Testimony Submitted to the Library of Congress
Working Group on the Future of Bibliographic Control

Part I
Organizational and Institutional Submissions

To: Library of Congress Working Group on the Future of Bibliographic Control
From: Cataloging and Classification Standing Committee, American Association of Law Libraries
August 7, 2007

The following testimony was written on behalf of the Cataloging and Classification Standing Committee of the American Association of Law Libraries (AALL). It does not represent any official position of AALL itself.

As the de facto national library for the United States, the Library of Congress has had tremendous influence over the development and use of bibliographic standards and structures used throughout the library community.

Furthermore, standards and procedures adopted by the Library of Congress, as the provider of cataloging to the Law Library of Congress, have significant impact upon the law library community. As we saw with last year's change in policy regarding the creation of series authority records and series headings, there was a cascade of decisions made by libraries everywhere, with some conforming to the new Library of Congress practice and others continuing to apply previous policy. For libraries continuing to control series entries (including the libraries participating in the Project for Cooperative Cataloging) there are significant challenges. A significant stream of records coming from the Library of Congress are now not standards-compliant and need intervention at the end of the bibliographic control chain, significantly raising expenses and lowering quality for everyone, and often resulting in duplicative work at many libraries.

1) Users and Uses of Bibliographic Data

Bibliographic metadata is inherently complex because of the myriad variations present in the materials being described. The current MARC standards have evolved over the years to accommodate new variations and formats. However, current ILS systems do a poor job of searching the data included in the bibliographic record and of making use of the inherent relationships between the various data elements. In other words, we need much better data-mining capabilities. It is not enough to rely simply on "word search" capabilities. We should not discard current practices of content designation just because our current systems don't do enough with the data. Systems come and go, but data is forever.

Users may only be interested in the Google search box to begin with, but probably want much better ways to filter the results to get at what they need. North Carolina State University's Endeca catalog suggests what could be accomplished by using facets to limit search results, but facets depend on the presence of accurate metadata in catalog records.

The "one size fits all" model of information retrieval does not work. There needs to be a variety of ways to search the data, and at different levels. Keyword searching might be adequate if you have very specific words to search. Even then, as has been pointed out by Thomas Mann, in a multilingual catalog, word searching can be problematic and

limiting. If we are waiting for the semantic web, we are going to be waiting for a long time.

Classification systems are the most underutilized mechanism we have for finding like materials in an ILS. The only thing most systems do is to provide a call number index that just puts the records in a shelf order. Users of catalogs have no idea what the numbers mean because there is no apparent underlying data structure, i.e. the MARC classification record, to support it. This could be extremely useful in the field of law. Wouldn't it be nice to search by terms in the classification hierarchy and retrieve results that way? Dr. Jolande Goldberg's conception for the KIA-KIX classification for the law of indigenous peoples is a step in this direction, i.e. the integration of web resources into the class schedule. The next step would be to integrate it into the catalog.

Better integration of catalog-based bibliographic data with external resources is needed, e.g. searching for catalogued and uncatalogued resources at the same time.

2) Structures and Standards for Bibliographic Data

MARC may be old, but it still works and can be translated to a variety of different machine formats. It isn't perfect, but it's the content designation that is important. Machines don't know a statute from a loose-leaf service.

We need better standards of interoperability between MARC records and other formats. We also need to separate the data itself from the presentation aspects of the data. Machines can do a lot if we tell them what we want them to do and exactly how to do it. We need to leverage the huge investment we have in ILS systems and bibliographic records to get these systems to do more.

3) Economics and Organization of Bibliographic Data

The greater library community faces ongoing economic challenges including:

Recruitment, training, and retention of staff and catalogers
The skills that are desirable in catalogers and cataloging staff are in demand not just in libraries but throughout the job market.

Shared cataloging and maintenance
Being at the end of the chain of bibliographic record creation and use, all actions that we take locally become that much more expensive when not shared in meaningful ways.

Dissimilar information systems
The information systems that we use to do our work are often difficult to use, proprietary, or non-standards-compliant. The tools that we use daily are often in different systems, implemented in different ways, meaning metadata and catalog records are created at differing levels and standards. This poses a significant ongoing expense throughout the entire cycle of record creation and maintenance.

Enrichment of cataloging data
Enrichment of cataloging data has many barriers. Many of these are economic, including staffing shortages and disparate systems which make integration expensive. Another barrier is the fact that enriched bibliographic records purchased from vendors are often, by contract, not allowed to be shared on a bibliographic utility. Other barriers, such as OCLC's enhance process, appear to be policy-driven.

Controlled access points
The creation and maintenance of controlled access points and their associated authority files are part of the essential core of what is considered cataloging. These are expensive operations because they require human intervention at many points, but significant value is added. Just because authority control is expensive does not mean it is not a worthwhile investment of time and effort. The collocation of materials under controlled access points enables users to more readily find related works. If this concept is extended to a "work," as it is in the FRBR model, then many variations and access points can be applied to a single work and controlled through a single record. It was also provide for more consistent application of access points.

Cost of maintenance
The cost of bibliographic maintenance is one inordinately placed upon local institutions. There is a lack of structure and standards for the shared updating and enhancement of records and their subsequent re-consumption and use by all cooperating institutions.

If the Library of Congress is to abandon or radically change its historic role in maintaining MARC21, the Library of Congress Classification System, Library of Congress Subject Headings, and other standards and practices, the impacts upon the greater library community have the potential to be monumental. These standards and practices, developed over the last century, would shift the greater library community from a unified environment to a system of scattered standards and practices, much like what is currently plaguing the Internet. We view the abandonment of established bibliographic structures and practices as a step backward, not forward.

Much attention is paid to the idea of libraries as **cost** centers – certain processes and policies are deemed too expensive and therefore "unsustainable." Perhaps this focus on cost has come about because it is much easier to measure cost than to measure **value**. We cannot know precisely what good has come about in the world because of libraries; we must satisfy ourselves with subjective quality assessment surveys as gauges of value. Like Lord Darlington's cynic, we risk becoming people who know the price of everything and the value of nothing.

We do not resent or fear change. If anything, libraries have been on the forefront of change, especially when it comes to technology. Catalogers have been implementing the "wisdom of crowds" since the advent of OCLC and shared cataloging. What concerns us is change justified solely by short-term gains in productivity or other limited goals.

Respectfully submitted by:
Jean M. Pajerek
Cornell Law Library
Ithaca, NY 14853
jmp8@cornell.edu
Chair, AALL Cataloging and Classification Standing Committee

The Peloponnesian War and the Future of Reference, Cataloging, and Scholarship in Research Libraries

By

Thomas Mann

Prepared for AFSCME 2910
The Library of Congress Professional Guild
representing over 1,600 professional employees
www.guild2910.org

June 13, 2007

Thomas Mann, Ph.D., a member of AFSCME 2910, is the author of *The Oxford Guide to Library Research*, third edition (Oxford and New York: Oxford University Press, 2005) and *Library Research Models* (Oxford U. Press, 1993).

The judgements made in this paper do not represent official views of the Library of Congress.

Abstract

The paper is an examination of the overall principles and practices of both reference service and cataloging operations in the promotion of scholarly research, pointing out important differences not just in *content* available onsite and offsite, but also among necessary *search techniques*. It specifies the differences between scholarship and quick information seeking, and examines the implications of those differences for the future of cataloging. It examines various proposals that the profession should concentrate its efforts on alternatives to cataloging: relevance ranking, tagging, under-the-hood programming, etc. The paper considers the need for, and requirements of, education of researchers; and it examines in detail many of the glaring disconnects between theory and practice in the library profession today. Finally, it provides an overview of the whole "shape of the elephant" of library services, within which cataloging is only one component.

What is involved in providing library service to the academic community? Is our purpose merely to provide "something quickly"? What, exactly, is wrong with promoting that end as our goal? What is the role of reference work? How does library cataloging fit into a larger scheme of necessary services? What *is* the larger scheme of which cataloging is only a part? What should research instruction classes strive to cover? What is a good outline for a basic research class? Does anything need to be explained at all if our "under the hood" programming and federated searching capabilities are adequate? In short, what idea of "the shape of the elephant" of research, and of library resources as a whole, do we wish to convey to an academic clientele?

Users of public and special libraries have different needs; my concern in this paper is the future of research libraries. Much of what the latter do, of course, spills over into public and special library practices.

A wide range of important issues and distinctions is involved here:

- **Differences in *content* available onsite and offsite**

  - copyright restrictions on what can and cannot be digitized
  - digitized sources restricted by site licenses or password use

- **Differences in *search methods* available onsite and offsite**

  - the variety of search methods, beyond keyword access (e.g, controlled vocabulary searching, citation searching, related record searching, browsing classified book stacks, use of published bibliographies), available onsite: their different retrieval capabilities

- **Differences between cataloging (conceptual categorization at scope-match level[1], vocabulary standardization within and across multiple languages, systematic linkage of categories) vs. relevance ranking of keywords, tagging, folksonomies, etc.**

  - the need for search methods enabling recognition of relevant sources whose characteristics (and keywords) cannot be specified in advance

- **Differences between scholarship and quick information seeking**

- relationships, interconnections, contexts, and integrations vs. isolated facts or snippets
- the need for successive, sequenced steps (with feedback loops) vs. "seamless one-stop shopping"

- **The problems of federated searching**

  - misrepresenting the full contents and search capabilities of individual databases
  - masking the existence of non-included sources

- **The inadequacy of the open Internet alone for scholarly research**

  - its inability to provide overviews of "the whole elephant"—i.e., not showing all relevant parts, not distinguishing important from tangential, not showing interconnections or relationships, not adequately allowing recognition of what cannot be specified

- **The need for education of users, not just improvements in "under the hood" algorithms**

  - education not just on how to use subject headings, but on how to do keyword searching itself
  - education on *multiple* search techniques other than keyword *or* subject-heading searching

- **The need for increased one-to-one connections with reference librarians, not just the digitizing of more material for direct full-text searching**

- **The disconnects between library theory and practice**

  - the assumption that library catalogs/portals should "seamlessly" cover "everything" to begin with
  - the assumption that library catalogs—or any other access mechanism—can operate efficiently without any prior instruction or point-of-use reference intervention
  - knee-jerk dismissals of enduring cataloging principles only because they originated in times of earlier technologies

- disregard of the importance of vocabulary control and cross-referencing because it cannot be accomplished by algorithms
- disregard of the significance of scope-match subject cataloging as the major solution to the problem of excessive irrelevant retrievals at the "granular" level
- disregard of the importance of shelving books in classified order, on the assumption that everything relevant can be identified online
- disregard of the extensive web of integral interconnections between LC subject headings and LC class numbers in providing access to book collections
- disregard of the increased utility of precoordinated strings of subject terms, and catalog browse displays of them

The problem with any discussion of such issues lies in the complexity of their interrelationships. It's like trying to pin down a warped piece of linoleum—flattening a bulge in one area immediately causes other bulges to pop up elsewhere. I cannot claim to have a system that flattens all the lumps, but I am concerned that many of the more important problems facing scholars are being ignored because a "digital library" paradigm puts blinders on our very ability to notice the problems in the first place.

I think the best way to clarify what I mean is to provide a concrete example, as a kind of central spine (I'm changing the metaphor) to which all of these issues are attached; I will discuss the various offshoot "ribs" as they arise in a real-world research situation. A major problem with much of the discussion in our profession these days is that many of us are indeed speaking from different paradigmatic frameworks. The only way to determine which is the better frame is to examine which one *works* best "at ground level"–i.e, which most readily enables the library profession to serve its scholarly clientele in ways that solve the full range of their problems.

Getting a researcher efficiently *from* what he or she asks for *to* what is available in a research library is a much more complex operation than most non-librarians realize; it is also more complex than too many library managers themselves seem to understand. Most of it cannot be done remotely through searching the open Internet, no matter how much under-the-hood programming underlies the utopian "single search box." As the following example will illustrate, the work involved also escapes description in quantifiable or measurable terms; but when it is done properly it nonetheless makes an enormous difference to the quality of the research that gets done. (It also justifies the expense of investing in costly resources that would otherwise be overlooked by most researchers, but which can indeed be brought efficiently to their attention.)

I am going to insist on differences between what I'll call "scholarship," on the one hand, vs. "quick information seeking" on the other. Obviously there is a spectrum of

continuities between the two–no one disputes that–but there are also big differences that are too often swept under the rug. Scholarship requires linkages, connections, contexts, and overviews of relationships; quick information seeking is largely satisfied by discrete information or facts without the need to also establish the contexts and relationships surrounding them. Scholarship is judged by the range, extent, and depth of elements it integrates into a whole; quick information seeking is largely judged by whether it provides a "right" answer or puts out an immediate informational "brush fire." Because of the range of elements involved, and the complexity of their integration, book formats are unusually important for scholarship (especially outside the hard sciences); more than any other medium, they allow an amplitude of coverage in ways that screen displays (especially of lengthy texts) make much more difficult to grasp.

For scholarly inquiries, the extent and depth of relationships *matter*–indeed, they are crucial to any judgment of the quality of the research product. Judging the result of a "quick information" search does not require an assessment of whether–or how successfully–it *integrates* the information discovered within larger expositions or narratives; the adequacy of an overall argument or survey does not arise in the same way it does in scholarly inquiries. There is a tendency in much current library literature to conflate "knowledge" and "understanding"–levels of learning that require interconnections to be made–with "information"; but they must be distinguished.

The example: Tribute payments in the Peloponnesian war

A graduate student came into the reading room where I work and asked, "Where are the books on ancient Greece?" It was evident this was a new user who was not familiar with closed stacks policy of the Library of Congress. I explained that particular books or other resources had to be identified through subject searches in the computer system (or other sources) and requested through call slips. Equally important, I turned this explanation of the stacks policy into a reference interview which elicited the fact that what the student *really* wanted was information on "the system of tribute payments among the Greek city-states during the Peloponnesian War."

The student said he had already done Google searches. Today, a search on "tribute" and "Peloponnesian" produces these results:

Google: 78,400 Web sites
Google Book Search [full texts of some digitized books]: 674 hits
Google Scholar [full texts of some digitized journals]: 2,030 hits

In each case, even months ago (when the retrievals were somewhat smaller), the student was overwhelmed with too much information: he "could not see the forest for the trees" or discern if he was finding the *best* relevant sources. A search on Wikipedia turned up

nothing right on the button, although it does have brief articles on th "Peloponnesian League" and "Peloponnesian War" that have the word "tribute" in them.

Most researchers–at any level, whether undergraduate or professional–who are moving into any new subject area experience the problem of the fabled Six Blind Men of India who were asked to describe an elephant: one grasped a leg and said "the elephant is like a tree"; one felt the side and said "the elephant is like a wall"; one grasped the tail and said "the elephant is like a rope"; and so on with the tusk ("like a spear"), the trunk ("a hose") and the ear ("a fan"). Each of them discovered something immediately, but none perceived either the *existence or the extent of the other important parts–or how they fit together*.

Finding "something quickly," in each case, proved to be seriously misleading to their overall comprehension of the subject.

In a very similar way, Google searching leaves remote scholars, outside the research library, in just the situation of the Blind Men of India: it hides the *existence* and the *extent* of relevant sources on most topics (by overlooking many relevant sources to begin with, and also by burying the good sources that it does find within massive and incomprehensible retrievals). It also does nothing to show the *interconnections* of the important parts (assuming that the important can be distinguished, to begin with, from the unimportant).

In this Peloponnesian case, my thinking was, first, to try to guide the student to an intelligible *overview* of the relevant literature, so that he could indeed see "the whole elephant," and not just "something" on the topic. This is the most important function a reference librarian can serve in a large research library.

My first thought was of encyclopedia articles (rather than whole books or journal articles) because their very purpose is to provide concise overviews of topics, with manageably small bibliographies of highly-recommended sources (rather than printouts of "everything"). So I started by searching an obscure subscription database, *Reference Universe*, which indexes all of the individual articles in over 12,000 reference sources; it is particularly good in its coverage of specialized subject encyclopedias. (As with so many subscription services, the title of the source does not begin to convey what it can do—even if the reader, working on his own, did come across this title in the Library's list of proprietary database subscriptions, he still would probably not have bothered to explore it.) The indexing in this file immediately identified an article o "Tribute lists (Athenian)" in a highly reliable source, *The Oxford Classical Dictionary*. This volume was right in the Main Reading Room reference collection; its article provided exactly the concise overview of the topic that the student wanted—without knowing how to ask for it, or even that it was *possible* to ask for a concise overview. The article also mentioned

at its end that "the standard work on the tribute records is B.D. Meritt, H.T. Wade-Gery, and M.F. McGregor, *The Athenian Tribute Lists*, 4 vols. (1939-53)."

Whenever there is a "standard work" on a topic, it is better to find this out sooner rather than later in the course of one's research (as many grad students–myself among them–have discovered "the hard way"). Armed with this information, I showed the reader how to search the computer catalog for that standard work. The LC cataloging record for the book then provided crucial information for the *next* step of the search–i.e., the record found through a known-item *title* search indicated that its most promising *subject category* is "Finance, public–Greece–Athens" (i.e., not "tribute" AND "Peloponnesian"). A search under this standardized LC subject heading retrieved a roster of directly relevant works whose keyword variations could never have been specified in advance:

Tribute Assessments in the Athenian Empire (1919)
Studies in the Athenian Tribute Lists (1926)
Treasurers of Athena (1932)
Athenian Financial Documents of the Fifth Century (1932)
Athenian Assessment of 425 B.C. (1934)
Documents on Athenian Tribute (1937)
Vorschlage zur Beschaffung von Geldmitteln, Oder, Uber die Staatseinkunft (1982)
Finances Publiques et Richesses Privees dans le Discours Athenian au Ve et IVe Siecles (1988)
Pathogene Syndroma sto Demosionomiko Systema tes Archais Athenas (1991)
Money, Expense, and Naval Power in Thucydides' History 1-5.24 (1993)
Money and the Corrosion of Power in Thucydides (2001)
Poroi: A New Translation / Xenophon (2003)

Advantages of controlled vocabulary use

Note several things about this retrieval:

A) Again, not one of these titles would have been retrieved by a keyword search on "tribute" combined with "Peloponnesian" (let alone "ancient Greece"–the words initially used by the researcher before I did the reference interview).

B) The works found through an LC subject heading search in the Library's catalog include both *current* and *older* works–from 1919 through 2003–together *in the same set* (not just recent, in-print works).

C) The works found through an LC subject heading search in the Library's catalog also include *both English and foreign language* sources–German, French, and

Greek–together *in the same set*, without the searcher having to specify any foreign language terms. (I should note that this subject heading was not the *only* one relevant to the topic.)

D) The retrieval was of manageable size, not overwhelming.

E) The works identified were actually owned by the Library, immediately accessible without the delays of borrowing or interlibrary loan. (The Principle of Least Effort needs to be kept in mind: because sources that are readily available are more attractive than those requiring greater time or effort to secure, we need to make high-quality sources as readily retrievable as possible–while we continue to operate in the real world, where paper-copy books are essential to scholarship because copyright and site-license restrictions will never vanish; nor is it likely that future scholars will readily *read* 300-page texts online. If our goal is to promote scholarship, then "least effort" on the researchers' part *means* "most effort" on *our* part, in our acquisition efforts, in creating high quality cataloging, in providing proactive reference service, and in assuring the long-term preservation of our material.)

F) Each of these books is *substantially* about the tribute payments–i.e., these are not just works that happen to have the keywords "tribute" and "Peloponnesian" somewhere near each other, as in the Google retrieval. They are essentially *whole books* on the desired topic, because cataloging works on the assumption of "scope-match" coverage–that is, the assigned LC headings strive to indicate the contents of *the book as a whole.* (Any single assigned heading may not, by itself, indicate the content of the entire work, but any heading will at least indicate the subject-content of a *substantial portion* of it. Scope-match cataloging aims to summarize the major overall content of a book, not its individual chapters or smaller subsections. It is the antithesis of "granular" level indexing, as provided by the book's index pages or by keywords from the entire text.) In focusing on these books immediately, there is no need to wade through hundreds of irrelevant sources that simply mention the desired keywords in passing, or in undesired contexts. The works retrieved under the LC subject heading are thus *structural parts* of "the elephant"–not insignificant toenails or individual hairs.

