bid textos universitaris de biblioteconomia i documentació
ISSN 1575-5886
DL B-19.675-1998
número 22
juny de 2009
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

21st Century Description and Access

[Versió catalana]

Roy Tennant
Senior Program Officer
OCLC Research
tennantr@oclc.org

Many recent articles, reports, and presentations in the library profession have identified major environmental, technological, and philosophical changes that are requiring us to completely rethink how libraries perform bibliographic control.[1]

Even the term *bibliographic control* is an anachronism:

> *Bibliographic*: This term has a lot to do with published literature – mainly book literature but to some extent also journal–, but not much at all to do with many of the forms of communication now being used on the Internet. When we harvest web sites, what is the part that we can call *bibliographic*? How bibliographic is, for example, a collaborative blog? We desperately need a more general and generic term for this, which for lack of a better idea I've decided to select *descriptive*, since no matter what the resource we are basically talking about describing a resource.
>
> *Control*: What part of this do I need to explain? *Google* is mass digitizing the contents of our largest libraries and making the output available for full-text searching. *LibraryThing.com* and others are being quite successful at letting anyone and everyone assign whatever terms pop into their heads to the books they own. If this was all garbage, and easily ignored, it would be one thing, but it isn't and we can't…
>
> I no longer believe in the future of bibliographic control. I no longer believe that the term *bibliographic* encompasses the universe in which we should be interested, and I no longer think *control* is either achievable or even desirable. We have entered the age of *descriptive enrichment* and we'd better get bloody well good at it.[2]

By *descriptive enrichment* I mean a set of procedures, human, machine, and combined, by which we capture descriptive information about an item or collection and continuously and iteratively enhance it. Perhaps an even more important distinction from past practice is that this process of continuously and iteratively enhancing the data should include not just professional librarians but also library users.

At the American Library Association Annual Conference 2008, Joe Janes described hearing someone say "what if a book got better every time it was read?", which he reformulated as "what if a library got better every time it was used?"[3] It is just this process of getting better through use that I'm attempting to describe. This is also why I avoid the *cataloging* term, which denotes a professional process performed by trained catalogers. I wish to suggest that by using the broader term *metadata* we can encompass not just library cataloging but also other activities that contribute and enhance descriptive information about items.

The goal of bibliographic description

To know whether we are eventually successful in our endeavor to re-engineer our bibliographic practices, we must establish what our goal is in doing these activities. I believe our goal should be *to enable independent, self-sufficient people to find what they seek*. I specify *independent, self-sufficient people* on purpose, since I believe that is what virtually everyone wishes to be. Most library users prefer to not speak to a librarian. Most prefer to find what they want on their own, even if it is painful and they aren't good at it. Our job, then, is to make the finding process as easy and painless as possible.

This means that the most effective bibliographic strategies will be those that are unapparent to the user – that is, they are so intuitive that they recede into the background and *things just work*. To accomplish this requires a great deal of complexity to be hidden from the user. That is, in order to be simple on the front-end, our systems must be complex on the back end. Heretofore, most library finding tools are the exact opposite of this. Users of our various catalogs and indexes are typically required to select a field to search, such as author or title, which allows the system to not have to function well without this foreknowledge. Present day systems such as *Google* and *Amazon* have now made it obvious to our users that systems can function well without such guidance, and they have little patience for systems that require such coddling.

Obviously systems design will be key in creating usable finding tools, but metadata will also be important. Despite the fact that more of the printed record is being digitized all the time, and therefore becoming keyword searchable, metadata is still essential for disambiguation, filtering, sorting, and ranking.

Although our need for bibliographic description continues unabated, we must find ways to create, enrich, and manage it more efficiently. Libraries have fewer resources to devote to painstaking processes. So our challenge is fairly stark: become radically more efficient in our descriptive procedures, use innovative techniques to tap into other sources for description, including our user communities, and use the power of machines more effectively, all with fewer resources.

