The Future of Cataloging

Address to the Ebsco Leadership Seminar
Boston, Massachusetts
January 16, 2005

Deanna B. Marcum
Associate Librarian for Library Services
Library of Congress

My career in librarianship has included work in cataloging, which I have always understood to be a major part of library functioning. But I didn't fully realize how major until I made a discovery when I became associate librarian of the Library of Congress. The discovery was financial—the Library of Congress is investing in cataloging at the rate of forty-four million dollars a year! You can well appreciate that a cost of that magnitude really got my attention.

If such an expenditure produces great benefits for the Library of Congress, libraries across the country, and others around the world, then we can justifiably argue that the forty-four million is well spent. But in the age of digital information, of Internet access, of electronic key-word searching, just how much do we need to continue to spend on carefully constructed catalogs? That is the question I have come here this evening to pose—how should we think about cataloging in the Age of Google?

I have not come to say that we no longer need the cataloger-produced bibliographic entry. I recognize that my own institution, the Library of Congress, created the bibliographic structure that is used by nearly every library in this country and by many around the world. Before starting any revolution against that structure, I want to take care to consider the potential consequences.

But I have many questions about cataloging , and I believe we must face them together and begin answering collectively. I therefore welcome the invitation to speak here as an opportunity to begin that discussion. I need your advice, your judgment, and that of others in the library and research communities to consider what the technologies that all of us are now adopting mean for cataloging in the future. I ask you to think of this evening as the first step in a longer exploration of a difficult issue.

Let me begin with a practical demonstration of the question's importance—an example of how digital-era students work.

Let us suppose that you are a librarian at a small college near the middle of the continental United States. Let us even suppose that yours is the library whose Web site I recently picked at random to see what digital resources it was offering. I am pleased to tell you that I was impressed. In addition to an electronically searchable catalog of your own physical holdings, I found that you offer fourteen EBSCOHost Online Databases, thirteen online databases from OCLC First Search, eleven InfoTrac Online Databases, five Lexis Nexis Online Databases, three Proquest Online Databases, and at least nine other online resources, including encyclopedias, dictionaries, electronic books, and materials for research on current issues. Consequently, users of your library have online access to literally hundreds of scholarly journals and other resources on all kinds of topics in a wide range of academic fields.

Now let us suppose that I am one of your college's students with a term paper coming due. And let's also suppose that I've been assigned to write about the foreign policy of President Fillmore. (I don't know why I am using this subject as an example, except that I can't get out of my mind the name of an amusing recording of political

ditties that a friend recently told me about. It's entitled "Sing Along With Millard Fillmore.")

Now, in the old days, I might have walked to your library, looked in an encyclopedia there for "Fillmore," then searched your paper card catalog to identify books on Fillmore, located these books by call number on a shelf, and looked through their tables of contents and maybe indices to find what they contained on foreign policy. But today I don't want to go to the library. I want to stay in my cozy dorm room, where I have a computer, which your college may even have provided me. So I decide to use it to do my research. One option, I find, is to do it through your library's Web site.

I click on your library's Web site (that is, on the Web site of the actual library that I selected). There I find the term "Online Catalog," and click on that. Then I see a menu of five aggregations of leased databases, identified by company names or as "other." Not knowing which aggregation will contain databases of use for research into Millard Fillmore, I click on one of the aggregations at random. There I find such references as *American National Biography*, *Encyclopedia Britannica*, and *World Book Encyclopedia*, which, I discover through more clicking, have information about Millard Fillmore. After clicking on each in turn, I find some short articles to use.

But I am trying to get an A, and therefore I want information in more depth on Fillmore's foreign policy. So I go back and click on another database-aggregation company button (I won't say which one), where I find another database menu. There, I click on a database from the History Resource Center, which provides access to full-text journals, reference articles, and historical documents, including—as I find out after a lot more clicking—some useful stuff about Millard Fillmore and foreign policy. Now maybe

I can write my paper. Note that I have not set foot in your library, or even checked your online catalog for print resources.

But today, for me as your student, there's an alternative to all this clicking, this navigating, that I have done on your site. I also have the option, sitting there in my cozy computer-equipped dorm room, of ignoring your library entirely, and going online to, say, a commercial search service such as Google. With Google, all I have to do is type my subject—"President Fillmore Foreign Policy"—into a search box and click on "Go." If I have used "Advanced Search" to get only references containing all four words, up will come what Google calls the first ten references out of literally thousands. I don't have to go through multiple organizational layers to get to something about Fillmore's foreign policy.