To change the metaphor for a moment, consider a mosaic picture of an elephant made up of thousands of small individual colored tiles. Keyword retrieval in a full-text database is like searching at the granular level for individual tiles; if you specify that you want all of the gray pieces (needed for the legs, sides, ears, tail) and all of the white pieces (tusks, teeth) they can indeed be retrieved together in one set. But searching at this level cannot retrieve *the image as a whole* with all of the parts properly interrelated; it cannot combine just some of the grays into legs or ears or tails, to the exclusion of other gray pieces that belong elsewhere. Nor can it exclude tiles from thousands of other entirely different pictures (rhinoceroses, skyscrapers, dirigibles), which are also retrieved because they happen to have gray and white pieces within their own makeup. *For these purposes you need the equivalent of "scope match" cataloging,*

which both defines what "the whole" object is to begin with and sets conceptual boundaries on what is or is not a legitimate part of that whole. Within these *scope boundaries* various keywords (from titles, contents, or full texts) are contextually relevant, but outside of them the same words become irrelevant "noise." Merely giving more weight to certain words tagged as metadata, so that they will be ranked by the software as more important within an overall keyword retrieval, will still not assemble an overall picture with any scope boundaries, or segregate structural from tangential elements within the picture, let alone separate the elements within the desired picture from the same elements appearing in entirely different pictures.

Pictures, of course, don't contain cross-references to other illustrations; so here the analogy breaks down. But controlled-vocabulary LC subject headings, unlike mosaic tiles or keywords, are indeed linked to broader, related, and narrower terms to establish a road map of relationships to other conceptual headings–a mapping frequently crucial to scholarly overviews that is *not provided at all by "ranked" metadata terms, or provided reliably by democratic tagging*. Moreover, this cross-reference network itself functions in a way that refers users to other headings that are themselves at scope-match (rather than granular) conceptual levels–a level that is also lost when precoordinated LCSH subject strings are decomposed into their individual "facet" elements.

The point needs emphasis: some theorists have a knee-jerk aversion to scope-match subject cataloging because they unthinkingly regard it as simply a carry-over from card catalog days. (Cards could not provide granular-level access without making catalogs much too physically large.) What they apparently lack is any experience in dealing with actual researchers, for whom *this level* of cataloging *solves* the otherwise intractable problem of retrieving so much chaff with keywords that the whole books they want become buried indistinguishably in huge retrievals–e.g., Google Book Search's 674 hits combining "tribute" and "Peloponnesian." Keyword searching at granular levels "overshoots the mark," as does faceted searching of LCSH elements that must be combined into wholes by searchers who barely know which keywords to enter in the first place, and who also often don't know what the "whole" *is* until they *recognize* it in a precoordinated string. (Would any searcher working entirely on his own know that "Finance, public" needs to be chosen to begin with, and then combined with "Greece" and "Athens"? As a reference librarian, I can say it is *much easier to teach how to find the precoordinated string* than to teach how to think up all of the individual facets that need to go into a Boolean combination.) Increasing the granularity of searching to keyword levels, *and* robbing LCSH "facets" of their conceptual contexts in precoordinated strings, are both practices that directly undermine the scope-match level of traditional indexing–but it is precisely this feature of cataloging that brings about the quick retrieval of the "elephant's" *structural* parts (the whole books on, or substantial treatments of, the topic). These are the books readers want to find first, unencumbered by the clutter of thousands of irrelevant hits having the right words in the wrong contexts, outside the desired conceptual boundaries.

Note that neither I nor anyone else is arguing against granular levels of access being provided in addition to scope-match; it is the replacement of one by the

other that is objectionable. We need both.

Scope-match cataloging hits the bull's eye at the level of retrieval most needed for distinguishing structural from ephemeral relevance to a topic. While it is true that the subject-content of a book (or other record) as a whole can indeed be indicated by a combination of individual index elements ("Finance" AND "public" AND "Greece" AND "Athens"), researchers have much more difficulty thinking up all of the terms that go into such combinations; it is much easier for them to simply *recognize* strings that have already been combined. ("Least effort" is a reality–again, it's easier for them on the retrieval end if we do more of the work on the input end.) Theorists who assert that simply "digitizing everything" eliminates the need for cataloging[2] evidently have minimal experience with the actual results produced by implementing their theory. Full-text searching is indeed extremely valuable in many situations; but if a researcher wishes to get an *overview* of the important works on a topic, that kind of searching is positively counterproductive–it cannot segregate whole books from fragments of books, nor can it separate substantial treatments from trivial. It buries high and low quality sources in huge sets without the discriminations that users need. Granular access precludes overview perspectives unless librarians also provide *alternative* search mechanisms that solve the problems *created* by granularity.

G) The problem of keyword variations (see the list, above, of titles retrieved) would not have been solved by "throwing more keywords into the hopper"–i.e., so that words which don't "hit" within titles (appearing on brief catalog records) can nonetheless be found because they do indeed "hit" within larger digitized full texts. In addition to erasing the necessary conceptual boundaries for determining the *relevance* of English-language hits (again, Google Book Search: *674* hits), the same keyword searches of English terms would fail to retrieve the relevant French, German, and Greek texts.

H) The catalog could assemble this group of highly-relevant resources, to begin with, because it makes direct use of the subject expertise of the professional catalogers who had previously brought about *conceptual categorization* of the relevant books in one grouping (under the standardized heading)–*and* done it at the level of the book as a whole–through *vocabulary control*. A retrieval system based on controlled conceptual categorization of sources is radically different from one that relies on *relevance ranking* of keywords done by machine algorithms. The latter can take the words specified by a researcher and change the display-order of the retrieved results according to various criteria for weighting the keywords; but such a system cannot find, to begin with, keywords other than those specified. (Claims for automated "query expansion" need to be examined skeptically; there is usually much "less there than meets the eye." Demonstrations–as with this Peloponnesian example–are called for, rather than mere assertions lacking concrete examples.) We all need to be very skeptical of the phrase "relevance ranking"–"term weighting" would be more accurate–because it radically changes the very meaning of the word *relevance*. It entirely divorces its definition from the notion of *conceptual* appropriateness, across both variant expressions

and variant languages, and from the notion of *substantial* (rather than tangential) appropriateness.

This point illustrates one of the major disconnects between theory and practice–or between competing paradigms–in our profession: some theorists dismiss the principle of vocabulary control (specifically LCSH) as outdated, apparently because it was developed under a technology (card catalogs) that *could* not provide granular-level access. The fact that thousands of professional catalogers created a system that *solves* the problems that today are *created* today by granularity, however, indicates concretely that the principles they developed (e.g., vocabulary control, scope-match indexing) are *not outdated simply because technologies have changed in the meantime*. Our professional forebears "created better than they knew"–or perhaps, more accurately, "better than many of us know today"–because the principles and practices they developed in the 20th century *provide the best solution to a major, and growing, problem of the 21st century.* If there is a problem of blinkered vision, it is not attributable to our predecessors; it lies with our own failure to recognize their genius, due to the constricting blinders of the digital library paradigm.

Additional search options beyond the catalog: *browsing classified shelves*

But there is much more to this Peloponnesian example. While the searcher was looking at the online catalog, I quickly inspected the reference collection's volumes for those that might be shelved adjacent to *The Oxford Classical Dictionary* (at DE5.O9 1996). Right nearby was another reference book: *Ancient Greece: Social and Historical Documents from Archaic Times to the Death of Socrates* (DF7.D55 1994); this contains full texts of relevant sources on the tribute payments, translated into English; and it also confirms that "the basic starting point for research on tribute" is same *Athenian Tribute Lists* work identified as "standard" by the *Oxford* source.

Additional search options beyond the catalog: *format searching for a literature review article*

While the researcher looked at this second reference book, I took yet another tack toward guiding him to an *overview* of "the shape of the elephant." At this point he had already gained an excellent sense of what are the most important *books* to start with (without the cluttering presence of hundreds of irrelevancies, as in Google Book Search); but I wished to get him to a similar overview, if possible, of the relevant journal articles. There is a mechanism for doing precisely this, which no general researcher has ever heard of. It is the *Web of Science* database, which indexes 9,000 of the highest-quality academic journals worldwide, in all subject areas–i.e., not just "science" areas, as its title seems to indicate. (This is another source that most humanities researchers would not bother to open, even if they saw it listed, without a reference librarian's intervention.) What I knew, in particular, was that *Web of Science* has a feature enabling searches to be

limited to "review" articles. These are not book reviews; rather, they are "state of the art" *literature review* articles written by knowledgeable scholars, to survey and summarize the entire literature of a topic, with extensive bibliographies–thus providing a more comprehensive and in-depth overview than that provided by encyclopedia articles. The *Web* database, searched initially by the Boolean combination "tribute AND Peloponnesian," and limited to the "review" document type, immediately turned up the following citation:

> **Title:** Athenian finance, 454-404 BC
> **Author(s):** Blamire A
> **Source:** HESPERIA 70 (1): 99-126 JAN-MAR 2001
> **Document Type:** Review
> **Language:** English
> **Cited References:** 105 Times Cited: 0
> **Abstract:** This paper presents a survey of Athenian financial history from the transfer of the Delian Treasury in, probably, 454 to the end of the Peloponnesian War some fifty years later, in the hope that future research will profit from an overview of the achievements of 20th-century scholarship.
> **KeyWords Plus:** PARTHENON; TREASURY; TRIBUTE
> **Addresses:** Blamire A (reprint author), 5 Caulfield Close, Bury St Edmonds, Suffolk 1P33 2LA England

Note that this "Document Type: Review" article has *105 footnotes*. This is the desired overview source for relevant journal articles. With this, along with the reference-book articles and the LC catalog retrieval, the reader was beginning to get a very good overview of the *whole* shape of the elephant rather than just a hodge-podge of "something" having the right keywords and retrieved quickly. (Note further that this citation also provides a mailing address for contacting the author–a regular feature of this database [and one that I anticipated] that is frequently valuable even apart from other considerations.)

All of the above steps were accomplished in less than fifteen minutes. It takes much more time to explain what is involved, and the reasons for doing one thing rather than another, than to just *do* it. (This, by the way, is the kind of "speedy" retrieval scholars really want, as opposed to another kind, discussed below [see **II**].)

Additional search options beyond the catalog: *related record searching*

There is still more: the citation retrieved by this *Web* database offered a clickable icon to "Find Related Records"; pursuing this link provided a list of other articles whose own footnotes *overlap* with the105 footnotes of the review article. Right near the top of this list (arranged in descending order by the number of overlapping footnotes) is the following reference:

Title: Epigraphic geography - The tribute quota fragments assigned to 421/0-415/4 BC
Author(s): Kallet L
Source: HESPERIA 73 (4): 465-496 OCT-DEC 2004
Document Type: Article
Language: English
Cited references: 43
* * *
E-mail addresses: kallet@mail.utexas.edu

This "related record" article (along with others) appears because it has *six footnotes in common* with the starting-point review article–i.e., related record searching identifies articles having *shared footnotes*. The important point here is that this latter article is indeed talking about tribute during the period of the Peloponnesian War (431-404 B.C.)–but nowhere does its citation or abstract contain the keyword "Peloponnesian." This directly-relevant source would have been missed entirely by a conventional keyword search; it was retrieved because it had shared footnotes *rather than* shared keywords with the starting-point source. (This citation, further, provided its author's *e-mail address*!)

Additional search options beyond the catalog: *citation searching and published bibliographies*

The same *Web* database also provided a means to do not just *keyword searches*, and not just *related record searches*, but also *citation searches*: in this case, I could quickly show the reader that it provides a list of twenty-nine scholarly articles (since 1997, the retrospective limit of LC's subscription) that *cite* "the standard work" by Meritt in their footnotes, as follow-up discussions of it.

Still more: while the reader was looking into the citation and related record search features that I brought to his attention, I also checked to see if there is a published subject bibliography on the topic, by searching *Bibliographic Index Plus* (yet another title not likely to draw any layperson's attention). This proprietary database turned up the same "Epigraphic geography" article already found (above), because it has forty-three footnotes in its bibliography. (Although the existence of this citation was not "new" information at this point, it is a good sign when more than one search avenue leads to the same source–just as the two reference books independently agreed in identifying "the standard work." Such *convergence* on the same sources is an excellent indication that one's literature review is not missing the most important material—i.e., that important parts of "the elephant" are not being overlooked.)

More again: at this point the reader essentially said "Enough for now!"–he wanted to start with that literature review article. But I informed him of many additional proprietary databases (not on the Internet) that could provide still more citations: *Digital Dissertations* (which immediately turns up a thesis that explicitly disagrees with "the standard work"), *Periodicals Index Online* (an index of 4,720 periodicals in multiple languages from 1665-1995) *, L'Anee Philologique* (the best index to classical studies journals) *, WilsonWeb* (including *Humanities Full Text, Humanities & Social Sciences Retrospective, Readers' Guide to Periodical Literature,* and *Readers' Guide Retrospective*). All of these sources provide scores of additional references to works that are "right on the button" in discussing the tribute payments—but the titles of these databases, too are such that most would not draw attention to their relevance to the Peloponnesian topic.

The need for multiple search techniques rather than one "seamless" search

Note that as a reference librarian I could bring to bear on this question a whole variety of different search techniques, of which most researchers are only dimly aware of (or not aware at all): I used not just *keyword searching*, but *subject category searching* (via LC's subject headings), *shelf-browsing* (via LC's classification system), *related record searching*, and *citation searching*. (I also did some rather sophisticated *Boolean combination searching*, with truncation symbols and parentheses, discussed below.) Further, as a librarian I thought in terms of *types of literature*–specialized encyclopedia articles, literature review articles, subject bibliographies–whose existence never even occurs to most non-librarians, who routinely think only in terms of *subject* searches rather than *format* searches. And, further, one of the reasons I sought out the *Web* database to begin with was that I knew it would also provide *people contact information*–i.e., the mail and e-mail addresses of scholars who have worked on the same topic.

The point here needs emphasis: a research library can provide not only a vast amount of *content* that is not on the open Internet; it can *also* provide *multiple different search techniques* that are usually much more efficient than "relevance ranked" and "more like this" Web searching. And most of these search techniques themselves are not available to offsite users who confine their searches to the open Internet.

Results such as those achieved in this example cannot be duplicated by a "single search box" Google-type inquiry, no matter how much relevance-ranking, query expansion, post-Boolean probabalistic connecting, federated searching, and under-the-hood programming it brings to bear on the specified keywords. We are doing a very serious disservice to our patrons–and to our own library science students–if we encourage them to believe that "everything" they need can be provided by a "seamless, one-stop" inquiry in a single blank search box.

Differences between scholarship and quick information seeking

The disservice consists in assuming that there are no differences between scholarship and quick information seeking, and, as a result, in failing to show patrons whole ranges of options that they would indeed pursue if they knew how to articulate their own desires in light of a better overview of available options. Scholars, especially, want more than they know how to ask for. Anyone who does reference interviews with them will find this to be true. These are the some of their major unarticulated concerns–the differences between scholarship and finding "something quickly":

I) Scholars seek, first and foremost, as clear and as extensive an *overview* of *all* relevant sources as they can achieve. They want to see "the shape of the elephant" of their topic–the full extent of its different important parts *and* how the parts fit together. Librarians who actually work directly with them can testify that they do in fact want this, *even if they don't articulate this desire explicitly in user surveys.* Unintegrated information may be adequate for those who just want "something" quickly; it is not adequate for scholarship.

II) Speed in cataloging is *not* the hallmark of quality service, especially if relevant books that are catalogued quickly at "minimal level" or in "batch processing" fail to show up within the conceptual categories *and* webs of cross-references that are defined by standard (and more time-consuming) cataloging practices. When the standardized category designations (i.e., LCSH headings) are lacking on minimal-level records, we are faced with having to deal with an utter wilderness of unpredictable keywords across multiple languages. *Systematic* retrievals, *integrations* of resources in conceptual categories, and *overviews* become impossible.

Indeed, researchers who merely want "something" *quickly* will not seek lengthy and complex *books* to begin with when much shorter sources (Web sites, articles) are easily available. Books are for those who do *not* want just fast information. The difference in clienteles needs to be kept in mind. Scholars pursuing in-depth information or knowledge need something *other* than speedy retrieval.

Patrons who call for "speedier cataloging operations" in user surveys have no idea that such requests are being *interpreted* by library managers as *also* calling for the elimination of the conceptual categorization mechanisms (vocabulary-controlled subject headings, cross-reference linkages, and classification numbers) that provide them with the *overviews*–at scope match conceptual levels–which they actually value much more than quick delivery of individual, isolated items. (Any scholar can ask him- or herself at this point: do I really want to publish something, which may be read widely by my peers, that completely overlooks many of the most important books that have already been done on my topic, just so that I can finish faster?) If survey questions spelled out the concealed trade-off, I strongly suspect they would produce markedly different views of the importance of using speed as "the gold standard of processing."[3]

Another problem with surveys is that they ask only for what the users "want" at a point where most users do not know the extent of options available to them; once a librarian shows them what they are missing, as in this Peloponnesian example, *they do indeed want a great deal more than they previously realized they could get*.

The more intellectual effort catalogers put into the system at the front end (in creating, defining the scope of, and linking [via cross-references and browse menus] conceptual categories), the less effort is required by researchers at the retrieval end, to achieve the overviews they want of "the shape of the elephant." Cataloging systems that dis-integrate the cataloging information do not in fact "make the data work harder"–they make the *users* work harder, and take more steps, to reconstruct on their own the range of necessary *relationships* whose existence they cannot anticipate, and which they could otherwise have simply *recognized*. (Note, however, that cataloging itself, while necessary, is not sufficient by itself to provide all of the overview perspective that scholars need. Cataloging has a niche to fill, which must be supplemented by a variety of other search mechanisms created by people other than library catalogers, as the Peloponnesian example demonstrates.)

III) Scholarship is necessarily iterative, proceeding in successive steps that change depending on feedback provided by previous steps; it cannot all be done simultaneously. Again, we need to get away from the advocacy of a single catalog (or Internet) search box providing "everything" in "seamless one-stop shopping." (In the movies, such delusional behavior is dealt with by a glass of cold water to the face, or a vigorous shaking; in the library field, I'm not sure what is required to bring us to our senses on this point.) The world of informational resources is much too complex to be dumbed down to this level. *There is much more to refining a search than simply typing more, or different, keywords into the same search box. Frequently an entirely different search technique is required*—browsing book stacks, talking to experts, using published bibliographies, using controlled vocabularies and browse displays rather than keywords, using "limit" options, doing citation or related-record searches, thinking in terms of reference formats rather than just subjects—many of which searches cannot be reduced to *any* "box" on any computer screen.

An experiential awareness of this fact signals another of the biggest disconnects in all of library science, between theorists who fantasize that "everything" can be retrieved through a single online search box, and practitioners who know that the real information universe is much too varied, too extensive, and too complex to be viewed all at once from *any* such single vantage point. No single window of access can possibly show the entire "shape of the elephant" in any scholarly field; indeed, it is the inadequacy of relying on any single vantage point that is the very point of the Six Blind Men fable.

IV) Scholars are especially concerned that they do not overlook sources that are unusually important, significant, or standard in their field of inquiry. It does not do them any good if standard works are included but buried indistinguishably within huge retrievals. (Meritt's *Athenian Tribute Lists*, for example, is indeed among the 674 hits retrieved by

Google Book Search–although its copyrighted full text is *not* digitized for online reading. But Google does not have the mechanisms available to reference librarians for *singling out* this work as the best starting point for research on the topic, amid all the chaff that gets retrieved at the same time. Neither, be it noted, does traditional cataloging single out this source as "the standard work"–which means, again, that cataloging is itself [like Google] only one avenue of access, among many others, to some [not all] resources, and that the *several other* search mechanisms are *also* important.)