Cooperate globally

Libraries have a lot in common throughout the world. Many of our procedures, services, goals and strategies are similar. By cooperating on a global scale we can begin to bring the benefits of webscale computing to libraries worldwide. Our best chance at this is offered by OCLC, a nonprofit membership organization that is "dedicated to the public purposes of furthering access to the world's information and reducing the rate of rise of library costs. More than 71,000 libraries in 112 countries and territories around the world use OCLC services to locate, acquire, catalog, lend and preserve library materials".[4]

The flagship service of OCLC is *WorldCat*, the union catalog of library holdings. Searchable by anyone at *WorldCat.org*, presently holds over 135 million records with over 1.4 billion holdings attached.[5] As the holdings of more libraries from around the world are added, including major national libraries, it is coming to represent the linguistic and cultural diversity of world culture.

From this vast database a variety of services can be created. For example, *WorldCat Identities* (available either directly at <http://worldcat.org/identities/> or integrated as a part of *WorldCat.org*) offers a single web page that aggregates a great deal of information about each author, which has been extracted by software from the data contained within *WorldCat*. Links are also made from these pages to other sources of author information.

There are many other opportunities for cooperation on a regional or global scale. Certainly the *Europeana* project is one such example (available at <http://www.europeana.eu/>). By working together we are much more likely to all achieve our aims than if we work alone.

Go up the chain

For many years libraries have been creating metadata (cataloging) by inspecting the title page and title page verso of a book and recording various bits of information about the work, assigning subject terms and classification, etc. This has been necessary because there was no other way to obtain the information in a usable form (i.e., computer readable). Those days are fast disappearing and yet many of our processes remain the same.

What has changed is how books are produced, distributed, and sold. Now metadata moves along the entire book selling chain from publishers to wholesalers to retailers. *Amazon* requires it, as do many retailers. What *Amazon* requires is that publishers provide information about the books they have for sale in the ONIX XML format.[6] This enables *Amazon* to effectively deal with the large number of products it needs to manage to make a profit. Meanwhile, although publishers are often providing rich descriptions of books in a machine-readable form (many of which include a summary, a brief author biography, pull quotes from reviews, etc.), libraries have until recently ignored these descriptions.

In an attempt to change this, and at the same time potentially create a new service for publishers, OCLC created the Next Generation Cataloging Project. The web site describes the roles of the various partners

and what they hope to achieve:

> *The role of publishers and vendors*: Publisher and vendor pilot partners provide OCLC with title information in ONIX format. OCLC crosswalks the data to MARC for addition to *WorldCat* and, where possible, enriches the data in automated ways through data mining and data mapping. Enriched metadata is returned to publishers and vendors in ONIX format for evaluation of OCLC enhancements.
>
> *The role of libraries*: Library pilot partners evaluate the quality of metadata added to *WorldCat* through this process and provide feedback on its suitability for use in library technical services workflows.
>
> *The role of other partners*: Publishing industry partners such as BISG (Book Industry Study Group, Inc.) assist OCLC with publisher industry data standards and terminologies, as well as providing a forum in which to share ideas and results with the industry.[7]

Through projects such as this, libraries will hopefully be able to make more effective use of metadata that publishers create – to both reduce our workload as well as enrich our records with previously unavailable content such as summaries.

Mine the data

Any large aggregation of data is likely to afford opportunities to "mine" it for information that can only be discovered when it exists in large quantities in one place. For example, several years back OCLC began using the number of libraries that hold a book to increase the relevance ranking of search results in *WorldCat*. That is, the more libraries that held a book the higher it would rank in search results. Only when you have a large aggregation of such data does an opportunity like this become available.

OCLC is also examining the use of MARC fields and subfields across the entire *WorldCat* database, to see what can be learned. My colleagues in OCLC Research have a number of projects related to this, under the broad category of "Renovating Descriptive and Organizing Practices".[8]

Enrich the data

The typical library catalog record, as impressive as the work required to create it may be, still lacks many things that our users often wish to see. Should it not be possible to see the full-text online, any portion of the work, such as the table of contents, index, cover art, etc., would be useful to enable the user to determine if they really need to see the book.

As mentioned above, ONIX records from publishers offer libraries a potential method for enriching our bibliographic descriptions with additional information that users find helpful. There have also been various attempts to digitize tables of contents and make them more widely available. Perhaps the most widely known of these projects is by the Library of Congress Bibliographic Enrichment Advisory Team (BEAT), which has provided enhanced information for hundreds of thousands of library records.[9]

As more library data is aggregated in one place, various data mining opportunities present themselves, as noted above. One such opportunity is being explored by OCLC is an experiment with an "audience level" rating. As the project web page states, "This research project explores using library holdings data in *WorldCat* to calculate audience-level indicators for books represented in the *WorldCat* database, based on the types of libraries that hold the titles."[10] Such an indicator might be used to filter search results based on relevance to a particular user. A college researcher, for example, will likely not be interested in items aimed at an elementary audience.

