users, and what we might use for other purposes--to rank search results, for instance. Other important questions might include, "When was the statement made?" or "Is the statement still valid? Remember that no assumption can be made that any, or all, of the mudball statements were made by catalogers.

I came across a good example of what this might do for us while discussing a possible development project a few weeks ago. The presenter was showing how library data might be used to populate a Blackboard system, with descriptive information about the resources available to particular classes. I asked, why wouldn't the system that supplied the library data want to know that it was used in the Blackboard system? Wouldn't it help rank materials as to their importance to other users if the information gathered for Blackboard was actually fed back, along with the details of the class, when it was used, who was the teacher who chose it, etc.?

As this column is being written, the LC Working Group on the Future of Bibliographic Control has just released its recommendations, though not yet the full report. Astoundingly, many of the ideas that have been considered too "out there" to be real are now becoming mainstream. Discussion is just beginning on what this all means, so stay tuned, and keep your seatbelts fastened—it's going to be a bumpy ride.

References

[1] Library of Congress Working Group on the Future of Bibliographic Control.

Available at: http://www.loc.gov/bibliographic-future/

[2] Open Archives Initiative Protocol for Metadata Harvesting. Available at: http://openarchives.org

[3] Schneider, Karen G. "Our exhilarating new mix and match, slice and dice world," October 11, 2007, *Free Range Librarian*. Available at: http://freerangelibrarian.com/2007/10/11/our-exhilarating-new-mix-and-match-slice-and-dice-world/

[4] The University of Illinois OAI-PMH Data Provider Registry. Available at: http://gita.grainger.uiuc.edu/registry/searchform.asp

[5] "Improving Metadata Quality: Augmentation and Recombination," paper by Diane I. Hillmann, Naomi R. Dushay, and Jon Phipps, presented at the DC2004 Conference, October 2004. Available at: http://purl.org/metadataresearch/dcconf2004/papers/Paper_21.pdf