V) Scholars do not wish to duplicate prior research unnecessarily or to have to "re-invent the wheel." This is just common sense; but it needs to be said, because simply finding "something quickly" does not even begin to solve *this* very serious problem. Indeed, if mechanisms that provide *only* "something quickly" *replace* (rather than supplement) those existing mechanisms (such as cataloging) that do provide systematic access, then the problem of scholars unnecessarily re-inventing the wheel will be enormously exacerbated rather than solved.

VI) Scholars wish to be aware of cross-disciplinary and cross-format connections relevant to their work. Even though they may not articulate this desire explicitly, they are eager to pursue such connections if the avenues for doing so are pointed out to them by people (reference librarians and curators) who have a greater knowledge of the existence of those avenues. And most of the problems of cross-disciplinary searching are not solved by simple federated searches of multiple databases, especially when such inquiries dumb down the search possibilities to only keyword access, and when such keyword searching itself is likely to bury important sources within huge masses of irrelevancies.

An exorbitant faith in federated searching is yet another of the major disconnects between theory and practice that plague our profession. Such searching does indeed serve a useful purpose in some situations–no one denies that–but it is not a panacea that eliminates the need for tailoring inquiries to the peculiar capabilities of individual databases. (See the further discussion below.)

VII) Scholars wish to find *current* books on a subject categorized with the *prior* books on the same subject, so that the newer works can be perceived in the context of the existing literature–not just in connection with the much smaller subset of titles that happen to be currently in print. (Quick information seekers who do wish to see only current books can usually re-order their search displays to "most recent first" without radical changes to the cataloging content that is necessary for more in-depth searching.) This is one of the main reasons that we subsidize research libraries through taxes and endowments that shield them from market forces of supply and demand–so that they can provide free access to works not currently in general demand, and which profit-seeking bookstores would readily discard. (Second-hand bookstores that have some of the out-of-print sources do not make them freely available any more than the in-print stores do.) No one denies that research libraries need to be fiscally prudent; but there is a big difference between being fiscally responsible vs. allowing business concerns to determine the very goals of the

library (e.g. “increasing market share” over “promoting scholarship”). The “profits” generated by the research libraries that make their holdings *freely* available to all comers accrue to the individual authors and researchers who make use of them, not to the “bottom line” (or “market share”) of the libraries themselves.

VIII) Advanced scholars also wish for similar categorization of English and foreign language books–i.e., they want subject-category searches to retrieve relevant materials in all languages together, so that a worldwide context of resources on their subject can be easily discerned. They do not wish to be straight-jacketed within retrieval systems that are good *only* for finding English-language sources. (Those who want sources in only one language can usually limit their searches to the language designation of their choice, again without destroying the additional capability [i.e., vocabulary control] of the system required for more extensive searching.)

IX) Scholars particularly appreciate mechanisms that enable them to *recognize* highly relevant sources *whose keywords they cannot think up in advance*, to enter into a blank search box. (Such mechanisms are provided by subject heading searches, shelf-browsing [i.e., using the LC classification system], citation searches, related record searches, and published bibliographies–not by uncontrolled keyword searching. Putting readers in contact with knowledgeable people also gives them a way to find information whose exact characteristics they have trouble articulating. Keyword searching has wonderful advantages of its own–again, no one denies that–but its very real weaknesses need to be counterbalanced by many other, and different, search capabilities.)

X) Although they are more cognizant of the need for diligence and persistence in research, and of the requirement to check multiples sources, and of the need to look beyond the “first screen” display of any retrievals, scholars also wish to avoid having to sort through huge lists or displays–from *any* source–in which relevant materials are buried within inadequately-sorted mountains of chaff having the right keywords in the wrong conceptual contexts. Even minimal experience with Google shows that its relevance-ranking software does not solve this problem; in fact, it *creates* the problem–which must then be solved by *other* search mechanisms.

One hopes that the Working Group on the Future of Bibliographic Control[4] will give serious attention to these concerns, because it is not enough to simply characterize the users of libraries’ resources as “consumers” and “managers” without a much better analysis of the peculiar needs of *scholarly* “consumers.” Indeed, among the “managers” today there are apparently many who believe that all, or even most, of the above difficulties can be overcome by a combination of (a) “digitizing everything” for full-text searching, which involves (b) increasing federated searching to that “all” databases can be searched simultaneously, and (c) relying on “under the hood” programming (with automatic relevance ranking), along with democratic tagging and folksonomy referrals, to provide adequate subject access to book collections—to the extent that controlled-

vocabulary cataloging can be eliminated in the library's catalog and classified shelving can be done away with in the bookstacks.[5]

In fact, however, it is not a solution to the problems of most scholars simply to give them more digitized full texts to search on the open Internet. Just putting more content online exacerbates rather than solves the problems of information overload if the mechanisms for *finding* that content are inadequate to sort, filter, categorize, organize, and display it.

Keyword search problems

Google-type retrievals will be especially disappointing, and off the mark, if the researcher types in the wrong keywords to begin with, or not enough of the right keywords. Uninstructed users routinely make such mistakes; but it is only reference librarians who are in a position to see how badly they've formulated most of their searches to begin with–it is when those searches fail, and the readers ask for help, that we can retrace the ground and find out what they actually typed in, in comparison to their actual goals as elicited by a reference interview. (User logs by themselves do not supply the latter information.) While it is often pointed out that readers don't know how to do subject searches via LC subject headings, *it is equally true that most researchers do not know how to do effective keyword searches either.* The very same objection leveled against the use of LC subject headings *also* applies to most keyword searches themselves. *Education is required all around.* (See below.)

The fact that LC headings are not used efficiently indicates that basic instruction is required–just as it is for efficient keyword searching–not that vocabulary control should be eliminated. The standardization of terms, and especially of subject strings at scope-match levels, with linkages of concepts through cross-references and browse displays, *solves* too many of the serious problems that are *created* by excessively-granular keyword searches in full-text databases to be cavalierly dismissed as no longer useful. The technologies have changed, but the principles of providing efficient access are still valid. And yet cataloging is indeed dismissed[6]–one can only conclude that those who do not recognize the solutions have, themselves, too little acquaintance with the serious problems scholars experience, which cry out for exactly the remedies that good cataloging provides.

Indeed, in this same "tribute in the Peloponnesian war" example, the results actually produced by Google's "single search box"–even in the separate Book and Scholar components of its site–are nothing short of a professional embarrassment compared to what a scholar can find when working with a skilled librarian, in conjunction with a real reference collection (shelved according to LC Classification), a good online catalog (using controlled LC Subject Headings), and an array of proprietary databases (not freely available to everyone on the Internet)–all backed up by an actual onsite

collection of book and journal volumes shelved in browsable order. With a combination of such onsite resources, a researcher can indeed be led to discern the overall "shape of the elephant" of the literature on his topic. In contrast, any direct search of huge full-text databases, with access only via keywords (regardless of how they are weighted) through a single search box, cannot even begin to show searchers "the shape" of the relevant literature, or the conceptual interrelationships of its various parts, or the relative importance of some parts over others.

Relevance ranking is not conceptual categorization

Term weighting–a.k.a. "relevance ranking"–of results is not at all the same as scope-match conceptual categorization via vocabulary control with cross-references to related categories (see **F, G** and **H** above). It improves, up to a point, the display of retrieved records having the specified keywords–that point being the first two screens and not much beyond–but it does nothing to retrieve, in the first place, alternative expressions for the same concept in either English or multiple foreign languages. Again, see the above list of related titles collocated under the LC subject heading "Finance, public–Greece–Athens," a cataloger-assigned term that does indeed round up widely variant phrases for the same idea.

Let's not sweep this issue under the rug: how many of these books would have been brought to a researcher's attention by term-weighted retrieval of the keywords "tribute" and "Peloponnesian"? A scholar in this area does not need merely *something*; he or she needs an *overview* of "what the library has" (in Cutter's words). And here we have yet another disconnect in our profession: the knee-jerk dismissal of Cutter's principles of cataloging overlooks the fact that scholars even in a "digital age" do need to know what *their home library* has, locally and easily available–rather than "everything anywhere"-because scholarship does indeed progress through a sequence of steps that start with the most readily available sources, and most scholarly books cannot be read online because of copyright restrictions.

Further, would term-weighting *segregate* these few *whole books* on the subject—the structural parts of "the elephant"–from hundreds of others that merely have the right keywords in irrelevant contexts? Answer: demonstrably "No." Look at the actual results. *Term-weighting does not set conceptual "boundaries" that define the extent of the desired context, outside of which the right words become "noise."* While mechanisms such as Google's PageRank system of counting links as "votes" of importance are useful, they (again) effectively change the very meaning of the word *relevance.* Re-arranging some of the right keywords in a particular order does nothing to *find* the many *conceptually* relevant works that are overlooked to begin with, or that have become buried within thousands of hits that are in fact *irrelevant* even though they share the specified keywords.

Limitations of tagging, and of breaking subject strings into separate facets

"Tag" terms (i.e., keywords added by users) can be useful. Good results can indeed be brought up, in many situations, when untrained people contribute their own indexing suggestions to catalog records; but results will be negligible in relating seldom-used books (those that don't attract many tags to begin with) to others on the same subject. Moreover, tagging by the general public in not an adequate *replacement* for vocabulary control (although it is indeed a good supplement, just as granular keyword searching is a good supplement to scope-match cataloging); numerous indexer-consistency studies have demonstrated repeatedly that untrained indexers attempting to come up with descriptive terms for a document agree in their choice of words only ten to twenty per cent of the time.[7]

To keep this discussion grounded in reality, let's look again at the Peloponnesian example, particularly at the variety of keywords other than "tribute" and "Peloponnesian" that would have to be specified to turn up the sources actually retrieved above: Assessment [singular], Assessments [plural], Athenian, Athena, Archais Athenas, Treasurers, Financial, Finances, Money, Expense, Power, Quota Fragments, Syndroma, Demosionomiko, Geldmittein, Staatseinkunst, Richesses, Fifth Century, Ve et IVe Siecles, 425 B.C., 421/0-415/4 BC, 454-404 BC, Thucydides, Poroi. Is it any wonder that untrained indexers do not arrive at the same keywords any more than authors themselves do?

Further, tagging by non-librarians is not as good as standard cataloging in revealing the extent of a subject's unanticipated aspects. For example, although this did not come up in the present Peloponnesian case, the LC subject heading "Finance, public–Greece–Athens" is actually part of a large catalog *browse display* that provides a greatly extended context of relationships–one that might well be relevant to other researchers with different questions in mind. A very small sampling of that catalog browse display includes the following:

Finance, public
    Search also subdivision Appropriations and expenditures under names of countries, cities, government agencies, institutions, etc.
    Narrower Terms:
        Budget
        Claims
        Customs administration
        [etc.]
Finance, public–Accounting
Finance, public–Accounting–Law and legislation–Pakistan–Punjab
Finance, public–Arab countries–Dictionaries, Arabic
Finance, public–Dictionaries

Finance, public–Europe–History
Finance, public–Germany–History
Finance, public–Great Britain–History
Finance, public–Greece–Athens
Finance, public–United States–History–1801-1861–Sources
Finance, public–United States–History–1801-1861–Speeches in Congress
Finance, public–Yugoslavia–History
Finance, public–Zimbabwe–Statistics

The "democratic" addition of multiple uncontrolled keywords to a record cannot provide an overview map of *relationships* like this that "surround" the subject of the book being tagged. Tagging addresses only the subject of book in hand–not the relationships of that subject itself to other "outside" or "surrounding" topics that may well be of interest if they are *recognizable* in a menu display. Another major shortcoming of democratic tagging is that it will not systematically provide links to all of the little-used and foreign-language books that research libraries have a responsibility to collect.

The shortcomings of tagging as a replacement for (rather than a supplement to) LCSH are particularly clear when we consider the contrasting advantages of precoordination of subject heading strings.

The continuing need for precoordination in Library of Congress Subject Headings

Why is the precoordination of LCSH strings highly desirable to maintain, in addition to our newer capacities to do post-coordinate combination of individual terms or facets? For several specific reasons:

First, precoordination of terms is necessary to convey the very meaning of many subjects; for example:

Motion pictures for women as a precoordinated string has a precise meaning that is not captured by the post-coordinate combination of (motion pictures AND women)
Violence in women is not the same as (violence AND women)
Women in development is not the same as (women AND development)
Women-alcoholics is not the same as (women AND alcoholics)
History–Philosophy is not the same as **Philosophy–History**

Tens of thousands of such phrase headings would lose their meaning if broken up into their component words. (Of course thesauri for various subject disciplines do not have similar precoordination; but those disciplines do not require coverage of all subject simultaneously and their relations to each other, which is the universal field which LCSH must cover.)

Second, breaking up subject heading strings into individual words or facets, to be re-combined post-coordinately, drastically undermines researchers ability to *recognize* relevant aspects of a topic that they could *not* combine because it never occurs to them that such aspects exist until they see them listed (e.g., Accounting, Arab countries, Dictionaries, Law and legislation, Sources, Statistics, etc.). Separate groupings of faceted elements do not make the data work harder; they make the researcher work harder to see relationships that are no longer presented for easy recognition.

Third, the precoordinated strings provides more focused conceptual contexts for the individual faceted elements, *without which the scope-match level of cataloging is lost*. Above all, it is the scope-match level of retrieval that is most necessary for a scholarly overview of the structural parts of "the elephant"–the *whole books* on the topic, not the ones that simply mention the desired topic. The retrieval becomes much more time-consuming and complicated if multiple individual terms have to be re-combined to achieve the scope-match level. Post-coordinate combinations to reach this level are all the more difficult to bring about if multiple different menus of terms (topical, geographic, chronological, form) have to be separately examined to see the array of terms that are available *for* the combinations.

Fourth, it beggars common sense to believe that the use of multiple separate menus of facets is easier to work with than a browse display of all of them arrayed in a single roster. Separating subdivisions from the topics they subdivide can readily lead to confusing irrelevancies, and to entirely overlooking combinations that ought to be made. For example, in the string "Finance, public–United States–History–1801-1861–Sources" the individual facets lose their necessary conceptual context if they are separated from each other. Combining the form subdivision with the topical heading alone will produce confusing irrelevancies; the geographic and chronological facets must *also* be included for the retrieval results to be on target. Providing strings of *interconnected* subdivisions for easy recognition in browse displays–coupled with an explanation from reference librarians of how the displays work–is much more effective, and more easily teachable, than requiring multiple pointing/clicking operations among entirely separate menus for geographic, topical, chronological, and form aspects. (Note: these comments do not apply exactly to the Endeca system[8], which does provide access to precoordinated subject headings, although not on the first screen of a retrieval. My concern here is more with the attitude expressed by Beacher Wiggins, the Director of Acquisitions and Bibliographic Access at the Library of Congress, which is LC's cataloging department; Wiggins has openly questioned the practice of continuing precoordination at all.[9] His views, of course, have unusual weight in determining LC cataloging policies. They are all the more puzzling because Wiggins presided over the Bicentennial Conference on Bibliographic Control for the New Millenium only a few years ago [2001], which conference specifically considered and rejected the idea of abandoning precoordination in favor of faceting.[10])

Fifth, the vertical browse displays of subject heading strings (as above) show the relationships not only of individual elements within any string, but also the relationships of *whole strings themselves to each other*, enabling researchers to recognize a wide variety of other aspects of their subject that are "outside" (but still related to) the subject defined by any single string. Moreover, these "surrounding" precoordinated strings are themselves at scope-match subject levels–i.e., they will not lead to excessively "granular" and irrelevant works having the right words in the wrong conceptual contexts; they, too, will lead efficiently to *whole books* on *their* subjects. .

Sixth, the entire (and crucial) cross-reference structure of LCSH is dependent on linkages already established between tens of thousands of precoordinated headings, for example:

Women–Psychology
RT Women–Mental health
NT Achievement motivation in women
Animus (Psychology)
Anxiety in women
Assertiveness in women
Body image in women
Cooperativeness in women
Helplessness (Psychology) in women
Leadership in women
Self-esteem in women
Self-perception in women

This entire *network of relationships*–the kind necessary for systematic and scholarly retrieval–would be lost if researchers could search **Wome**n AND **Psychology** only as individual "facet" terms. Without the network, researchers will be relegated to the condition of the Six Blind Men, enabled to grasp only isolated parts of "the elephant" without having any mechanism enabling them to perceive the connections of those parts to other structural elements of their subject.

Seventh, tens of thousands of precoordinated subject strings are formally linked to specific LC classification numbers. Since the subject strings themselves are at scope-match conceptual levels, *so too will be the classification areas to which they point.* That is, researchers who go to the designated subject classes in the book stacks will be browsing in *whole books* on the topic of interest–not merely in snippets of text having the right words in the wrong contexts.[11] Cataloging and classification, once again, provide a solution to the problem of overly-granular retrieval. In order to find which areas of the bookstacks to browse, however, researchers need the subject headings in the library catalog to serve as the index to the class scheme. But the linkage between a subject

heading and a classification number is usually dependent on the precoordination of multiple facets within the same string. For example, notice the specific linkages of the following precoordinated strings:

Greece–History–Peloponnesian War, 431-404 B.C.: **DF229-DF230**

Greece–History–19th century: **DF803**

Greece–History–Acarnanian Revolt, 1836: **DF823.6**

Greece–History–Civil War, 1944-1949: **DF849.5**

Such formal connections between LCSH and LC Classification (LCC) not only make browsing in large collections much more effective for researchers; the same linkages–already *formally established* between tens of thousands of precoordinated headings and class numbers–also make class number assignments themselves much easier for catalogers to do. (Note that thesauri in specific subject areas do not need to serve this extra purpose of indexing a classification scheme in addition to indexing documents directly. LCSH cannot be reduced to a conventional thesaurus because it has to do things that are beyond the latter's scope.) And yet the elaborate webs of relationships between LCSH and LCC that have been created over the course of a century, by thousands of extremely perceptive professional catalogers, are *not even noticed* by "digital library" theorists. When we show no awareness at all of the very *structure* of our research libraries, our profession is effectively encouraging bulls to run rampant through china shops.

Eighth, most of the standard subdivisions of LCSH terms are not recorded in the printed "red books" set of subject headings–the thousands of heading-subdivision combinations that have been created show up only on browse displays such as those above. Without these browse displays, there is no way to know in advance the array of combinations that are possible in a given subject area; naive researchers cannot specify beforehand even a fraction of combinations that have already been established. Without the vertical browse displays of the precoordinated headings arrayed in sequence, the catalog has lost most of its basic *vocabulary control.* Too many valid headings are not recorded at all in the red books because they follow pattern-rules without being individually listed. Without *systematic access* to those headings, too, the catalog does not have a *controlled* vocabulary–and systematic access in such cases is not provided either by the cross-reference structure or by outright guessing of which elements exist, as potential elements for postcoordinate combinations. Browse displays are an *integral* component of LCSH vocabulary control.

Yet another "disconnect" in our profession needs emphasis here: just as many theorists have a knee-jerk aversion to the goal of aiming at scope-match cataloging levels

(because the newer technologies make "granular" access easier to provide), many also have a tied-in aversion to browse-displays of precoordinated subject-heading strings, such as the above–for the same "reason," that providing them is regarded as merely a carry-over from card catalog conventions. Again, however, there is a huge gap between theory and practice; such theorists evidently lack the experience of seeing how many real research problems are *solved* by these subject strings, and menu/browse displays of them, in online catalogs. *The fact that precoordinated headings were developed under the technology of card formats does not mean that the rationale behind their creation is outdated or no longer important.* They do much more than merely "break up large files"; they *also* solve the different and *more important* problem of providing systematic overviews of "the whole scope/shape" of their subjects–and they do it by enabling researchers to *recognize* search possibilities that they could never have specified in advance, and which they cannot easily reconstruct from multiple separate menus of facets.