Empower users to add value

As many organizations have discovered, Internet users often have information of value to contribute if they are provided the opportunity. Cultural heritage institutions are beginning to enable users to contribute tags, reviews, ratings, and other types of enhancements to standard descriptions by librarians, archivists, and curators.

The *Steve.museum* project has performed research into user tagging of digital museum objects, and the

findings so far seem to indicate a great potential to enhance retrieval of museum objects by allowing users to tag items with their terminology.[11] "Social tagging and folksonomy could make a positive contribution to the accessibility of on-line art museum collections," states Jennifer Trant of the project.[12]

In the figure below, Trant illustrates that while museums may wish to document one set of features about an object, users can embellish the description with words more likely to be meaningful to them.

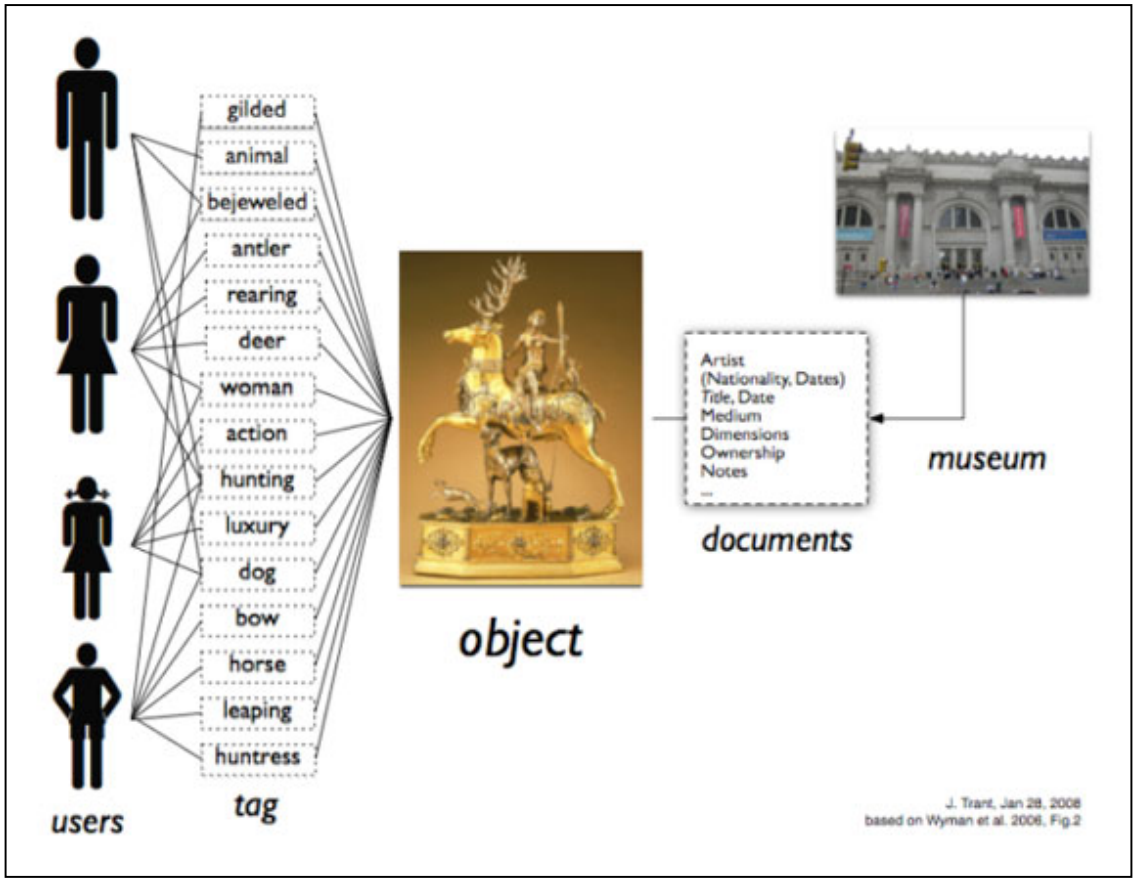

Figure 1. Differing views on object documentation: while users tag from multiple perspectives, the museum documents from a single, institutional point of view.[13]

As those who have taken on the responsibility of collecting, managing, and preserving our cultural heritage, we may sometimes fail to acknowledge that those not trained in these tasks still have much to offer. In particular, those who live in a community depicted by our collections of historical photographs may have much to offer. For example, Figure 2 depicts a comment from a user about an historic photograph on the *Maritime History of the Great Lakes* web site managed by the Halton Hills Public Library in Ontario, Canada.

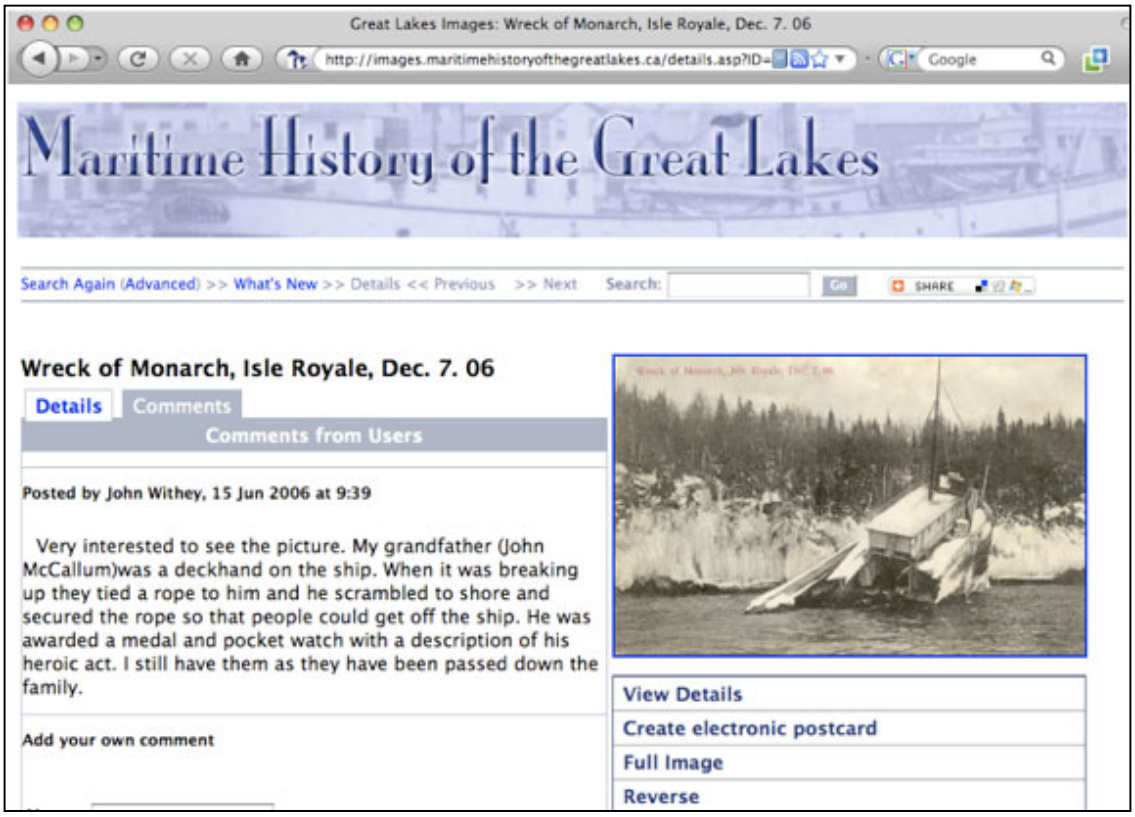

Figure 2. User comment on an historic photograph in the *Maritime History of the Great Lakes* web site, managed by the Halton Hills Public Library of Ontario, Canada.

Make the data available for others to use

Bibliographic data can be used in many different contexts to provide a variety of useful services. To better enable this data to be "mashed up" with other data in new kinds of services it needs to be made available to software through protocols and structured formats; i.e., XML.