The computerized browse displays that show us these *overview maps* of the extent of a subject's aspects (as in the above examples) are one of the major breakthroughs in cataloging technology in the last generation–in the card catalog days, it was much more difficult to *see* the arrays of subject aspects. And yet while reference librarians and researchers use these maps to gain the best overview perspective on the "shape" of the book literature on their topics, too many digital library theorists fail even to notice their existence–or they dismiss them out of hand because the system that created them was developed under a non-computerized technology that *must* be regarded with contempt by anyone who wishes to maintain social standing in the digital library world. The issue of whether precoordinated strings *actually solve real retrieval problems better than the proposed alternatives* is swept under the rug, for motives of not wanting to appear "out of date" amid the cutting-edge technologists. Once again, however, our predecessors in the cataloging profession "created better than they knew"–they left us a solution to problems of 21st century information overload, the excessive granularity of which they could not have anticipated. And their solution works better for *scholarly book retrieval* than any that are based on relevance-ranking, faceting, or algorithmic manipulations that destroy indexing and cross-referencing at the whole-book/scope-match level of subject conceptualization. It is only the blinders of our own digital library paradigm that prevent us from seeing the much-needed existing solution that is staring us right in the face.

I find it very easy to teach the use of browse displays such as that above–once a good example is pointed out, students pick up on the "recognition" possibilities of the displayed subdivisions immediately. (I especially advise them to look for form subdivisions "Bibliography," "Encyclopedias," and "Sources.") But education is still required, no matter what display technologies we come up with. The only way to justify a lack of formal educational effort on our part is to change the very goal of service, away from the promotion of scholarship to, instead, the promotion of just finding "something

quickly"–i.e., *endorsing* research having the *lack* of perspective exemplified by the Six Blind Men of India.

The objection that maintaining precoordinated strings of LCSH terms is "too expensive" or "too cumbersome" to deal with the Internet is easily handled: *don't try to catalog the entire Internet in the first place.* Confine cataloging primarily (though not exclusively) to more manageable collections. (See points **i-v**, below.)

Limitations of folksonomies

Folksonomy lists of related sources, based on assemblages of democratically tagged results (as in LibraryThing[12]) are also desirable supplements but terrible substitutes for the retrievals brought about by controlled vocabularies. How many of the "Peloponnesian" books (in multiple languages, in and out of print) listed above under the LC heading would have been found in folksonomy lists derived from uncontrolled tags? Folksonomies do not *adequately* show the contexts and webs of relationships that scholarship requires–which linkages can be and are provided by professional catalogers who maintain the *controlled vocabulary* of the LC system. And let's not forget–as many seem to have done–that beyond the standardization of terms for individual subjects, vocabulary control also entails the maintenance of scope notes, cross-references, and browse displays (like that for "Finance, public" above) which explain and exhibit the conceptual connections *among the many related search terms that have not been applied to the book in hand*, but which, once brought to the searchers' attention, are often of equal or even greater interest in expanding their horizons. Subject headings show not just books in the same category, but also whole webs of other, different (but related) categories.

Notice especially that the boundaries and interrelationships among LC subject headings, and between headings and class numbers, are spelled out explicitly for examination, so that we can see for ourselves what is and is not being connected–quite unlike automated "query expansion" mechanisms that operate "under the hood" in "black boxes," leaving users without any possibility of understanding *what* has been expanded, how extensively (or how inadequately or naively), in what conceptual contexts, and in what languages.

While folksonomies have severe limitations and cannot replace conventional cataloging, they also offer real advantages that can supplement cataloging. Perhaps financial arrangements with LibraryThing (or other such operations) might be worked out in such a way that LC/OCLC catalog records for books would provide clickable links to LibraryThing records for the same works. In this way researchers could take advantage of that supplemental network of connections without losing the primary network created by professional librarians.

Problems with "seamless" federated searching

Another element entailed in the utopian vision of "seamless, one-stop shopping" is the naïve belief that education can indeed be replaced by federated searching—i.e., that the simple combination of multiple (even "all") databases together into a single search pool is enough, by itself, to make them "accessible." Here is yet another disconnect between theoreticians and actual researchers. The problem is that lumping together multiple databases, with different search softwares, different controlled or uncontrolled vocabularies, different field search capacities, and different limiting features, dumbs all of them down to a lowest common denominator of keyword searching. This again may be adequate for finding "something quickly"–the unacknowledged "default" goal of librarianship, according to many of the new theoreticians–but it is utterly inadequate for promoting scholarly research, with its very different requirements (**I** through **X** above).

To keep this discussion once more grounded in reality, let's continue with our Peloponnesian example. And let us assume that the online catalog of the Library of Congress could be included in a federated search with just two other titles: *Periodicals Index Online* (an index to 4,720 periodicals in 58 languages internationally from 1665 to 1995), and *Web of Science* (indexing 9,000 academic journals internationally). The online catalog offers subject headings lacking in the two subscription databases, and *PCI* and *Web* offer very different search and limiting features. Reducing all searches to "lowest common denominator" keyword inquiries is, in fact, likely to exacerbate rather than solve the problem of the Six Blind Men–it will lead researchers to think that the few keyword "hits" they immediately get represent everything that exists about "the elephant."

Specifically, searching the book catalog with "tribute" AND "Peloponnesian" would miss all of the variant titles retrieved under the LCSH heading "Finance, public–Greece–Athens." The same search in *Periodicals Index Online*–strictly a keyword index–would also miss most of what is available in that database, because many other keywords are necessary: "(Athens OR Athenian OR Athenian* OR Delian OR Peloponnesian OR Greek OR Greece) AND (tribute* OR financ* OR payment*)" would be only a start. If one truncates "Athen*" the results will include a great deal of chaff having the terms "Athenia," "Athenaeum," "Athen," "Athenagoras," and "Athenais." If, however, one does not truncate, the search would miss foreign-language articles with terms such as "athéniennes," "athénien," "Athéna ," "Athènes," "Atheniensium," and "athenischen." "Attischen" would be missed entirely. Similarly, the truncation of "tribute*" after the "e" would be sufficient to bring up English language singular and plural forms; but it would then miss the German forms "Tribut" and "Tributquotenlisten." And other citations having terms such as the many others listed above (Treasurers, Financial, , Syndroma, Demosionomiko, Geldmittein, Richesses, Ve et IVe Siecles, etc.) would also be overlooked. (This is why keyword searching itself, like controlled-vocabulary searching, requires some prior instruction.) Nonetheless, a "federated"

searcher would probably conclude that he or she had indeed "covered" both the LC catalog and *PCI*, no matter what he typed in.

The same student, by including *Web of Science* in the federated search, would also miss the wide variety of keywords within that database, too–but, equally important, the student would have no clue that this particular source, when searched singly, would enable him to do citation and related record searches, with the impressive results given above; nor would he realize that this file offers the capability to zero in immediately on literature review articles, which otherwise tend to be buried within much larger retrievals. Again, the searcher would probably assume that he had "covered" the database because it was "included" in the federated pool.

The *primary* niche for library cataloging: books

I would be the first to agree that the inexpensive indexing methods of term weighting, tagging, and folksonomy referrals–none of which requires expensive professional input–are entirely appropriate for dealing with most of the Internet's Web offerings. With billions of sites to be indexed, it is out of the question to think that traditional cataloging can be applied to all of them. No one in his right mind would say otherwise.

But there is a crucial distinction that is being swept under the rug: the difference between quick information seeking and scholarship. The latter, especially in all subject areas outside the hard sciences (but within them, too, in many cases), requires *books*. The book format, more than any Web site, can accommodate the lengthy attention spans needed to fully grasp the extent and interrelationships of arguments and evidence pertinent to highly complex issues. (Digitizing a full book has the undesired side effect of making it virtually unreadable as a whole.) It is no accident that the University of California's landmark "How Much Information?" study assumes that the average book is 300 pages long.[13] My own attempt to survey the extent of resources available in research libraries–to provide a map of "the whole elephant"–came out to this same length.[14] *The Oxford Guide to Library Research* could have been longer; but anything shorter would not have done justice to the complexity of the topic. (Nor can its scores of recommendations for researchers be reduced to improved algorithms behind a single search box. Apparently, however, there are people in our profession who, with their fixed idea of "one box" searching, actually believe that everything in a research library [both content and search techniques] *can* be found efficiently, with ease and precision, through one box. Moreover, some of the same theorists regard such a massive dumbing down of search capabilities as the very goal of "updating" one's skills "for the 21st century."[15])

The universe of books published every year is much smaller, and much more manageable, than the universe of Web sites; *this* is the "niche" of sources to which professional cataloging should be *primarily* devoted. Books also merit the extra work

involved in cataloging and classification because of their greater importance to scholarship, and because of their long-term preservability. Most of the billions of Web sites do not merit this level of attention to begin with; they are too inconsequential and too ephemeral.[16] If we are going to promote scholarship, it is not enough to simply digitize the books for immediate retrieval if term weighting of keywords, tagging, and folksonomy referrals are the only mechanisms we provide for finding them. It is not at all unrealistic to propose that research libraries fill *the niche of providing the best, most systematic, access to books*–the alternative avenues of access (i.e., other than professional cataloging) may indeed be adequate for finding "something quickly" on the Internet, but they are not adequate for showing "the shape of the elephant" of relevant *book literature* on a topic. (N.B.: I would not confine cataloging *exclusively* to books; see below.)

We need to be clear about what is at stake here. The undeniable fact that there are too many Internet sites to be controlled by traditional cataloging leads some theorists to leap to the conclusion that *therefore* the library profession should abandon traditional (and expensive) cataloging entirely, even for books, and rely instead on inexpensive automated algorithms and tags/folksonomies supplied by others, which can be applied to greater volumes of material at less expense. A better solution is available, however; but it is necessarily more complex. It is the "niche" strategy that is dismissed out of hand by the Calhoun Report; it may be schematized as follows:

i) Do not attempt in the first place to control all of the Internet by means of traditional cataloging and classification; accept the obvious fact that this is impossible.

ii) Abandon the goal of having library catalogs provide "one stop, seamless access to everything." Confine cataloging and classification to a more limited *niche*, that of providing systematic access *primarily* to the library's own book collections—not to the entire Internet. Do not, however, limit cataloging solely to books; also catalog selected, high quality Web sites so that they show up in the same categories as the books, under the same headings, at scope-match levels, and in the same networks and webs of relationships defined by LCSH. In this way, users will be enabled to discover both books and quality Web sites (or other formats [e.g., maps, motion pictures, etc.] deemed worth the expense of cataloging) all in the same search–with the full recognition that vast amounts of other resources (individual journal and newspaper articles, individual manuscripts, *most* Web sites, etc.) will *not* be retrieved in the same search, even with federated searching.

iii) Rely on the abundance of sources created outside libraries, such as Internet search engines and commercial databases, to provide access to *all* of the other resources that lie beyond the niche of the library catalog's coverage. (Published bibliographies and carefully assembled reference collections, and browsable book stacks, in addition to Web sites, search engines, and subscription databases, must also be relied on.)

iv) Recognize that cataloging is itself only one function of research libraries, and that abandoning the coverage of "everything" *through the catalog* simply means that *other* parts of the total library system must step in to provide the additional access that is needed. Specifically, *reference service rather than cataloging* must steer researchers to the hundreds of subscription databases (with idiosyncratic vocabulary and limit options), thousands of hidden Web sites not visible to conventional search engines, tens of thousands of reference sources, dozens of unanticipated literature formats, and untold people-contact sources that patrons would miss entirely if they relied only on "one stop" computer searches. (Can any catalog search, now or in the future, duplicate the results in the Peloponnesian example?)

v) Recognize that no matter what we do in mounting and maintaining access systems of *any* kind, most researchers who work on their own without prior education or point-of-use instruction will *still* routinely miss *most* of what is available to them, without realizing they have missed anything. They will not see "the shape of the elephant" on their own. There is no circumventing the fact that high quality research requires *education and instruction*; this can only be supplemented and never replaced by better under-the-hood programming. The goal of providing free access to everything, from anywhere (outside library walls), at any time, by anyone, without any professional cataloging of important sources, and without reference intervention or education from librarians, is not only impossible, it is positively damaging to scholarship: it creates the false impression that researchers never need the kind of overviews provided in the Peloponnesian example, and that all of the requirements of scholarship (**I** through **X**, above) are no longer worth bothering about or worth striving to provide. It encourages potential scholars to believe that whatever few fragmentary parts of the elephant they happen to touch on their own *constitute* the whole animal. Should our profession continue to move in this direction, we will effectively be propagating exactly the kind of ignorance exemplified by the Six Blind Men.

We cannot continue to let the new technologies set their own agenda of what needs to be done, especially when that agenda calls for "lowest common denominator" and "one search box/one size fits all" searching that positively undermines the requirements of scholarly research. All of us–particularly the younger members of our profession–need to aim for goals higher than this. We have to remember cataloging *principles* that are still vital to efficient knowledge organization–even though they may have been first used under now-outdated *technologies*. Too many of us are failing to do the critical thinking needed to disentangle the principles from the technologies. The former still solve real problems today–problems of information overload, of haphazard and non-systematic retrieval, of inability to grasp "the elephant" as a whole–problems that are greatly *exacerbated* by the *lack* of traditional cataloging and by the *inadequacies* of the new technologies.

The need for education

If our profession does prudently abandon the unrealistic goal of providing access to "everything" through "seamless one-stop shopping via a single search box," it should by no means abandon the ultimate goal of providing efficient access to "everything" *via different and more realistic methods*. Since we cannot rely on computer algorithms to replace human intelligence, since we must assume that neither copyright nor licensing restrictions will ever vanish (thereby allowing free digital access to "everything"), and, further, since we ought to aim for a goal of promoting systematic scholarship rather than merely providing "something quickly," then–in the absence of single search box that will bring about utopia–we need to provide education (classes, publications, and point of use instruction) as an integral part of our overall professional program. Since we cannot make the complex, extensive, idiosyncratic, multi-lingual, and multi-format universe of knowledge records give up all of its secrets to "under the hood" programming, we must therefore *teach* what our algorithms cannot show automatically.

As I said above, I am convinced after 30 years of reference work that most users at all levels—undergraduates through full professors–will, if left only to their own devices, miss most of what *any* access system can deliver, most of the time. For example–over and above the specific "Peloponnesian" case discussed so far–I have on many occasions shown to historians and biographers who have already published books the existence of the databases *Historical* Abstracts and *America: History and Life*–the two basic databases in the field–with which they had no prior familiarity. In all such cases, they are *delighted* to have these resources brought to their attention–and are often dismayed that they did not find them sooner. The same patrons are, usually, equally ignorant of browse displays in library catalogs–and are equally delighted to be introduced to their use, to see how books "surrounding" to their topic are discoverable much more efficiently than they had realized. The same researchers never know how to limit computer searches efficiently by time periods (i.e., subject periods, not dates of publication) or geographic areas (subject areas, not places of publication). Back to the "Peloponnesian" example, most researchers are equally ignorant of the ways to zero in on "standard" works, encyclopedia articles, literature reviews, and subject bibliographies–or of the possibilities of doing citation or related record searches. Nor do they know how easy it is to find knowledgeable people, outside their own circle of acquaintances, to talk to about their topics. Nor do they have any idea of the range of disparate databases that provide coverage of their subject areas. They usually don't know how to do efficient keyword searching: specifically, not only do they not understand the differences between keywords and controlled vocabulary subject-category headings, they also don't know about truncation, nested Boolean combinations, word proximity searches, or use of quotation marks for phrase searching. Nor do they grasp the differences between term-weighting ("relevance ranking") and conceptual categorization. In all cases they greatly appreciate being shown–by reference librarians–both content and search techniques that they knew nothing about beforehand.

None of these very real problems will be solved simply by improvements in federated searching and under-the-hood programming. As library professionals we truly need to think outside the box of the Internet.

The range of files and search techniques available–and the differences among them–as well as the solutions to persistent search problems provided by quality cataloging, all need to be *taught* or *demonstrated* to researchers.

The reason that federated searching and under-the-hood programming are not panaceas is that scholars can never determine what they are *not* getting when their searches are handled by "black box" operations whose workings are not transparent. They are prevented from seeing how the "the shape of the elephant" is being determined. In *The Wizard of Oz*, Dorothy gets to the truth only when she disregards the advice to "pay no attention to the man behind the curtain." We should all have a similar mistrust of Great Floating Heads who tell us that we, too, need not look either "behind the curtain" *or* "under the hood"–that all our problems are being solved for us automatically by the higher authority of a Great and Powerful computer algorithm. At the very least, such wizards should demonstrate–not assert, but demonstrate–what their systems do on actual problems such as "tribute in the Peloponnesian war," for which considerably more than "something" retrieved quickly is required. Real questions such as these might serve as additional test cases:

"I am interested in the gods of the Mayas–what do you have on that?"

"What do you have on the foreign policy of Millard Fillmore?"

"What can I find on the Bay of Pigs invasion?"

"What can I find on the history of Yugoslavia?"

"What is available on landscape architecture?"

One especially hopes that the Working Group on the Future of Bibliographic Control will *test* its recommendations by their success in dealing with such real-world inquiries. Indeed, the group might *start* by examining how, specifically, its proposals would deal with the "Peloponnesian tribute" question.

Since we cannot rely on term-weighting/relevance-ranking, democratic tagging, folksonomy referrals, or federated searching to solve the problems of scholarship, and since most students are just as ignorant of how to do efficient keyword searches as they are of how to use LC subject headings, it is reasonable to conclude that a minimum of education must be imparted to them, no matter what content and software we offer in our online systems. But what, specifically, should our educational programs cover?

What should we teach in research instruction classes?

The Association of College and Research Libraries has proposed a set of five standards for information literacy, with, under each, a host of specific Performance Indicators and desired Outcomes for measuring successful implementation.[17] The document, however, is rather diffuse in terms of explaining what, specifically, needs to be taught–as it must necessarily be, given that it wishes to cover a very wide range of desirable outcomes.

I would like to propose a narrower, and more teachable, specification of topics to be covered in Research Orientation classes. The ACRL goals could not, I suspect, be covered in less than a semester. What I am proposing can be covered in one or two classes. My emphasis is on conveying to students–some of whom I hope will become scholars–the range of search options available to them within research libraries, which are not freely available from anywhere, at any time, by anyone, on the open Internet. In other words, I am offering an outline that portrays libraries as essentially *alternatives* to the Internet, rather than as "information reserves" that Google or Open Content Alliance "just hasn't gotten around to digitizing yet."

My experience with the outline is that it does indeed work best with graduate students and professionals who are engaged in doing substantive research. I say that because, truth be told, I've sometimes had experiences with undergraduate classes in which no one took any notes at all until I gave everyone my e-mail address. I believe it is simply a truism that the more experience anyone brings to a research class, the more he or she will get out of it–i.e., those who have never experienced the real problems that researchers run into will not recognize the importance of the solutions being offered, while those who do have the experience will sometimes almost literally slap their heads with the reaction, "Oh, there's a way to *do* that–I wish I'd known this before."

The scheme I propose is structured around different *methods of searching* that are applicable in any subject area. The overall point is that each has peculiar strengths, but also weaknesses and blind spots–no one search technique will enable a researcher to see "the whole elephant." (I make explicit use of the Six Blind Men fable.)

Of particular importance is that this outline situates library cataloging and classification within a larger context of other avenues of access to resources, that research libraries must also provide. It is an attempt to provide a larger intellectual framework for the whole profession–"the shape of the *whole* elephant," of which cataloging and classification are the legs and the tusks.

I will present the outline first, then add some comments on its noteworthy limitations, and on possible modifications of it. A fuller discussion of the individual

sources mentioned may be found on the Web site of the Library of Congress at < http://www.loc.gov/rr/main/research/ >, although the ordering of elements is slightly different there.

Basic Approaches for Subject Access

Initial Overview Sources: *Reference Universe* (encyclopedia articles) and *Web of Science* (review articles) databases

1) Controlled Vocabulary Subject Heading Searches

a) *Library of Congress Subject Headings*—multi-volume annual "red books" set–for finding *books*
b) Look for most specific or tightest fit–not general–headings
c) Four ways to find best terms:
  i) Narrower Term (NT) and Related Term (RT) cross-references in red books
  ii) alphabetically adjacent narrower/related terms in red books
  iii) subject tracings on catalog records
  iv) browse displays showing arrays of subdivisions not recorded in red books

2) Keyword Searches

a) Often more precise, but big trade-offs: loss of synonyms and variant phrases, hits in wrong contexts, blindness to foreign language sources
b) Relevance ranking/term weighting is not the same as conceptual categorization

3) Citation Searches

a) Will tell you if any starting-point source has been cited by subsequent journal articles
  Arts & Humanities Citation Index
  Social Sciences Citation Index — All three in *Web of Science*
  Science Citation Index
b) Advantage: circumvents vocabulary problems
c) other databases providing citation search capabilities

4) Related Record Searches (*Web of Science* Web database)

a) Will tell you which articles have footnotes in common with starting-point article
b) Advantage: Another way to circumvent keyword synonyms problem

5) Searches through Published Bibliographies

a) Different from computer printouts
b) Forms to use in online catalog:
[Subject heading]–Bibliography
[Subject heading]–[Geographic or Topical subdivision]–Bibliography
c) *Bibliographic Index Plus* (1982-); paper set (1937-)

6) Using People Sources

a) *Encyclopedia of Associations, Washington Information Directory*, etc.
b) authors of relevant articles
c) Internet contacts

7) Systematic Browsing Using Subject-Classified Bookstacks

a) Depth of access to full-text information; enables recognition without prior specification
b) Scattering likely: find *LCSH* heading(s) in catalog first (Note: Bibliographies are in Z)
c) Online catalog allows searches of catalog records (not full texts) by classification number

8) Computer Searches (truncation, Boolean combinations of terms, proximity searches, limits)

a) Online library catalogs
b) Online Subscription Services - licensed Web Sites
- Cannot be tapped into freely from anywhere, at anytime, by anyone
c) Internet search engines

I usually preface and conclude the entire presentation with advice to the effect that "If you remember nothing else, remember to talk to the reference librarians–if you work entirely on your own you will probably miss more than you find, and you won't know that you've missed anything. It's not only okay to ask questions; in a large research library, it's necessary."

It is immediately obvious that this is not a discussion of "how to think critically about Web sites." A major purpose of the talk is to wean students away from the open Internet by showing them the amazing resources available in research libraries–i.e., to present libraries as *preferable alternatives* to the Web when the goal is scholarship rather than quick information seeking. (If the research orientation is a whole semester course rather than a "one shot" talk, then of course the Internet would have to be discussed in detail.) Professors routinely lament that their students use only the Net; we librarians are

only exacerbating the problem if our own instructional efforts tacitly confirm the students' predispositions by ignoring the sources available only via libraries. Moreover, one learns to do critical thinking primarily by writing papers–clarifying nebulous thoughts by putting them into specific words, grammatical sentences, and coherent paragraphs–and having the results criticized; one does not learn it by passively listening to lectures. In other words, the professors themselves bear the responsibility to teach critical thinking skills, in ways that "one shot" talks by librarians cannot effectively address. (For the same reason, I do not think it is the job of librarians [in such lecture situations] to use our limited time to discuss style manuals or formats for footnoting. We need to concentrate of telling people how to *find* the information they need; how skillfully they read the sources and write them up is a matter for their professors to judge.)

It will also be obvious to experienced teachers that this "methods of searching" scheme avoids focusing on any particular subject area (Anthropology, Literature, Nursing, Psychology, etc.), because all of the search methods potentially work in *any* subject. That's one of the major strengths of this outline–it provides numerous "fall back" alternatives if one's first or second search attempts don't produce good results. I am not saying that all eight methods need to be employed on any given inquiry–as in the Peloponnesian example–but students who grasp only these few alternatives will be able to get farther into a topic, and will also be able to ask better questions in the first place, than those who are left at the stage of simply typing keywords into a blank search box, no matter how they evaluate the results. Presentations geared to audiences in particular subjects areas, however, should obviously concentrate on examples of research questions within the disciplinary area of concern.

The scheme also avoids discussion of most of the conventional types of reference literature (Almanacs, Atlases, Directories, Chronologies, Concordances, Dictionaries, Gazetteers, etc.) that form the structure of many traditional research classes. This overall "type of literature" framework does not show *enough* of "the whole elephant." I regard searching by such formats to be a "ninth" method, which I usually omit because any discussion of a dozen or so such types, in addition to the eight search techniques already given, is a sure way to make eyes glaze over. I also think that learning research via types of literature is something that requires a whole semester, and actual practice–it just doesn't "take" well without extended *experience* in working with the various formats. But other instructors may wish to include this search method, or even substitute it in place of some of the other eight.

The scheme starts with a discussion of two particular databases, *Reference Universe* and *Web of Science*, because of their utility in zeroing in an overview encyclopedia and literature review articles. (The importance of these has already been demonstrated in the Peloponnesian example.) It is true, however, that some academic libraries may not have subscriptions to either file; but in their absence some discussion of

alternative ways to find such "overview" sources, right at the beginning of a research project, is highly desirable.[18]

The eighth element in the list, "Computer Searches," allows for considerable wiggle room. One important overall point is that computer searching makes use of many of the elements discussed previously–controlled vocabulary subject headings, keywords, citations, etc.–but also enables them to be combined and limited in a variety of ways. The discussion at this point could go in either of two ways: one would be to exemplify search features such as truncation, Boolean operators, word proximity searching, quotation marks for phrase specification, and limiting options (by language, date, document type, etc.). The other would be to discuss the coverage of the more important subscription databases–Wilson, EBSCO, ProQuest, Factiva, LexisNexis, FirstSearch, and others not on the open Internet–that are accessible locally to the students, within the library walls or via their I.D. passwords. (I opt for an overview of particularly useful individual databases. Unless one has a great deal of time–pr more than one class session–I think the complexities of truncation, Boolean searching, *et al*., are best explained by reference librarians at the point of use, "over the student's shoulder.")

My colleagues and I have been offering such "Research Orientation" classes, in sessions of (usually) an hour and a half, every week for over a dozen years. It is more than noteworthy that, in the feedback sheets we get from the attendees, the one thing that they have told us most frequently, most explicitly, and most heartily is, essentially, "thank you for explaining how the subject headings work." (One attendee recently told me, regarding the subject headings explanation, "Research that took me two weeks before, I can now do in two minutes.") Admittedly the people who attend the talks given at the Library of Congress are a self-selecting group of researchers who actually want to use the Library's resources–they are people who already know that the Internet will not provide everything they need to find. They wouldn't be in the class to begin with if the Net were solving all their problems. It is an audience that is more scholarly to begin with than any class of undergraduates who are there because they've been assigned to attend. But that's also why we get such encouraging feedback from them–they do indeed have experience of the problems of substantive research, and they recognize solutions to those problems when they are presented with them.

Conclusion

The essayist William Hazlitt once wrote:

> The most trifling objects … assume the vividness, the delicacy, and importance of insects seen through a magnifying glass…. Ask the sum-total of the value of human life and we are puzzled with the length of the account and the multiplicity of items in it: take any one of them apart, and it is wonderful what matter of reflection will be found in it!

("The Letter-Bell," 1830)

A single reference question on "tribute payments in the Peloponnesian War" may indeed be trifling in the grand scheme of things, but when we take it apart and look at its implications for the future of both scholarship and librarianship, it takes on quite a bit more significance. Any such open-ended inquiry in the world of scholarly research is fraught with similar wide-ranging implications on what are the goals we librarians ought to aim for, and on what range of mechanisms we need to create ourselves, or provide from other sources, for attaining those goals. We need to make the best possible use of our principles, our experience, our tested practices, and our technologies, and not yield to the temptations to let either the technologies themselves or transient fashions constrict our vision of what needs to be done to promote scholarship of the highest possible quality–and that is a goal very different from striving to provide "something quickly."

———

Notes

[1] "Scope match" is the term used by Francis Miksa to describe the the level of specificity aimed at in traditional subject cataloging; see his *The Subject in the Dictionary Catalog from Cutter to the Present* (Chicago: American Library Association, 1983). The term refers to the practice of Library of Congress catalogers to sum up the content of a book (or other record) *as a whole* in assigning subject headings. (In other words, subject cataloging did not aim to indicate the content of individual chapters within a book, or to bring to researchers' attention the level of detail found in the book's index.) Indicating the subject of the book as a whole, if it could not be done by a single subject term, could be accomplished by providing as few separate headings as possible that, in combination, covered the whole scope of the book (e.g., Finance, public; United States; History; Sources; etc.); or it could be brought about by creating precoordinated subject headings whose subdivisions, in combination, indicated the content of the book as whole in a single string (e.g., Finance, public–United States–History–1801-1861–Sources). See the ensuing discussion.

[2] The Library of Congress is attempting eliminate its costly subject cataloging operations at the "scope match" level in exchange for digitizing more full texts at the granular level of keyword retrieval. "[U]sers increasingly want the content itself not a cataloging record"–Deanna Marcum, Associate Librarian for Library Services, in her testimony to the House Appropriations Committee, March 20, 2007. It is characteristic of Marcum to portray the digitization of full texts vs. cataloging as a zero sum game in which one can be done only at the expense of the other, rather than as complementary avenues of access that are both desirable. See Marcum's other,

similar statements, and a review of the "Calhoun Report," commissioned and endorsed by her, in the several discussion papers at < www.guild2910.org > (accessed May 1, 2007).

[3] The "Calhoun Report" on the future of cataloging cataloging (< http://www.loc.gov/catdir/calhoun-report-final.pdf > (accessed May 1, 2007), which was both commissioned and highly praised by Library of Congress management, explicitly calls for this in its Recommendation 4.3.5.

[4] See < http://www.loc.gov/bibliographic-future/ > (accessed May 1, 2007).

[5] See "What Is Going On at the Library of Congress?" and "More on What Is Going On at the Library of Congress" at the Web site of LC's professional union < www.guild2910.org > (accessed May 1, 2007).

[6] The "Calhoun Report", ibid., explicitly calls, twice, for the elimination of LC subject headings (page 14: "eliminate LCSH"; page 18: "Abandon the attempt to do comprehensive subject analysis manually with LCSH in favor of subject keywords; urge LC to dismantle LCSH.")

[7] For a summary of these studies see Thomas Mann, "'Cataloging Must Change!' and Indexer Consistency Studies–Misreading the Evidence at Our Peril," *Cataloging & Classification Quarterly* 23 (3/4) 1997, 3-45.

[8] See < http://www.lib.ncsu.edu/endeca/ > (accessed May 1, 2007).

[9] "The Future of Cataloging," *Library of Congress Information Bulletin*, 65, 9 (September, 2006), 206.

[10] See its recommendations on "What Can the Library Community Offer in Support of Semantic Interoperability? " < www.loc.gov/catdir/bibcontrol/TDG_5.pdf > (accessed May 1, 2007).

[11]The continuing importance of being able to browse book collections is insisted on by scholars even today; see the list of user studies appended to my review of the Calhoun Report at < www.guild2910.org/AFSCMECalhounReviewREV.pdf > (accessed May 1, 2007).

[12] For a good introduction to LibraryThing see its site at < http://www.librarything.com/ > (accessed May 1, 2007).

[13] "How Much Information?" < http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/print.htm >, Table 2.3 (Accessed May 1, 2007).

[14] *The Oxford Guide to Library Research*, third edition (New York: Oxford University Press, 2005).

[15] See Notes 2 and 3, above.

[16] For a scheme to integrate the cataloging of selected, high-quality Web sites in library catalogs, so that they show up in the same conceptual categories as book records, see "Is Precoordination Unnecessary in LCSH? Are Web Sites More Important to Catalog than Books? A Reference Librarian's Thoughts on the Future of Bibliographic Control," in *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millenium* (Library of Congress, 2001); available online at: < www.loc.gov/catdir/bibcontrol/mann_paper.html > (accessed May 1, 2007).

[17] Information Literacy Competency Standards for Higher Education, < http://www.ala.org/ala/acrl/acrlstandards/informationliteracycompetency.htm > (accessed May 1, 2007).

[18] *Oxford Guide*, ibid., Chapter 1, "Initial Overviews: Encyclopedias," and Chapter 8, "Higher-Level Overviews: Review Articles," offer other ways to gain overview perspectives.

July 31, 2007

To: Dr. José-Marie Griffiths
Working Group on the Future of Bibliographic Control

From: ACRL Slavic and East European Section

Dear Dr. Griffiths,

We would like to offer our comments to you on behalf of the Slavic and East European Section (SEES) of ACRL.

SEES represents over 200 librarians and specialists involved in Slavic and East European studies. In addition to Russia and the countries of Eastern and Central Europe, the section is concerned with those aspects of library service relating to the study of the Baltic, Central Asia and the Caucasus. Because we support research using many foreign languages and many resources not easily identified or obtained in the West, we very much recognize the value of bibliographic control.

We share the concerns of our colleagues in the ACRL Western European Studies Section (WESS), and strongly support all the points they made in their letter to you. SEES members and the researchers they support continue to rely on controlled vocabularies. This includes not only LCSH (for all the reasons WESS listed in their letter), but also the controlled name headings created by the NACO program. Given the variety of transliteration schemes for non-Roman scripts (such as Cyrillic) and the many possible language forms (e.g., a corporate name that may appear in Russian, English, or Ukrainian), authority control and cross references remain vital to users in our field.

We also want to reiterate WESS's call for continuing support for foreign language skills and subject expertise among the Library of Congress's cataloging staff. LC remains a fundamental player in cataloging in the United States, and the expertise and resources of LC will be almost impossible to duplicate elsewhere.

Finally, we would like to emphasize that our bibliographic infrastructure is an important factor in making the United States a leading research center in the world. Cataloging and metadata play an essential role in that research infrastructure. SEES members and the researchers they support often work in other parts of the world, places that are not so fortunate when it comes to bibliographic control and research tools. We feel that we have perhaps a unique perspective from which to compare a future with strong descriptive cataloging, subject analysis, and controlled vocabularies, and a future without those things.

Some of the very real advantages that technology has brought us may be deceptive. Keyword searching and easy access to electronic resources are wonderful, but they are

not adequate replacements for the structure and organization that content standards and controlled vocabularies provide. We urge you to consider the role that bibliographic control plays in the larger research infrastructure, especially with regard to foreign language resources and international studies.

Yours sincerely,

SEES Executive Board

George Andrew Spencer, Chair, aspencer@library.wisc.edu
Brad Schaffner, Past Chair, bschaffn@fas.harvard.edu
Jackie Byrd, Secretary, byrd@indiana.edu
Sandra Levy, Chair Newsletter Committee, slevy@chicago.edu
Diana Brooking, Chair Automated Bibliographic Control Committee, dbrookin@u.washington.edu
Cathy Zeljak, Chair Access and Preservation Committee, czeljak@gwu.edu
Terri Miller, Newsletter Editor, ticklet@mail.lib.msu.edu

July 26, 2007

Dear Dr. Griffiths,

On behalf of the more than 600 members of the ACRL Western European Studies Section (WESS), I would like to take the opportunity to offer our comments to the Working Group on the Future of Bibliographic Control.

WESS members specialize in acquiring, organizing, and providing reference assistance for information sources originating in or relating to Western Europe. Working so extensively with materials in foreign languages, we are concerned with recent discussions about the continuing need for LCSH. It appears that some in the library profession believe keyword searching and mass digitization renders controlled vocabulary and precoordinated subject headings obsolete. The collective experience of WESS members, many of whom are reference librarians or bibliographers, does not support this view.

Certainly, keyword-searching an OPAC will often locate a large number of titles. However, this approach typically cannot identify holdings on a given topic as efficiently, systematically, comprehensively, or quickly as the use of controlled vocabulary. A large number of results baffles users, who often plod through one extraneous hit after another with little sense of any intellectual relationships among the various titles retrieved. Conversely, too narrow a keyword search leaves a user frustrated at not finding material when, as so often happens, a research library owns plenty on the topic of interest.

As WESS members know from assisting countless researchers, keyword searching alone is even less satisfactory when research involves materials not in English. Excepting subject headings, a keyword search in English will generally retrieve only sources in English. Our faculty and students, however, very often need sources in a variety of European languages. Without LCSH, they—or the librarians assisting them—would have to conduct multiple keyword searches in each language relevant to a given topic. Moreover, they would have to hope that they were correctly guessing the relevant keywords in each language. Furthermore, the serendipity of discovering a work in Spanish relevant to the study of a German author would be lost. LCSH eliminates guess-work and iterative searching, saving time and reliably identifying the full array of material on a specific topic.

LCSH is, therefore, far from obsolete. Even when library users do use keyword searching, we have found, by and large, that relevant results are largely due to controlled vocabulary in subject heading strings. Although keyword searching is often the first step for most users, many quickly turn to subject headings once they realize the increased relevancy they offer over keyword searching. While users do not always understand how controlled vocabulary works, in order to serve them, WESS librarians certainly need precoordinated subject headings. It is critical that LCSH be retained and further developed as a fundamental tool for both patrons and librarians.

Given the need for LCSH, there is a continuing need for foreign language skills and subject expertise among LC's cataloging staff, continuing LC's strong tradition of providing excellent descriptive and subject cataloging. If LC's catalogers were to lack the necessary subject and language expertise, the system of name authority and uniform title access could seriously deteriorate and thus render new editions of European authors, irrespective of when they wrote, essentially invisible in a catalog. WESS strongly urges that LC continue to require relevant subject expertise and foreign language skills in its cataloging staff.

We appreciate your consideration of our concerns. Please contact me should you have questions or require additional information about WESS' position on these issues.

Yours sincerely,

Sarah G. Wenzel, Chair
Western European Studies Section, ACRL
(sgwenzel@uchicago.edu)

WESS Executive Board Members (2007-2008):

Sarah G. Wenzel, Chair, sgwenzel@uchicago.edu
Laura Dale Bischof, Vice Chair, Chair-Elect, bisch004@tc.umn.edu
Bryan Skib, Past Chair, bskib@umich.edu
Jonathan Marner, Member-at-Large, j-marner@tamu.edu

Committee Chairs:
David Lincove, lincove.1@osu.edu
Tom Izbicki, izbicki@jhu.edu
Beau David Case, bdcase@umich.edu
Rowena L. Griem, Rowena.Griem@yale.edu
Rebecca R. Malek-Wiley, malek@tulane.edu
Shawn Jeremy Martin, shawnmar@umich.edu
Brian Vetruba, bvetruba@wustl.edu

Paul Vermouth, Newsletter Editor *ex officio,* vermouth@fas.harvard.edu

Reinhart Sonnenburg, WESSWeb Editor *ex officio,* sonnenburg@dartmouth.edu

Association for Library Collections and Technical Services (ALCTS), a division of the American Library Association

Written Testimony for the First Public Meeting of the Library of Congress Working Group on the Future of Bibliographic Control – "Users and Uses of Bibliographic Data"

Mountain View, California, March 9, 2007

ALCTS is pleased to have the opportunity to comment on the March public meeting of the Library of Congress Working Group on Bibliographic Control, on the theme of "users and uses of bibliographic data." The documentation of that meeting, cited at the end of this statement, presents a consistent picture of the material presented by the speakers and the themes addressed. We would like to focus on two themes: the concept of "users," and a striking call for an increase in the scope of metadata rather than its reduction.

I. Framing the concept of "users"

With regard to the Working Group's question, "Who is using current bibliographic data and how are they using it?", it is first necessary to understand how the very concept of "users" should be situated. For example, on a conceptually simple but operationally impossible level, we could examine the current situation of multiple identifiable user groups and attempt to describe the status quo for these groups. The obvious difficulty, of course, is that this implies hundreds if not thousands of separate answers to the Working Group's question.

While the approach above is clearly not feasible, there is also the temptation of assuming that a given type of user should be taken as typical, and framing a discussion on the basis of that implicit assumption. At least in casual exchanges among librarians, this approach to imagining "the user" is remarkably pervasive. For example, the hypothetical user is frequently assumed to be an impatient undergraduate, a research university faculty member, a genre fiction aficionado, or a person outside academia with a quick information need, among other possibilities. It is far too easy to demonstrate, in consequence of such an assumption, that the user's needs are easily knowable and clearly met (or not met, more frequently).