The three main ways to expose bibliographic data for use by others is via an application program interface (API), harvesting via the *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH)[14], and as linked data.

Application Program Interface (API)

APIs are a useful way to expose bibliographic data when you do not wish to provide the data for harvesting or downloading. However, every time an application uses your data your server is called by that application. Libraries may not wish to host such traffic, in which case harvesting may be a better method to expose data. *TechEssence.info* maintains a list of library related APIs.[15]

A good example of API use in a library context is the growing set of Web Services being exposed by OCLC through its *Grid Services* initiative. These services are supported by the OCLC Developer Network, where a number of developer support services can be found: service documentation, sample code, a mailing list and a blog, etc.[16] There are about a dozen separate services available, with more being added all the time.

Harvesting

Providing bibliographic data for harvesting means supporting the *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH), which is fairly easy and is already part of many library software applications. Probably all institution repository software applications support this protocol, with some of the major ones being DSpace, Fedora, ePrints, and CONTENTdm.

The protocol itself is fairly simple, with only six verbs transported via HTTP as a Representational State Transfer (REST) request. The only metadata format required is Dublin Core, but some sites support other, richer metadata formats for downloading, such as MARCXML and/or MODS.

Linked data

As stated by the "Linked Data FAQ" by Structured Dynamics, "Linked data is a set of best practices for publishing and deploying instance and class data using the RDF data model, naming the data objects using uniform resource identifiers (URIs), thereby exposing the data for access via the HTTP protocol, while emphasizing data interconnections, interrelationships and context useful to both humans and machine agents".[17]

A simpler way of describing it might be as a set of best practices to expose structured data on the web in a form usable by software applications. Also, the purpose of linked data is to enable connections between related data sets so that a variety of potential applications are enabled that would not be possible without these linkages.

A prime example of linked data is the *Authorities and Vocabularies* web site by the Library of Congress.[18] This site presently provides the Library Congress Subject Headings vocabulary for downloading, searching, and linking. Software can request a particular subject heading record, have it returned in one of several different structure formats, and discover broader, narrower, or related terms as well as the identifiers (URL) of those terms within the data set.

It isn't clear yet what exposing structured data in this way will enable, but unless it is done we will never know. At least now anyone who wishes to have access to this important vocabulary can obtain it in a way useable by software.

Expose the data where people congregate

Although libraries have invested a great deal of time and money in constructing portals for users to discover library content – both print and digital – it is clear that the vast majority of Internet users are not aware of these sites and fail to discover much of the digital riches that we are beginning to accumulate online.

To try to address this problem, last year the Library of Congress began an experiment to expose some of its digital content on *Flickr.com*. Dubbed the *Flickr Commons*, the Library of Congress submitted about 3,000 black and white and color photographs from two collections. The results were stunning:

> In the first 24 hours after launch, *Flickr* reported 1.1 million total views on our account; a little over a week later, the account had received 3.6 million page views and 1.9 million total visits. That included over 2 million views of the photos, and over 1 million views of the photostream. By early October, LC photos were averaging approximately 500,000 views a month and had crossed the 10 million mark in total views and the 6 million mark for visits.[19]

The reason for this astounding traffic is simple: *Flickr.com* is where many more users frequent, and therefore exposing the previously "hidden" content on a portal where they frequently visited brought this material to their attention.

Web sites that measure traffic can illustrate the disparity between visits to the Library of Congress web site and *Flickr.com*. For example, Figure 3 illustrates the traffic ranking of these two sites by *Alexa.com*. *Flickr.com* is ranked as 33 in traffic at the time of this writing, and the Library of Congress is ranked as 2,904. Although a ranking under 3,000 is still respectable, the difference between the two is telling.

Libraries, museums, and archives have compelling content, but without getting it exposed where people can be found it will remain undiscovered and unused.

Figure 3. *Alexa* daily traffic trend for *Flickr.com* and *Loc.gov*, November 2008 – early May 2009.