From the documentation of the Working Group's public meeting in March 2007, there are promising signs of ideas about "users" which steer clear of both of these dangers. An intelligent approach may be based on the research presented by Karen Markey, particularly the concept of "double novices." As Janet Swan Hill notes, "anything we do to improve the quality of discovery/access for people searching at a double novice level will improve the quality of searches for people searching in the other three quadrants," and again, "given the range of types and missions of libraries and other information repositories, and the almost incomprehensible range of needs that libraries and other such institutions need to try to satisfy, that the most critical group to make sure that we serve well are the people searching in the two middle quadrants." A discussion which focuses on improving searching and discovery for "double novices" validly generalizes the question of "the user" in a way that avoids projecting characteristics of one's own most familiar group onto users in general. It may allow for research into user behavior that is less anecdotally bound by specific circumstances than such quantitative research tends to be. It also avoids the essentialism implied in the phrase "The user is the king," from the Google Scholar presentation (as quoted by Kathy Winzer). Although this seems like a common-sense expression, it is at the same time a feel-good sentiment that sidesteps the difficult questions arising from real-world exchanges with actual people in wildly varying contexts.

The Working Group's description of the "consumer environment" and the "management environment," as a framework for discussing user groups and conditions of use, also avoids the pitfalls of attempting to try to answer "who are the users" with reference to specific contexts, and the tendency to project forward from that starting point. However, we must treat the "consumer" metaphor with care. Consumers of products such as food and clothing have, in general, a much better understanding of the nature and quality of what they consume than do consumers of metadata. There is a commonly observed disconnect between user desires as articulated to themselves, and the ability to articulate and understand what it means to satisfy those desires operationally. While a consumer may judge that a piece of fruit has gone bad, and communicate that judgment with a floor manager based on common knowledge, the typical consumer's capacity for describing what specifically contributes to either a satisfying or an inferior information exchange will normally be much rarer.

We note that, from the available documentation of the Mountain View proceedings, it appears that the "users" represented were entirely members of academic library communities. Prof. Timothy Burke, the single non-librarian speaker, was characterized by Karen Coyle as "the thoughtful user that we all hope to meet," and from the evidence, this must certainly have been true. Nevertheless, the concerns expressed by a scholar of Prof. Burke's standing (e.g., identifying "clusters of intertextuality") may not represent those of other library-using populations. This is not intended to downplay those concerns, of course. The statistics presented by Andrew Pace regarding "What Patrons Are Doing," although intriguing in themselves, are again drawn from an academic library setting, as was the testimony provided by Bernie Hurley of University of California at Berkeley. Although it is true that the Working Group encourages written testimony from all interested parties, it appears that few outside academia have provided testimony thus far. Those who have concerns have the responsibility of expressing them, rather than expecting to be "spoken for" by the Working Group itself. Still, in the time remaining, we urge the Working Group to more actively solicit the views of a broader spectrum of public, school, special, and small-academic (including community college) library users into its discussions. We appreciate the fact that the Working Group is taking steps in this direction, particularly in preparation for its third public meeting, and hope that it continues this effort while preparing its final report.

II. The call for an increase in metadata

A significant theme, present in the documentation from the Mountain View meeting was the need for increased, rather than reduced, structured and controlled metadata. Karen Markey called for metadata elements indicating discipline, appropriate knowledge level, authority of the author, genre/literary nature, accessibility (or "what can be done with the document": Markey p. 6), and user reviews or ratings, among other possibilities. Some of these could be supplied via importation of/linkage with metadata from sources outside traditional cataloging metadata. Some of these elements are available now in catalog records, and can be enhanced or mainstreamed in practice (e.g. genre, discipline via classification, audience coding). Timothy Burke's requested "seven tools" are similar in that some are either possible now, or could be available with an investment in strengthened practice. The indication of "lineage of publications" could be approached, at least in part, via FRBR implementation, for example. Nancy Fallgren, in her summary, cites the need for "more/better authority control for consistent data, the ability to evaluate and distinguish among resources, and schemes or formats that enable better interoperability among disparate collections." All of these involve investments built on traditional practices and existing metadata, in addition to collaborations with other communities.

Andrew Pace called for faceted classification, better name authority control for organizations, work identifiers (also with implications for authority control), and enhanced physical description (whether provided by catalogers/metadata specialist or vendors). Pace's "Paradox #1" is particularly of interest: "We finally have interesting discovery tools that make use of bibliographic data in ways that show us that the data are not completely adequate for use with the new discovery tools." This paradox may imply a

number of different paths forward, one of which is to ramp up the investment in more, and more rigorously supplied, metadata. Markey emphasized the need to "embrace post-Boolean searching (with relevance feedback) that gives high weights to subject cataloging data in bibliographic records" (Markey p. 4). This also implies making traditionally-supplied metadata work harder, with systems which dig more deeply into existing metadata.

The provision of enhanced or new forms of metadata may involve drawing metadata from other sources than standard cataloging practice, of course. As Nancy Fallgren notes, "the University of California Libraries' Bibliographic Services Task Force proposes creating minimal bibliographic records and enhancing them with metadata from other sources, such as content from publisher's data." While practices such as this may indeed prove to be sound and beneficial in the long run, a question arises as to whether the principle of "trust but verify" will be central. What will be the user- and use-centered criteria employed for both trust and verification? We are also curious as to the reliability of value-based metadata, e.g., that which indicates the authority of the creator of a document. How will the authority of the creators of such metadata be verified?

In short, we see consistent demands for a greater role for authority control and expanded metadata sets. At the same time, it is clear that end-user supplied subject data, often in the form of tags, has become increasingly important. Indeed, this phenomenon unifies the two concepts of "uses" and "users." We wish to note that dichotomous discussions, of authority-controlled metadata "versus" socially provided metadata, are unproductive. We should not frame such conversations in either/or, zero-sum terms. Instead, we should assume the complementarity of these two major types of metadata, and work to make them consonant. Let us take authoritatively provided metadata, such as LCSH or names established in the NAF, as a comparatively stable (though evolving) base for enhancements and "hooks" to shifting disciplines or user bases. The ideal would be a balance of authoritative metadata, created with specific attention paid by information professionals, with crowd data addressing the concerns of microcommunities. The base metadata can be migrated, upgraded and enhanced en masse precisely because of its comparative stability; at the same time, "users" are able to participate meaningfully in the metadata conversation. (Lorcan Dempsey's discussion of the Expert Economy and the Attention Economy may be analogous here.) This scenario will be best realized with the contributions of catalogers who recognize that their expertise is fully pertinent to the challenges presented by cooperation with a broad range of metadata communities. To serve a broader base of user types and diversified information resources, catalogers must act assertively to expand their roles beyond what is traditionally called "bibliographic control."

Sources:

Coyle, Karen. Coyle's InFormation blog: http://kcoyle.blogspot.com/

Fallgren, Nancy J. "Working Group on the Future of Bibliographic Control, Users and uses of bibliographic data meeting, March 8, 2007, Mountain View, CA: Brief meeting summary."

Hill, Janet Swan. "Terminology, Markey's paper and a few other things." Email message from Janet Swan Hill to members of the LC Working Group, March 19, 2007.

Markey, Karen. "Users & Uses of Bibliographic Data."

Pace, Andrew K. "Users and Uses of Bibliographic Data: The Promise and Paradox of Bibliographic Control. NCSU Case Study: Faceted Navigation" (PowerPoint).
http://www.lib.ncsu.edu/endeca/presentations.html

Winzer, Kathy. Informal summary of Mountain View meeting, prepared for the ALCTS Board Executive Committee.

Prepared by the ALCTS Cataloging and Classification Section on behalf of the ALCTS Board of Directors.

Submitted to the LC Working Group by the ALCTS Board of Directors.
June 19, 2007

Association for Library Collections and Technical Services (ALCTS),
a division of the American Library Association
Written Testimony for the Library of Congress Working Group on the Future of Bibliographic Control
Second Public Meeting: Standards and Structures,
Chicago, IL, May 9, 2007

The Association for Library Collections and Technical Services (ALCTS) appreciates the opportunity to provide written testimony for the Working Group's consideration, in advance of its upcoming public meeting in Chicago. A paper that responds to the reports of the Working Group's first meeting on the topic of "Users and Uses" will be forthcoming. This written testimony addresses the five questions asked by the Working Group in its background paper.

Q1. What kinds of standards and structures are needed to provide effective bibliographic control in the environmental spectrum spanning consumer uses and management uses? How can we make better use of current standards and structures in meeting both consumer and management user needs? What relevant communities need to have input and what organizational structures would best support this?

Standards and structures are needed that are global in terms of world-wide applicability (e.g., useful in scores of languages besides English) and in terms of user groups. This is a situation faced not only by libraries, but by multiple sectors, such as business, government, and academia. The way forward will mean collaborations among all parties in the information chain: publishers, governments, aggregators, authors and other creators of content, readers and other users of information, computer/information professionals, and others.

The enormous quantity of metadata based on current standards (AACR2, MARC, LCSH) has greater potential value than ever, particularly given the ease with which it can be easily "mined" for new applications. The consistency of use of those standards that characterized cataloging practice for the last half of the twentieth century is already proving to be of great benefit as our services develop. Briefly mentioned examples include OCLC's FictionFinder; Columbia University's HILCC (Hierarchical Interface to LC Classification), recently acquired by Serials Solutions; North Carolina State University's well-known application of Endeca for faceted searching. In the realm of social networking, LibraryThing's use of MARC record data and its implementation of LCSH is significant. This point should not be lost as our existing standards evolve and new standards become mainstream.

There is a movement away from reliance on ILS vendors and proprietary software for major enhancements to the "user experience." Nevertheless, the investments made by libraries in ILS systems will continue to be significant, and the systems themselves are the primary points for storage and manipulation of by far the greatest proportion of core metadata (bibliographic, financial, vendor and end user). To reinvigorate the library/vendor relationship, we suggest a new approach to system development. It is possible that a leadership group, drawn with care from across the entire spectrum of librarianship, could envision a "new ILS" with guaranteed basic functionalities, both of traditional types and of those needed to serve present and foreseeable future generations of end users. This set of guaranteed functionalities would become a basic set of specifications. At the same time, there would be, in every library/vendor contract, the guarantee of far greater fluidity of the ease with which traditional core functions may be enhanced, and the ability to link proprietary systems with populist emerging technologies. To put it another way, it is time for the concept "hard-coded" to become a thing of the past.

With regard to community input, it is essential that a broad spectrum of user constituencies is actively included as a normative practice. This spectrum needs to go well beyond academic (particularly research)

and government libraries, and the larger players in the information industry. Public, school and special library representatives should be centrally involved: leadership groups, which may of course be fuzzily bounded and informal, should represent the population at large. "User communities" should also be understood to include linguistic and cultural communities, not only professional communities defined by constituency. To allow for optimal participation, grassroots settings are required, in addition to national conferences and by-invitation meetings. Forums could include state and regional meetings, facilitated by state library associations, OCLC regional service providers, regional resource-sharing consortia, LIS schools, and so on. These can complement online activities, including the uses of wikis, blogs and online forums, virtual meetings, and collaborative development of open-source software tested using significantly large datasets.

Q2. Libraries and related cultural heritage organizations have made a major investment in controlled data. These include structures for organizing subjects, personal and corporate names, place names, roles and relationships, time periods, etc. What role will this data play in networked environments? What is its relationship to the semantic web, tagging, or other newer approaches? How does this data work across database silos? How are supporting infrastructure pieces (gazetteers, controlled vocabularies, etc.) situated and maintained?

It has become evident that, to effectively address these questions, a significant investment is required in interoperability at all levels. Many demonstration projects, as well as working implementations of limited scope, have shown the potential benefits of search, retrieval and display in environments where the user is not required to understand the idiosyncrasies of a given “silo.” What is required is a move toward implementation of the best practices and effective outcomes already discovered, in large-scale end-user environments. This in turn means acceptance of the “perpetual beta” ethos now characteristic of new product development and satisfying user experiences. It is presumed, of course, that further research and experimentation will continue to be essential, but the time is now to invest in production environments, beyond pilot projects and short-term specially-funded efforts.

Types of interoperability include, at least, among languages and scripts, among data structures, and among user community or disciplinary vocabularies. The latter involves building bridges among authority-controlled metadata systems per se, and between these systems and loosely- or unstructured end-user metadata such as tags. We would do well, in the United States, to learn from effective interoperability projects developed in areas of the world where multilingual discovery is an ordinary need. (An example in the area of subject metadata is the MACS project, now hosted at Brussels Free University.) The concept of interoperability may be extended to include combinations or “mashups” of multiple types of data: for instance, place names in all of the world’s languages with geospatial coordinates attached to those names.

The question of how supporting infrastructure elements are “situated and maintained” can be thought of by using the image of interlocking decentralized systems. Assuming the development of multiple levels of interoperability as a basic element of information retrieval systems, it should be possible for a greater variety of controlled data systems to be used in a common environment. In the case of subject metadata, for instance, there would be no question of “shoehorning” the subject vocabulary of a given community – whether that vocabulary is highly or slightly structured – into the constraints which are a necessary part of mainstream systems. The latter (such as LCSH, LCC and DDC) would not only continue to play major roles, but would likely gain in value when flexibly implemented in a variety of environments.

The expertise that librarians have developed in authority control is one of the most important elements that libraries have to offer to the larger information environment. The value of it is recognized by popular information sources, such as the Internet Movie Database and LibraryThing. We need to develop systems that enable authority control to work more effectively in multiple information environments, as well as to

make the underlying concepts more accessible for end users. Developing interoperability between controlled vocabularies and user-provided metadata may be crucial, not only for helping users make effective use of controlled metadata, but also in terms of user assistance and understanding.

The provision of authority control is by nature a complex process. As global interoperability between and among databases of controlled data becomes the norm, the value of this investment in trained human intellect will become increasingly evident. Nevertheless, there are barriers to increased effectiveness and efficiency in the practice of authority control which should be addressed. For example, NACO trainers state that the requirement to consult multiple sources in training (e.g., AACR2, LCRI, DCM, NACO manuals, rule summaries on the PCC NACO pages, etc.) presents as much difficulty, or more, than the complexity of the rules themselves. Other obstacles are presented by factors such as highly complex rule interpretations and apparently contradictory conventions for construction of qualifiers and uniform titles. On another level, there is a need for more effective sharing of authority-controlled metadata, beyond local creation and maintenance of individual databases. Just as the mechanisms which enable interoperability need not be replicated in every local system, so too should local systems be able to make direct use of central, international databases while still allowing for local needs. This holds the potential for a significant reduction in redundant effort.

Q3. Data is created to be processed by applications. We mine it for meaning; merge and manipulate it for display; use it to support supply chains and inventory control; share it between repositories and discovery environments. Are our standards and structures appropriate to this reality?

We would like to offer, to begin with, a caution in response to this question. Data is, in the first instance, created by human beings, directly or via mediated techniques, so that others may gain information and understanding through interaction with that data. (For example, the data underlying works of literary value is not created by authors for the primary purpose of being scanned by an automatic character recognition program.) However, when human beings create data in an information sharing environment pervasively mediated by computers in communication networks, then it is essential to create the data they wish to share with one another in forms that are easily and richly processed by computerized methods. It is in this sense that data is created to be processed by applications.

It is certain that the standards and structures which presently exist are not adequate to support many of the applications desirable and needed for sophisticated data processing in the service of multiple end-user tasks. This is true for the two generalized user environments described by the Working Group as "consumer" and "management," as well as those of identifiable user groups (e.g. parents of young children, community college students, information industry professionals). This brings us again to enhanced interoperability. It is unlikely that overarching "rules, guidelines, models, and structural schema" will be developed which enable the widest variety of applications to share and process data in the many ways described in the above question. It may not even be desirable to develop such universally controlling mechanisms. What is more likely is a continuation of what we have seen for decades, as described in the saying, "The great thing about standards is that there are so many of them." In this connection, our attention should continue to be placed on enabling standards and structures, those existing as well as to come, to interoperate flexibly.

Q4. What requirements are placed on our bibliographic structures through new application areas, such as mass digitization and greater off-site storage, or the desire to create richer user interfaces and integrated discovery environments?

The primary requirements are elegance and fluidity. Elegance implies simplicity, in the sense of facilitating communication among applications, not in the sense of losing richness of metadata content or granularity in coding. Fluidity implies the ability to translate metadata, with its associated coding, into

multiple interfaces and environments. Elegance and simplicity together imply a clean separation of the intended functions of different metadata elements and types from display conventions. At the same time, metadata is associated with context, or provenance, which must be preserved as it is frequently key to the meaning of a given term (e.g., subject terms established according to LCSH as compared with MeSH). Metadata from multiple sources, whether stringently or loosely controlled, should not lose differentiation even when usefully residing in the same system.

Q5. Libraries now manage different flows of data, created within different regimes, much of it outside the library environment. They also want their data and services to appear in other environments. At the same time, we see more reuse and flow of data across publishers, libraries, agents, other bibliographic services, etc. What does this mean for our bibliographic standards and structures?

This question recapitulates the general themes discussed. Standards and structures need to enable metadata drawn from different disciplines and communities to be reused intelligently in a variety of end-user environments. "Library-created" metadata, whether of established or recently developed types, should be able to flow easily into applications created outside librarianship proper. Similarly, metadata types with provenance outside librarianship should be able to be incorporated into whatever "library systems" become, for the purposes of building on the strengths of library-created metadata as well as addressing its weaknesses. Implied are the development of standards and structures which allow fluid sharing and mixing of data types, preservation of context (the metadata's "original intelligence"), granularity in indexing and display to any degree desired, decoupling of markup from display, and translation/transformation into the conventions of multiple end-user environments. It is important that standards of types such as RDA and MARC be further developed to enable "hooks" to multiple other standards. At the same time, these standards need to be supplemented by others whose sole function is to serve as multiplug adapters, in a sense. Analogously, "structures" or organizational bodies such as the JSC, MARBI or PCC would want to consider their standards-developing activities in this light.

References

HILTT: http://www.columbia.edu/cu/libraries/inside/projects/metadata/hilcc/
LibraryThing: http://www.librarything.com/. Note that the "book suggestion" engine provides suggested reading of interest based on several criteria, including "similar library subjects and classifications" drawn from MARC records.
MACS: https://macs.vub.ac.be/pub/
NCSU: http://www.lib.ncsu.edu/catalog/
OCLC FictionFinder: http://www.oclc.org/research/projects/frbr/fictionfinder.htm

Prepared by the ALCTS Cataloging and Classification Section on behalf of the ALCTS Board of Directors.

Submitted to the LC Working Group by the ALCTS Board of Directors.
April 24, 2007

Association for Library Collections & Technical Services (ALCTS),
a division of the American Library Association
Testimony for the Third Public Meeting of the Library of Congress Working Group on the Future of Bibliographic Control
Economics and Organization of Bibliographic Data,
Washington, D.C, July 9, 2007

The ALCTS Board of Directors appreciates the opportunity to provide testimony for the third public meeting of the Working Group on the topic of "Economics and Organization of Bibliographic Data." This statement emphasizes the intimate relationship between these two topics, focusing on the following key concepts: optimization (from the Working Group's question 1), relations among stakeholders (question 2), organizational arrangements (question 3), and the role of the Library of Congress (question 5). The broad ideas expressed here will, of course, have multiple ramifications at a variety of levels in different contexts. This statement addresses what we regard to be essential general principles.

Our primary points are as follows:

- Bibliographic control--cataloging and metadata-provides more than inventory control for single institutions; it provides shared access to information to all searchers over time and creates value as a common good for individuals and institutions throughout the economy.
- The economic benefits of bibliographic control greatly exceed the costs to particular institutions.
- The library community must control the costs of bibliographic control and maximize benefits by thinking and acting in common at a global network level.
- Advanced technology and professionally trained human intellect must be used together to optimize the benefits of bibliographic control and lower its costs.
- Since changes implemented by LC have a large economic impact on all other libraries, LC must be explicit about its plans for changing its contributions and must coordinate implementation with other stakeholders.

Organization

The library and information communities are at an important juncture. There are opportunities to more fully develop current digital technology while reinvesting in the application of trained human intellect. Will we, as the library community, have the foresight to capitalize on both resources, or will we view the most important elements of human intelligence as fundamentally replaceable by automation? In place of the latter barren vision, we should strive to develop a network-level organization to provide metadata in which both human intellect and machine processing complement each other to an extent not presently realized. We need to strategize and operate at the network level, taking advantage of distributed and deeply linked local intelligence. This is a vision of a decentralized system in which organizations involved in standards development and aspects of metadata creation coordinate and communicate their work so closely that maintaining standards and high quality are givens. It contrasts with the current situation, in which comparatively few institutions fund quality bibliographic information, an arrangement that is proving unsustainable. We are at a moment when products such as Endeca and Primo are beginning to manifest the real power latent in bibliographic metadata.

The importance of quality and depth in the creation and sharing of metadata is crucial in realizing the vision of the new technological opportunities. This vision involves optimizing the library community's enormous knowledge base to provide the more granular, flexible, and varied types of metadata called for in the first two public meetings of the Working Group. (The community is itself changing, as publishers and members of other information communities outside librarianship become involved in metadata

provision and cooperative standards development.) ALCTS sees the need for investments in more sophisticated tools to supplement human intelligence, instead of costly attempts to replace it.

One potential scenario for realizing network-level organization might be the following. The library and information communities can consider which metadata record elements should be standard for every (or most every) implementation, and which can or should be customizable according to local needs. This might apply to any type of metadata, not only bibliographic information. We could then strengthen the connection between global and local metadata via linking, instead of the current practice of replicating and updating entire duplicate records in countless individual local catalogs. The perennial vision of "catalog it only once" has always been attractive, but has foundered for the very good reason that local communities have a right to the metadata most appropriate for their needs. Deconstructing a bibliographic record into modular units, with appropriate distributed agencies responsible for different units (descriptive or rights metadata, classification and subject metadata, etc.) holds the potential of balancing network-level efficiencies with respect for diverse communities. This is an investment neither in "one size fits all" nor in pushing the artificial-intelligence metaphor beyond viability, but in the synergistic effects of effective networking combined with human intellect. There may be multiple ways to accomplish decentralized metadata provision; this is just one example. Another potential scenario would be to improve on the current accumulation of metadata from different sources into a coherent whole through shared enhancement, both automated and human.

As part of this vision of decentralized and networked intelligence, specific standards will continue to require specific agencies for their maintenance and development. Network effects take place in the spaces between and among standards. *LCSH, AACR2, AAT* and MARC 21 are examples of standards; *RDA* is a developing standard. (In contrast, "social tagging" is not yet a standard, but rather a loosely described set of practices. It may develop in a completely decentralized and uncoordinated fashion for the foreseeable future.) Standards organizations have the responsibility to maintain them for as long as the community at large finds them useful. Should an organization wish to withdraw from such a responsibility, it is imperative that this be done in a deliberate and tightly coordinated fashion, in collaboration with the relevant stakeholders.

With regard to the stakeholders, members of the cataloging community need to remain responsible for contributing to the variety, depth, and quality of metadata available in shared databases. This is not simply a matter of conscience or an abstract idea. When each institution does its part to contribute to the common good, such investments translate into widely shared strength and cumulative value. In the medium to long term, it is more prudent to contribute to the common good than to define policies and practices with an eye only for the short term and narrowly local. This is a form of network-level thinking, with great and still untapped synergistic benefits for library users and information seekers of all types.

An example of contribution to the common good is for libraries (of all types: school, public, academic, corporate) to enhance records for local use, by upgrading minimal level records, by providing subject analysis, and/or by adding notes or tables of contents, and by contributing those enhancements to a shared bibliographic environment in order that networked libraries and patrons may all benefit. It is vital that library managers recognize the central importance of stakeholders acting in concert rather than in isolation, and provide the leadership to develop appropriate economic and professional incentives.

The role of the Library of Congress in a network-level system of metadata provision is a matter of great concern to the library community around the world. ALCTS acknowledges the enormous complexity surrounding the Library's provision of bibliographic services over more than a century, its status as a Congressional library which is regarded as a *de facto* national library, and the economic challenges it faces. Drastic changes in Library of Congress policy and practice are likely to have destabilizing effects at the international level if not handled with great care, much advance planning, and wide consultation.

The library community needs the Library of Congress to state unambiguously what its plans are with regard to the provision of bibliographic services. We realize that plans change over time, and must be altered as circumstances change. Therefore, ALCTS asks that if the Library of Congress intends to withdraw from any of its traditional functions that the announcement be made with sufficient lead time to ensure a well-considered and orderly transition. This will enable the library community to develop and implement appropriate practices and procedures at the local level and cooperatively with other organizations.

Economics

Allocation of resources is fundamentally a political matter. Current economic conditions must be viewed in light of the actions taken to shrink public sector/civil society funding over the past thirty years, beginning at least with the "tax revolts" of the late 1970s. To assert the political nature of resource allocation, and therefore the mutability of any current condition, is of course not to solve the economic problems faced by any institution. Acting with professional responsibility requires more than making statements about "limited resources." Acknowledging limitations is one aspect of any worthwhile discussion about resource allocation, not the endpoint. The responsibility of management is to develop and implement plans that support institutional visions and values. The politics of advocating for the budgetary support of trained human intellect may be difficult to face, but is more important to our profession than assumptions such as that a radically decreased professional workforce represents some kind of valuable opportunity.

It is often claimed that cataloging is "expensive." This claim by itself means very little in isolation, outside of a broader context. What values are supported by any level of expense? What public goods are supported? What are benefits for the community at large, in the short, medium, and long term? Questions such as these must be addressed when evaluating whether or not any particular expenditure is appropriate or adequate.

The $44 million figure given in recent years for Library of Congress cataloging costs must be compared to the hundreds of millions in savings accrued by U.S. libraries that the Library itself has used in the recent past to justify these same expenditures. It applies, as well, to the expenses of all institutions belonging to the greater network. No institution is an island; disinvestment in the public sector means damage to the private sector as well. Conversely, a strengthened public sector benefits all. (The economic stimulus provided by city-wide wireless networking makes for an interesting analogy.) Incrementally modest investments in services provided outside the boundaries of one's own institution are likely to pay off well in benefits returned.

Although the Library of Congress does not receive funding specifically for providing bibliographic metadata to the nation and the world, ALCTS wishes to emphasize the inestimable value of the Library's long-term functions as a global standards-bearer and as a generator of substantial economic savings in and beyond the United States. The economic benefits are, in themselves, arguments for investment in bibliographic access through both a skilled professional and support staff as well as through technological innovation.

Prepared by a Task Group appointed by the ALCTS Board, based on substantive contributions by the ALCTS Cataloging and Classification Section.

Submitted to the Working Group by the ALCTS Board of Directors.
July 30, 2007

1. What kinds of structures and standards are needed to provide effective bibliographic control in the environmental spectrum spanning consumer uses and management uses? How can we make better use of current structures and standards in meeting both consumer and management user needs? What relevant communities need to have input and what organizational structures would best support this?

   The reasons for structure and standards have not changed significantly for libraries in managing their resources. Allowing customers to search and identify specific resources and maintaining inventories of all kinds of materials to allow for accurate purchase and circulation remain key issues for libraries. All organizations, libraries, museums, archives, who maintain resources, both digital and physical, all share similar needs.

   The problem in the past has been proprietary metadata standards, sometimes for specific kinds of organizations (libraries vs. museums); sometimes for kinds of materials (MARC vs. Dublin core vs. EAD); sometimes even built for specific systems, has led to closed, highly bureaucratic standards that are inflexible, impossible to share across systems, and impossible to maintain relevancy in the consumers' world. Most of the older standards are too complex and impossible to change without several years of discussion and a few more of implementation.

   The focus from our perspective needs to be on an easy, flexible structure that can be used for all resources, across systems, but allow for reliable identification of specific items in any form from the serious researcher. Although not the only pieces of data necessary, the key elements are names and subjects, in their broadest sense, and a reliable cross-reference system that will continue to keep new vocabulary in synch, if you will, with standard organizational vocabularies.

2. Libraries and related cultural heritage organizations have made a major investment in controlled data. These include structures for organizing subjects, personal and corporate names, place names, roles and relationships, time periods, etc. What role will these data play in networked environments? What is the relationship to the semantic web, tagging, or other newer approaches? How do these data work across database silos? How are supporting infrastructure pieces (gazetteers, controlled vocabularies, etc.) situated and maintained?

   The biggest concern we see in this area is keeping relevant terminology for all customers. Can this be done with a centralized "authority" system? It is frustrating to search Sam Clemens and come up with no hits. And, these events have the potential of becoming even more frustrating as social tagging expands. We need to be able to bring together like terminology from our customers: "detective story", "mystery", "who-dun-it", regardless of which is used, so that others searching for like materials can identify and retrieve them. How can we insure that we are not missing anything? Granted for the undergraduate, this may not be an important as for the PhD candidate who has focused on a dissertation subject that has more rigorous requirements. We don't believe this goes away even with everything available online. In fact, it would seem that focusing down to specific areas of research will become even more difficult without some means to synthesize cross-taxonomy searches, kind of like a "wikipedia of controlled vocabulary". Since access to many of the vocabularies used is either controlled by subscription or being part of an organization, the current environment allows for little cross-system use.

3. Data are created to be processed by applications. We mine data for meaning; merge and manipulate data for display; use data to support supply chains and inventory control; share data between repositories and discovery environments. Are our structures and standards appropriate to this reality?

   We've established that there are no current standards/structures in place that will deal with this situation. NISO hasn't even addressed the issues of cross system interoperability of taxonomies/vocabularies. And we have to get out of the forms for data entry, e.g., AACR2, 3 or 4. If names are deemed important, insure that anyone, including new local taggers, can do it correctly.

   First name: Organization name:
   Last name:
   Middle name:
   Birth date:
   Death date:

   And let the systems deal with putting it together.

4. What requirements are placed on our bibliographic structures through new application areas, such as mass digitization and greater off-site storage, or the desire to create richer user interfaces and integrated discovery environments?

   No matter where or what kind of resources we are talking about, relevance is still important to researcher. The system needs to provide relevant results, not quantity as much, because we know the customer won't look beyond the first screen or two, max. And currently, the way searching is done now some of the most relevant results seem to be on screen 99. Consequently, until our search engines become much more savvy, some kind of vocabulary control is necessary for our research customers.

5. Libraries now manage different flows of data, created within different regimes, much of it outside the library environment. They also want their data and services to appear in other environments. At the same time, we see more reuse and flow of data across publishers, libraries, agents, other bibliographic services, etc. What does this mean for our bibliographic structures and standards?

   It seems that we are still in need of some kind of metadata system that provides a set of standard elements to handle all resources. And, that the vocabularies have some kind of reliable authority [especially in cross-reference arena], especially as social networking increases and terminologies get more bizarre, or change with the times, as the case may be.

Working Group on the Future of Bibliographic Control

Library and Archives Canada

Written Testimony
July 20, 2007

Library and Archives Canada (LAC) appreciates the opportunity to submit a written testimony for consideration by the Working Group on the Future of Bibliographic Control. LAC's statement will focus on bibliographic control in a digital environment, cost factors, multilingual issues, and the implications of these factors for a national institution.

Organization of Bibliographic Data Creation

The concept of Universal Bibliographic Control first articulated in the 1970s by IFLA is rarely heard now, but nevertheless continues to underpin the international and national organization of the creation of bibliographic data. Creation of bibliographic data follows standards and structures which allow the data to be shared and exchanged. Under UBC, the "national bibliographic agency" in each country, usually the national library or *de facto* national library,, is responsible for documenting the publications of that country, in other words, for compiling a national bibliography. This bibliographic data is collectively intended to describe the publications of the world, with the ultimate aim of aiding worldwide sharing of information and knowledge through publications. The data is also made available to other libraries in the country and around the world, for re-use in their own catalogues, thereby achieving significant collective cost savings.

This highly structured, distributed international model for organizing bibliographic data creation and dissemination has been functioning more or less successfully for decades. However, with the change in scale of publishing brought about by digital publishing and digitization, and the emerging role of the Web in bringing digital information to consumers and citizens, the question arises: is this model sustainable? Use of metadata generated automatically or supplied by interested parties such as publishers, authors and vendors is increasing. Social tagging and other forms of citizen participation in metadata creation is also occurring. Sources of bibliographic data are becoming more diverse, and metadata is being created in increasingly rich and varied forms. Traditional clients are becoming not only the users of our services but also our partners in the area of metadata creation.

Digital Publishing

Library and Archives Canada has stated that one of its key strategic priorities is to adjust all aspects of its activities to adapt to the needs of the digital information environment and to benefit from the opportunities it presents. This strategy reflects the vastly increased and rapidly growing importance of digital resources and information sources in the lives of all, in Canada and elsewhere. Part of

LAC's digital initiatives includes the extension of legal deposit to Canadian online publications as of January 2007. Another is to undertake a series of crawls of the Web domain of the Canadian federal government, and of all the provincial and territorial government Websites. It is estimated that these crawls have netted for the LAC digital collection approximately 55 million digital objects representing about 900,000 digital titles, in addition to the 30,000 digital titles collected and catalogued individually. LAC is also embarking, in conjunction with national partners, on a major effort to digitize holdings on a large scale.

Bibliographic Control of Digital Publishing

In view of the scale of publishing this represents, LAC has developed a "Web Resource Discovery Policy" which outlines a strategy for providing access to digital publications through full-text indexing of the documents, in combination with automatically-generated metadata; metadata supplied by others (publishers, authors); minimal bibliographic records created for purposes such as acquisitions; and finally, full, standard bibliographic descriptions for a small subset of carefully selected digital titles of particular value or significance. Digitized titles would be added to the bibliographic record for the original, rather than receiving a separate bibliographic record. This policy is being implemented in phases as systems development proceeds.

The policy also recognizes that bibliographic data is not the only means of gaining access to publications. Digital or digitized documents can be accessed through full-text searching using user-friendly search engines. While this approach does not benefit from the rigour of traditional bibliographic description and the resulting collocation of works with the same author, title, or subject, to distinguish one work from another, it provides access that is "good enough" to the vast numbers of digital resources that LAC is mandated to acquire. User needs for search-friendly, timely access to large quantities of digital information was a key consideration in developing this policy.

The effect of this new policy will be to allow LAC to provide cost-effective and timely access to a large number of digital publications. This will mean that while LAC will not provide bibliographic data for all Canadian publications, it will provide bibliographic or other access to a larger total number of publications, including the important digital publications. However, it will also mean that Canadiana, the national bibliography for Canada, will not aim to cover all the publishing output of the country. If this implication is generalized to other national libraries, the bibliographic control of digital publications worldwide is compromised, at least as bibliographic control is understood traditionally.

Cost versus Value of Bibliographic Data Creation

The value of bibliographic data is under increasing scrutiny, and there is great pressure by publicly funded organizations to reduce the cost of data creation and to prove value for money. In light of the questions raised about the value of bibliographic data, and the expense of its creation, public institutions such as

national libraries may not be able to continue to create bibliographic data in the quantity and quality of past years, unless they can demonstrate its value for money in the eyes of their funding authorities. Public funding authorities base their decisions ultimately on the value and impact of this activity for the citizens of the country.

Multilingualism

Library and Archives Canada would like to emphasize the value of bilingual and multilingual access to the world's publications; while English is a useful common language for many international interactions, it can sometimes overwhelm the richness and essential qualities of other languages and scripts. Bibliographic and other access should become increasingly hospitable to the many languages and scripts found in the published voices of the world.

Moving from Control to Access

As the future of bibliographic control evolves, LAC hopes to maintain a balanced approach. LAC will continue to create bibliographic records for Canadiana in all its manifestations to meet the needs of users and other libraries, but will seek to reduce the costs of this activity by a variety of means, including the selective reduction of detail in description. At the same time, LAC will work to increase the perceived value of this data through the use of improved display and search techniques, enriched content in the descriptions, and the use of Web 2.0 social metadata features to involve Canadians in the description of their heritage. The task of increasing the perceived value of bibliographic data is of course shared with all libraries.

In the realm of digital Canadiana, LAC's approach will be to maximize use of non-bibliographic techniques for the provision of access (as described above), and to maximize to the extent possible, and through partnerships and a national strategy, the quantity of digital publications and other digital heritage available to users. While following this path, LAC will seek to increase its capacity to provide access to multilingual and multiscript Canadiana.

By employing the various strategies outlined above, LAC hopes to achieve a balance between meeting the requirements of its national role of providing authoritative bibliographic information about Canadiana to other libraries, and the provision of access to Canada's information resources, especially digital resources, in a cost-effective, results-focused way, for all users. This approach reflects an evolution of the concept of bibliographic control towards a model oriented towards bibliographic access and resource discovery.

9 July 2007
NAL testimony to the Library of Congress Working Group on the Future of Bibliographic Control – Third Public Session

Good afternoon, I am Christopher Cole the Associate Director for Technical Services at the National Agricultural Library (NAL). We thank the Working group for the opportunity to offer our comments. I would like to take a few minutes - a very few minutes - to present NAL's viewpoint on the Future of Bibliographic Control.

As both a library and indexing publisher, NAL has experience creating indexing records using basic metadata supplied by publishers. These records are enhanced by our staff with access points and quality control. Using publisher supplied metadata has saved considerable costs over creating records in-house from scratch and the quality has not suffered at all.

NAL believes that a similar approach to creating bibliographic metadata using publisher information is possible and necessary for LC and other libraries. The basic records can and should be created by the publisher. The librarians can then add value through access points and quality control.

The Library of Congress has played a critical and beneficial role in creating bibliographic metadata for new books through its Cataloging in Publication program. We at NAL are proud of helping contribute to the CIP program by cataloging agricultural related titles. CIP helps libraries of all sizes build their catalogs in a timely and affordable way. There are many small libraries that rely on the CIP data printed ion the books to create their local catalogs. Unfortunately, the process of creating the CIP records does not take advantage of the publishers' data to create the records in a fast, automated, and affordable manner. We have concerns whether the present labor intensive process is economically sustainable and what the effects of its failure would mean for America's libraries.

There are more opportunities for libraries to obtain basic metadata from the content producers other than just book and article publishing. The movie and recording industry creates solid metadata records for each item they produce to enable those items to be distributed and sold. If libraries can adjust their focus from creating entirely transcribed bibliographic description to adding value through authority control and access points, these and other metadata resources can be tapped. This is especially important as these items are increasing produced and delivered in digital format and without good metadata, the library will be unable house and retrieve the items.

The discussion of using non-library metadata is not simply a one-way exchange of vendor to library. NAL believes that we can and should focus on making our bibliographic data easily accessible and manipuable by other communities. To make this a reality, libraries will need to reconsider our catalog structures and practices. We need to focus on providing value to communities beyond our traditional users. We have developed controlled subject vocabularies and authority files that have applicability and utility outside libraries. The ongoing process of creating a new catalog code is a logical point to begin this transition. Unfortunately, it appears the focus of the RDA remains on traditional materials and procedures.

The Library of Congress in partnership with libraries across America has created an essential resource using our shared cataloging records. Our catalogs describe and provide access to the harvest of our society's creativity. The LC's role in the creation and implementation of standards has been and will continue to be essential. We are not recommending the abandonment of standards but a transformation. Our goal should be to describe all types of materials that we acquire in a simple and consistent way. We should focus not on a "record" but a clear set of data elements that can be assembled by libraries, vendors, indexers, and others into the specific display formats needed for their uses.

Testimony to the LC Working Group on the Future of Bibliographic Control

Submitted by
Dianne McCutcheon
Chief, Technical Services Division
National Library of Medicine
July 17, 2007

We would posit that cataloging is a "public good," but disagree with the corresponding implication that that means cost is irrelevant. All public services have a cost and managers need to determine the appropriate cost benefit ratio. As public institutions, supported by taxpayers, NLM and LC have a particular obligation to ensure that cataloging is being supplied in the most efficient, cost effective manner to achieve the needed goals. For NLM, the goals are to expose its collection of world's biomedical literature, and make it available to the public. NLM does this for material at various levels of granularity. Our largest database is MEDLINE/PubMed, which provides controlled subject access to material at the article level. Material at the book or journal level is provided through the online catalogs, L+ and the NLM catalog.