Enable Effective User Interaction

For many years our bibliographic search tools have hardly progressed at all in enabling effective user interaction. To this day, despite evidence from sites like *Amazon* that searching could be less painful, most of our bibliographic search tools are enmeshed in technology that was state-of-the-art in the 1980s. Many writers have thus recently disparaged library catalog systems (often called, as if to cement the anachronistic nature of these systems, online *public access* computer systems or OPACs) as being unintuitive, difficult to use, and ineffective.[20]

In recent years, however, there has been an explosion of experiments striving to fix this problem. One of the first successful experiments in recreating our primary bibliographic finding tool was the North Carolina State University Endeca-based catalog. In a partnership with Endeca, a company that before this project was mainly known for providing the software for catalog sites such as *L.L. Bean*, NCSU demonstrated the power of faceted browsing within a bibliographic context.[21]

Open source library catalog systems such as *Koha* (see <http://koha.org/>) and *Evergreen* (see <http://www.open-ils.org/>) are also both pushing the envelope for library bibliographic discovery tools and also enable anyone to see the code and make changes.

With the creation of a new text indexing tool built for faceted browsing called *Solr* (see <http://lucene.apache.org/solr/>), at least two library projects have created systems based on this platform. The *VUFind* project, by the William Falvey Memorial Library of Villanova University, has seen adoption by several libraries, including the National Library of Australia.[22] Another project constructed on the *Solr* platform is *Blacklight* from the University of Virginia (see <http://blacklightopac.org/>). Both of *VUFind* and Blacklight are finding tools only, which means libraries still need many of the functions of a typical integrated library system (ILS) as well.

Tools for the future

No matter what the future holds, we know that we face widespread, systemic changes in how we perform bibliographic description. Whether the particular strategies I describe above prove helpful or harmful, only time will tell. But it seems clear to me that there are some specific professional strategies that are likely to help us face an uncertain future with strength and resilience.

Relationships with others in the chain. Although libraries have never quite been alone in the creation of bibliographic data, one could argue that we have been isolated. Our professional standards are ours alone – no other profession has ever adopted MARC, for example, as its core bibliographic standard, or Z39.50 as its essential search protocol. Even the OpenURL standard, specifically written broadly in its 1.0 incarnation to encourage widespread adoption, has seen little or not take-up outside of libraries. This must

end.

Worldwide computer networks now make possible what before would have been problematic, if not impossible. We can now collaborate much easier with publishers, bibliographic data aggregators and vendors, and bookstores – all of whom share interest in bibliographic data. In this community libraries are a virtual backwater, where bibliographic data is often created from scratch using hand-tooled processes and arcane procedures. Meanwhile, publishers create a great deal of data before a book is even published – from basic bibliographic description to summaries, author biographies, and review quotes.

Establishing strong and mutually beneficial relationships with every player in the chain of bibliographic custody can only help libraries become both more efficient as well as more effective. But doing so may mean giving up some of our long-standing autonomy. It may increasingly make little sense, for example, for us to have our own bibliographic description format. If the ONIX format used by publishers can potentially fulfill our needs, why should we continue to use MARC? What could we gain by continued isolation? What opportunities will we miss if we remain in bibliographic isolation? These questions become tremendously compelling given the efforts to remake our bibliographic infrastructure through the *Resource Description and Access* (RDA) effort (see <http://www.rdaonline.org/>). We cannot afford to wait, and yet our professional direction appears set in on a completely different path.

Facility with different bibliographic formats. The conundrum of having multiple bibliographic formats (at least one used by those outside the library community; e.g., ONIX, and multiple formats within the library community; e.g., MARC, MODS) means that we must get very good at using them all. We are no longer a profession focused only on MARC – we cannot be. So we must get good at translating formats and understanding the weaknesses and strengths of each. We will need new tools for analyzing metadata, normalizing it, visually depicting it in useful ways, and making systemic changes.

Professional retooling. Long gone are the days when a librarian could exit graduate school complacent in the knowledge that she was fully trained for what she could expect to encounter during her career. A degree in librarianship should not be viewed as an endpoint, but a beginning – a foundation upon which someone builds over the long course of a career with constant learning and retooling.

Creating our bibliographic future today

There is no question that we face many challenges in a world where *Google* is digitizing, and providing full-text searching for, millions of books that libraries hold. The role of bibliographic description is forever changed in such a world. Discover is no longer limited to the descriptions that catalogers choose to create. Nonetheless, I continue to believe that structured bibliographic description will continue to be useful. But we need to do it very differently than we are doing it today, we need to work with others who have information and procedures to contribute, and we need to expose it in some very different places and in some very different ways.

The strategies described above are really only the tip of an iceberg that we have yet to fully understand. And our understanding will evolve with the environment, which means we must be content with always being at least one step behind where we need to be. But being more than one step behind means forfeiting our future to others who will create it for us. I prefer a different future.

Notes

[1] For example, University of California Libraries Bibliographic Services Task Force. *Final Report December 2005* <http://libraries.universityofcalifornia.edu/sopag/BSTF/FinalsansBiblio.pdf>; Calhoun, Karen, *The Changing Nature of the Catalog and its Integration with Other Discovery Tools*. February 21, 2006, <http://dspace.library.cornell.edu/bitstream/1813/2670/1/LC+64+report+draft2b.pdf>; Library of Congress Working Group on the Future of Bibliographic Control, *On the Record: Report of The Library of Congress Working Group on the Future of Bibliographic Control* (January 9, 2008), <http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf>.

[2] Tennant, Roy. "The Future of Descriptive Enrichment," blog post, <http://www.libraryjournal.com/blog/1090000309/post/1920018592.html>.

[3] Paraphrased in Stephens, Owen. "ALA 2008: There's no catalog like no catalog - the ultimate debate on the

future of the library catalog" blog post, <http://www.meanboyfriend.com/overdue_ideas/2008/06/ala-2008-theres-no-catalog-like-no-catalog—-the-ultimate-debate-onf-the-future-of-the-library-catalog.html>.

4 "About OCLC" web page, <http://www.oclc.org/us/en/about/> .

5 "Facts and Statistics" web page, <http://www.oclc.org/us/en/worldcat/statistics/>.

6 "ONIX for Books" we page, <http://www.editeur.org/onix.html>.

7 "Next Generation Cataloging" web page, <http://www.oclc.org/partnerships/material/nexgen/nextgencataloging.htm>.

8 "Renovating Descriptive and Organizing Practices" web page, <http://www.oclc.org/programs/ourwork/renovating/>.

9 "Bibliographic Enrichment Advisory Team" web page, <http://www.loc.gov/catdir/beat/>.

10 "Audience Level" web page, <http://www.oclc.org/research/projects/audience/>.

11 *Steve.museum* "Links and Resources" web page, <http://steve.museum/?option=com_content&task=blogsection&id=5&Itemid=14>.

12 Trant, J. "Tagging, Folksonomies and Art Museums: Early Experiments and Ongoing Research," *Journal of Digital Information*, v. 10, no. 1 (2009), p. 39, <http://journals.tdl.org/jodi/article/view/269/278>.

13 Ibid., p. 3.

14 *The Open Archives Initiative Protocol for Metadata Harvesting*, <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

15 TechEssence.info "Library Application Program Interfaces (APIs)" web page <http://techessence.info/apis/>.

16 OCLC Developer Network web site <http://worldcat.org/devnet/>.

17 Structured Dynamics "Linked Data FAQ" web page <http://structureddynamics.com/linked_data.html>.

18 Library of Congress "Authorities and Vocabularies" web site <http://id.loc.gov/>

19 Springer, Michelle, Beth Dulabahn, Phil Michel, et.al., *For the Common Good: The Library of Congress and the Flickr Pilot Project*, Washington, DC: Library of Congress, October 30, 2008, <http://www.loc.gov/rr/print/flickr_report_final.pdf>.

20 Schneider, Karen, "How OPACs Suck, Parts 1-3," *ALA TechSource Blog*, <http://www.techsource.ala.org/blog/2006/05/how-opacs-suck-part-3-the-big-picture.html> and University of California Libraries Bibliographic Services Task Force. *Final Report December 2005* <http://libraries.universityofcalifornia.edu/sopag/BSTF/FinalsansBiblio.pdf> are but two examples.

21 Antelman, K., Lynema, E., and Andrew Pace, "Toward a 21st Century Library Catalog," *Information Technology and Libraries*, 25(3): 2006, <http://eprints.rclis.org/7332/>.

22 VUFind web site, <http://vufind.org/>; "Open Source at the National Library of Australia Catalogue," web page <http://www.nla.gov.au/pub/gateways/issues/92/story02.html>.