Until very recently, there was only one way to provide access to this material and that was by manual transcription/keying of data, because no alternatives were available. However, now publishers and vendors are working in an electronic environment, and even print material generally originates in an electronic document. Therefore, basic descriptive data are available in an easily transmissible and ingestible format. Why would we not want to take advantage of this data and eliminate the rote keying tasks and allow our staff to devote their time and attention to more professional tasks, such as subject analysis and authority control? Obviously the ideal situation is that the descriptive data supplied early in the publishing chain is compatible with the needs of library catalogs so it can be ingested without the need for significant editing.

About 10 years ago, NLM went from manually keying article citation data, to using electronic metadata supplied by journal publishers in XML format. NLM has been able to realize considerable cost savings in the creation of MEDLINE records, even factoring in the quality control review of the publisher-supplied metadata. NLM was able to get the publishers to supply us with the data in a standard prescribed format because they want to be cited in MEDLINE. We estimate that even with increased quality control costs for reviewing the publisher's data, cataloging costs for monographs could be reduced by 20% or more if we could ingest the electronic metadata. If we are to get usable XML data for monographic material, we must either convince the publishers it is to their advantage to be cited in our catalogs, or make our descriptive standards so usable, logical, and straightforward that publishers will want to adopt them, rather than spend the time and effort to develop their own. NLM estimates that even with increased quality control costs for review, our cataloging costs for monographs could be reduced by 20% or more if we could ingest the electronic metadata.

In the early 70s many libraries created their own OPAC systems (LC's MUMS, Harvard's Hollis, NLM's Catline). Economically this proved to be unfeasible and

redundant work and libraries let commercial vendors take over this development, recognizing that it made more economic sense, even if they were not able to obtain every feature they had in the past or might desire. By the same token, using existing metadata for more of our resources makes economic sense, even if every record is not exactly the way a particular library might have created it itself. Giving up customization is clearly not easy or even feasible for everyone. NLM will always want to use MeSH and the NLM Classification. It is very revealing that when OCLC and RLG merged, the research libraries insisted they still needed access to their unique institutional records.

How do we get the metadata earlier in the life cycle chain of a bibliographic resource? It is highly unlikely that publishers and vendors will supply us with MARC data. We must be able to use more open and widely compatible formats like XML. Publishers are already using the ONIX standard to transmit bibliographic data to booksellers. While the bookselling community has some needs that libraries do not, the basic descriptive data should be the same in both environments. We need a simple, inexpensive way to get this data directly into our catalogs. Until such time as our OPACs can handle non-MARC data, conversions or crosswalks will need to be created to take in ONIX or similar types of data. Rather than each library having to redundantly write these specifications, this seems like a reasonable task for a central group to take on. OCLC is a logical choice since they are already receiving records from many large vendors.

However, we cannot rely solely on OCLC as the distribution mechanism for this data. It is important that we not forget the large number of small libraries in the US who are not OCLC members and cannot afford to purchase this data. LC has historically provided bibliographic data to these libraries at little to no cost, often by providing CIP data right in the item itself. If LC were to abandon this role, who else in the nation could pick it up? Officially LC is the library of *Congress* and their main responsibility is to support and serve the members of Congress. However, Dr. Billington publicly presents the library as "the nation's library" and many of the activities of LC have traditionally supported other libraries in this country. The LC legislation and budget needs to reflect this role.

Automating processes where possible generally increase efficiencies. Cataloging rules should be developed with automation in mind. Many worry that automated data is of lower quality than that of manually created data. Current cataloging practices are the equivalent of handcrafted materials, like antique furniture or handmade clothing. For most of us, while we may admire the old pieces in a museum, or order them for very special occasions, the fact is that the mass produced material serves us very well and economically and allows more people access to these goods. The same would be true of more automated cataloging. Too often this is approached as an all or nothing concept and rejected on that basis. Some say we can't accept ingested metadata because there is still a lot of material (older titles, foreign titles) that won't have electronic data associated with it. That is no reason not to take advantage of the material that does have existing metadata, and in fact, frees the library to reallocate its resources to the material that really needs human attention.

Others say that a machine can never replace human judgment, particularly in areas like subject analysis. In many cases, this is true, but that is not to say that machines cannot assist humans. Being able to quickly see and process large aggregations of data, a

machine may be able to pick up relationships and concepts that a human cannot. Critics of automated subject analysis compare the machine to a trained subject expert, but the truth is that in most libraries, economic realities mean that catalogers must catalog in a variety of subject areas and cannot be expert in all of them. NLM has developed a software product, MTI (Medical Text Indexer), to assist indexers in selecting appropriate terms from its controlled vocabulary, MeSH, for several years, and recently completed a study on its usefulness for catalogers. While the tool is not completely configured for cataloging purposes, rather than indexing, we found a about a 60% agreement in the terms chosen by MTI and the catalogers. While automated subject assignment may work better in the hard sciences, than in the social sciences, why not begin to use this more where it can work effectively? It doesn't have to be used on every type of material or without any human review or intervention.

The current economic model for traditional cataloging agencies needs to be re-examined. Each individual library catalog is a separate silo. Even when data is shared through OCLC, an upgrade or correction to a record must be made over and over again in each catalog. Libraries shoulder all the costs to create the cataloging data and then pay for the privilege of sharing that data in OCLC. While OCLC provides some financial credits for creating new records or upgrading others, that is a one time thing, and after that, regardless of how many times that record is used by other libraries, no additional money accrues to the libraries. In this model, libraries and vendors have little incentive to do original cataloging. Libraries traditionally have had an altruistic outlook and are working and sharing data for the greater good, but publishers and vendors are profit making institutions. Perhaps it is time to reconsider the OCLC model, and pay the record originators *each* time their records are used. This might encourage others to pick up the slack if LC is unable to create as many records as in the past.

It is important to remember that OCLC is not in the business of creating bibliographic records (with the small exception of its contract cataloging services). The database exists only because libraries contribute their records. If libraries eliminate cataloging services, then nothing new goes into the OCLC database. The tools for more automated subject analysis or authority control are dependent on the large aggregations of records in the LC or OCLC database. Much of the power of the Internet search engines to find bibliographic material is leveraged on the metadata already available from the libraries. As more and more full text goes online, users will be drowned in results if there is no way to distinguish more significant data from random text. One major advantage to a library catalog user is that the rules used to create the records and the search results obtained are open and available to anyone who wants to discover them. Internet search engines keep their indexing and search strategies secret, so users have no way of assessing the reliability and completeness of the results they get. (Eversberg, B. On the theory of library catalogs and search engines. 2005.)

Libraries also need to be exploring the possibility of enhancement of our records by our users. There has been an historic distrust of users in the library community, going back to the days when stacks were closed because "users might find the wrong thing" or not stocking fiction in public libraries because it wasn't educational material. However, as sites like LibraryThing demonstrate, there are motivated and educated users out there who can add value to library created data, but again, without the basic cataloging records created in libraries, these sites have nothing to work with. Obviously, we need to be

cautious in this approach, since we know there are groups with political and other agendas who might try to co-opt the bibliographic record's neutrality, which is still an important factor. One starting point might be for LC to cease providing subject analysis for adult fiction, and instead let users "tag" this material. These titles are unlikely to be used for research, so experimenting in this area should have no long-term repercussions for the serious scholar.

Internationalization has encouraged adoption of standards that can be used more widely than just in the US, or even the English-speaking countries. This expansion is advantageous, in that it potentially allows us to accept bibliographic data from a wider variety of sources with little editing needed. However, as the number of stakeholders increase, the time it takes to make any changes to standards seems to increase exponentially. This might have been acceptable in the past, but in today's environment, things are changing too rapidly to make a 2-5 year development cycle for a standard acceptable. By then the standard will be out of date because of changes in the computing or publishing environment. Our standards need to be made simpler and more flexible, based on general principles, rather than case by case instances.

In summary, we know that descriptive metadata is available for much of the material being added to our catalogs. We must make use of this metadata, without unnecessary modifications, wherever possible. We must look to automate routine activities, and use the power of the computer to assist people with the value-added activities that catalogs provide over search engines, such as controlled vocabularies and headings. We need to investigate new collaborations to change the models of metadata creation and distribution and standards development. Only this way can we survive and continue to provide control to the avalanche of information headed to our users.

U.S. GOVERNMENT
PRINTING OFFICE
KEEPING AMERICA INFORMED

August 6, 2007

Dr. José-Marie Griffiths
Dean and Professor
School of Information and Library Science
University of North Carolina at Chapel Hill
CB#3360, 100 Manning Hall
Chapel Hill, NC 27599-3360

Dear Dr. Griffiths,

The Government Printing Office (GPO) commends the Working Group on the Future of Bibliographic Control for soliciting input for its report and recommendations to the Library of Congress (LC). As the national cataloging authority for Federal Government publications and long-time member of LC's Program for Cooperative Cataloging, GPO has a keen interest in the outcome of Working Group's findings.

GPO's Future Digital System (FDsys) is a world-class system for managing the life-cycle of official Government content, which will verify and track versions, assure authenticity, preserve content, and provide permanent access. FDsys has the capability to employ multiple established extension schema and input standards, including MARC, for expressing metadata when possible and it supports the capability to employ additional established extension schema for expressing metadata in the future.

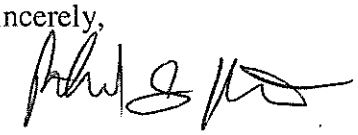
Metadata Encoding and Transmission Standard (METS) is the encoding standard for content packages in FDsys and metadata files are encoded in XML and conform to schema adopted by FDsys. Multiple extension schemas including, but not limited to, Dublin Core, PREMIS, Metadata Object Description Schema (MODS), are used to relate the administrative, technical, preservation, rights, and source information to the digital object.

GPO's integrated library system (ILS) will be a major component of FDsys. The ILS will feed bibliographic data in MARC format to FDsys and the *Catalog of U.S. Government Publications* (CGP), the online catalog module of the ILS, will remain as one means to search for publications that are within scope of the Federal Depository Library Program and GPO's Cataloging and Indexing Program (C&IP).

In addition to the metadata requirements associated with FDsys and the MARC records created for the C&IP, GPO is also exploring the use of automated data extraction tools to create XML records. Preliminary work indicates that these tools are particularly successful with publications that have recurring patterns in their format, e.g., technical reports and Congressional reports.

GPO is pleased that LC, through the Working Group, is exploring the use of alternative metadata schemas and additional tools for creating bibliographic records and providing bibliographic control in the digital information environment. Again, thank you for the opportunity to provide comments to the Working Group on the Future of Bibliographic Control. If I can be of assistance please feel free to contact me at rdavis@gpo.gov.

Sincerely,



RICHARD G. DAVIS
Director, Library Services and Content Management
and Acting Superintendent of Documents

York University Response to the Library of Congress Working Group on the Future of Bibliographic Control

Compiled by Stacy Allison-Cassin, York University, Toronto, ON, sacassin@yorku.ca

The following represents the responses of Heather Fraser, Head, Bibliographic Services, York University; Tim Knight, Head, Technical Services, Osgoode Law Library, York University; and Stacy Allison-Cassin, Cataloguing Librarian, York University. It does not represent the opinions of our institution.

It is clear that the cataloguing landscape is undergoing a radical shift, but it is unclear as to what the outcome will be. We have outlined some of the areas where we see increasing pressures below.

1. Cataloguing rules are complicated

Cataloguing rules have been very useful in helping us standardise bibliographic records, assisting both cooperative cataloguing efforts and the creation of good metadata however they are now more of a hindrance than a help.

- reflect the material catalogued and differences in publishing practices
    - includes rare pre-publishing industry materials and not-so-rare 'unpublished' post-publishing industry materials
- are mired in bureaucracy making them slow to change resulting in inability to keep pace with current changes in formats and terminology
- we have had to be more flexible in terms of interpreting the "rules" in order to provide access to our collections
- are difficult to learn and internalise and difficult to teach to paraprofessionals in a work-place setting

2. MARC coding is complicated

Like our cataloguing rules, MARC coding is overly complicated and does not easily translate to other systems. While MARC is a very rich source of metadata, its promise was never realised.

- reflects complicated practices (see 1)
- originally designed to facilitate printing on cards, and although robust and useful was not translated well into the world of the OPAC
- facilitates access through structured data structures
- not used to its full potential by cataloguers and/or ILS developers
- most current LMSs have not been able to keep up with changes to MARC and making changes to database policy can be complex
- It would be better to move to a more flexible system of encoding metadata such as XML.

3. Publishing and published resources are complicated

Publishing practices have always made cataloguing more complicated and it does not seem likely that this will change. Can we depend on publishers to provide clear, accurate

bibliographic information?

- each publisher presents their materials in their own way
- difference in carriers has further complicated description of resources
- rare materials, unpublished materials and materials published on the the internet muddy the water

4. Subject analysis and application of subject headings/descriptors is complicated

Subject analysis is one of our key professional activities as cataloguers, however it has become clear that we increasingly face challenges in our abilities to give good subject access to resources.

- LCSH is slow to change and falls behind current practice
- LCSH uses outdated and arcane terminology
- LCSH not a true thesaurus with proper broader, narrower, and related structure; what's there is incomplete and often plain wrong
- due to the economic advantages of accepting catalogue copy from LC we are overly dependent LCSH as the incorporation of other thesauri into our catalogues is expensive
- economic pressures prevent us from doing thorough and professional subject analysis (see
- purpose not well understood in the profession including cataloguers, reference librarians and library users
- where possible it is important to strive for accuracy in terms of access points and subject access to provide good data for retrieval
- research question-->subject terms-->resources – subject headings are the glue that potentially brings users to resources, 'every book its reader'
- 'tagging' can be more current but lacks any structure and consistent application

5. Classification of materials vs. 'mark and park'

Classification is a valuable activity and is should be retained.

- classification can facilitate browsing in and 'educative' way, i.e. users can at least learn that their materials are found at a certain class number
- library users need not understand classification in order to benefit from it
- can be used away from the resources, i.e. browsing through the catalogue
- it has been suggested that browsing occurs as users move through the library looking for the call number they've found, i.e. As one navigates the shelves and moves through the physical space, perhaps also true when negotiating the information space of the catalogue

6. FRBR is fuzzy, single vs. multiple records

FRBR is a useful data model and presents a challenge to libraries to decide how to handle multiple records.

- library users/librarians receive the multiple record option poorly
- we are increasingly required to provide access to different forms of the same content and must decide what works the best for our institutions: one record vs. multiple record approach

- FRBR combined with a well developed ILS could help simplify presentation of items in the catalogue
- a clearly thought out presentation of the work-expression-manifestation-item structure could successfully play out using tree-type structures (something everyone is familiar with) in a system that handles it properly

7. Increase in electronic publishing and digital documents

The explosion of digital content is a challenge to cataloguing. It is impossible to catalogue all the relevant digital content. We must decide what to provide access to, and how to do it, or, leave the provision of access to other entities such as Internet search engines.

- new publishing formats and self-publishing opportunities accessed via WWW
- perception that “business as usual” was not working and that we need help to deal with this exponential increase in access to resources
- transition of serial publishers (and now monographs) to digital formats
- very popular in academic environment
- allows publishers to aggregate successful titles with less successful ones
- New modes of publishing such as blogs, wikis and podcasts present new challenges

8. Research is complicated

The friction between the seemingly easy world of Internet search and the complicated “closed” world of library (re)search has made it more difficult convince people to use library catalogues.

- research is not easy
- however experience with Internet search engines has led people to believe that it is and as a result 'people don't do real research anymore', and they are allowed not to, i.e. spoon fed as undergraduates
- cataloguing can facilitate the research practice in libraries
- libraries have never been the first place to start research, more often occurs through external sources, e.g. colleagues, professors, references in existing research, etc.

9. ILS does not take full advantage of the data provided

Many ILSs have made our work more challenging

- not developed by/with cataloguers (see 17)
- updates and fixes are slow and require significant outcry before vendors even consider requests for enhancements
- system configuration should be more openly adjustable and not left to systems people who are not librarians or have close relationships with librarians
- bibliographic utilities also need to be more flexible in terms of accepting records from contributing institutions as with shared cataloguing it is easier to share expertise

10. Although considered the 'core' of the profession cataloguing is misunderstood by professionals and library users alike

- see 1-5
- periodically having to justify the importance of cataloguing to administrators

11. Cataloguing increasingly carried out by para-professionals and library technicians

The core of professional librarian cataloguers has been shrinking and the reliance on non-librarian cataloguing has been increasing

- librarian trained cataloguers are becoming rarer and retiring staff are not replaced, which doesn't make sense as amount of information increases number of information organizers decreases (but this is an elitist librarian view as there are now many more players involved outside of the profession)
- lack of professionally trained staff actually doing cataloguing has placed an extremely high reliance on LC cataloguing
- this has led to blind acceptance of 'LC' records and the gradual erosion of consistent, high quality, useful cataloguing records, especially in academic institutions where it counts the most
- LC and the leading academic institutions can't cope with the demand on the increase in available information
- cataloguing practice is buckling under the strain and has been for years

12. Traditional workflows must change

- "shelf-ready" and also purchases of files of MARC records have made it easier for some institutions to provide access to resources that they would be unable to provide with in house cataloguing staff, given that however, we have to be able to accept less than perfectly "clean" databases
- it is important to develop sensible and cost-effective workflow to suit local needs and equip and educate staff for change
- we cannot look to LC for the lead any longer either in terms of cataloguing "rules" or subject access, many LC decisions, especially in terms of subject access are "subject" to the influence of lobby groups in the U.S.

13. We are not cataloguing individual items we are building collections

Perhaps the most important function of cataloguing and the catalogue is to create and express the collection

- we are moving towards one single collection – the be all and end all union catalogue
- how can we continue to serve the library users specific to our particular institution?
- can we have one catalogue that serves all class of library user from kids to researchers?
- if MARC was used to its fullest and ILS was designed to really take advantage of it then this might be possible, e.g. 008/22

14. Inadequate/incomplete teaching of cataloguing

Cataloguing is not a core requirement at many library schools.

- not enough time spent on cataloguing in the library science curriculum, however the education is there for those who seek it

- balance between teachers who are active cataloguers vs. academics
- more practicum opportunities needed so students have an opportunity to learn real world cataloguing
- training for paraprofessionals is rooted in old practices

15. Information retrieval experiences based on Internet search engines (Google, etc.) and other commercially oriented businesses (Amazon, etc.)

- these systems are designed to sell, this is browsing in a bad way
- similar to bookstores and malls, they want you to get lost and find something else to buy
- Not effective at handling unusual formats or materials

16. How do we <u>really</u> move away from the card catalogue?

Most of our standards, rules and codes are based on the card catalogue.

- MARC and cataloguing rules based on card technology
- cards forced people to browse exposed them to the system
- still not well understood by library users but perhaps better understood by librarians of the time
- divided into author/title and subject most people used the author/title catalogue still finding themselves lost in the sea of subject headings
- this is where 'main entry' was developed, literally the main card that held information found on the other shorter brief cards
- the 'primary access' point as adopted by RDA has nothing to do with 'main entry' but can provide a collocation point for shelf organization

17. Cataloguing is expensive, but what is the alternative?

If we can no longer do the kind of cataloguing work we have done in the past, what options are there?

- working with publishers – good but the role of publisher is also changing as they loose grip on the printed word
- working with other metadata producers
- producing minimal level records and pulling in enhanced metadata from other sources
- letting the web organize itself with little or no guidance
- librarians work the semantic web

18. Need an interdisciplinary think tank to workout out these problems

- need to mix librarians, cataloguers, programmers and experienced researchers, e.g. like the Access conference but year round
- OCLC would probably consider themselves as such, but they are way too expensive
- Better understanding of the work of cataloguing would lead to greater support for the profession and assist us to finding solutions to the problems we are facing
- Need to get away from the “we know best” attitude in regards to the organization of knowledge and information

- it could/should be more casual like an opensource enterprise: 'free' and everyone can do it because they love what they do
- most/many librarians have tended to shy away from these types of relationships: is this gender related?, female-dominated librarianship vs. male-dominated computer science or is this a false dichotomy?; more of an excuse than a truism perhaps