Never mind that the first five references include articles from Encarta and LookSmart that come with commercial advertisements. Never mind that the second reference is a sketch about Fillmore by, quote, "Caroline," last name not given, who turns out to be a Pocantico Hills School fifth grader. And never mind that the fifth reference gives some information on Fillmore from a decade-by-decade outline of events, provided by some unidentified individual who records, rather shakily, that he or she, quote, "tried to make all the information as accurate as possible." Through the LookSmart Directory, which is third on the list, I can get to articles from the Columbia Electronic Encyclopedia. And I may also find other material of real value in those thousands of references to my subject. So, is it any surprise that many students just go Googling instead of to the library, virtual or physical, and use whatever turns up first in the key-word search?

In fact, we already know from many studies that students—and other researchers—are going first to Google and other search services rather than to library catalogs. The Pew Internet and American Life Project has just published a new study of Internet use entitled, *Counting on the Internet*. The report says that more than sixty percent of Americans now have Internet access, and 40 percent have been online for more than three years. High percentages of Internet users expect the Web to contain information they need about such matters as news, health care, electronic commerce, and government services. In the words of the study, "most expect to find key information online, most find the information they seek, and many now turn to the Internet first."[1]

Similarly with students. Last November, participants in the most recent meeting of the American Society for Information Science and Technology heard a paper entitled, "'I Still Prefer Google': University Student Perceptions of Searching OPACs and the Web." The paper reported on a study of a group of graduate and undergraduate students who performed searches in Google and on a university OPAC. In the words of the report, ". . . while students were aware of the problems inherent in Web searching and of the many ways in which OPACs are more organized, they generally preferred Web searching . . . students were able to approach even the drawbacks of the Web—its clutter of irrelevant pages and the dubious authority of the results—in an enthusiastic and proactive manner, very different from the passive and ineffectual admiration they expressed for the OPAC." Why? Essentially, the study showed, because they found searching with Google easier.[2]

Earlier, OCLC published a "White Paper on the Information Habits of College Students," which focused on their use of campus library Web sites and other Web

resources. An abstract of the report says that “college and university students look to campus libraries and library websites for their information needs and value access to accurate, up-to-date information with easily identifiable authors. They are aware of the shortcomings of information available from the Web and of their needs for assistance in finding information in electronic or paper formats.” Inside, however, the report says that “the first-choice web resources for most of their assignments are search engines (such as Google or Alta Vista), web portals (such as MSN, AOL, or Yahoo!), and course-specific websites.” In fact, nearly eighty percent use search engines “for every assignment” or “for most assignments.” Four-fifths are bothered “at least a little” by ads on the Web, but nearly three-fifths believe “that there is no difference in the reliability of information on websites with advertising.”[3]

Recognizing that students—and many other information users—increasingly go to Google before going to a physical library for what they need, libraries and publishers are converting their print collections to digital formats so that high-quality, authentic resources will be electrically accessible. We librarians, particularly those who serve students, believe this is important for educational reasons. But as we develop digital resources, the question arises—do we need to provide detailed cataloging information for these digitized materials? Or can we think of Google as the catalog?

Not everything can be converted to digital form, of course. We have to recognize that many of the resources that libraries have digitized are old books and other items, mostly from the nineteenth century, on which copyright has expired. But this material is useful and being used—digital librarians are amazed at the extent to which some material, once put online, gets visited more than it ever did on library shelves. Moreover,

great amounts of newer material of value to students, teachers, scholars, and others, also are electronically available, by license, whether from commercial publishers of journals, ebooks, and databases, or from nonprofit archives such as JSTOR. More and more is being made accessible on the Web, where it is discoverable, and in many cases searchable, not through library catalogs but through electronic search boxes.

I have only begun to describe this phenomenon. Commercially driven technological developments are taking digital word-level indexing even further, in leaps of great magnitude. Major new developments have come to light just in the past few months. A subsidiary of Amazon.com called A9.com says it is in an early stage of developing what it calls "innovative technologies to improve search experience for e-commerce applications."[4] Amazon.com, itself, has unveiled a "Look Inside the Book" feature, which allows potential buyers to do key-word searches that turn up, not just authors, titles, or publishers, but also *excerpts* from digitized texts containing the key words sought, thus providing another way for you to identify books that you might wish to buy from Amazon. Amazon intends to provide this search function for thousands of books.

Even more recently and amazingly, you have all seen the announcement by Google that it is now going to digitize and make substantial parts of the contents of five major research libraries searchable using its key-word search engine. On December 14, Google announced, quote, "that it is working with the libraries of Harvard, Stanford, the University of Michigan, and Oxford University as well as the New York Public Library to digitally scan books from their collections so that users worldwide can search them in Google."[5] Google plans to underwrite, and contribute technical expertise to digitizing,

tens of thousands of pages daily at the libraries over a decade, through an agreement that covers more than fifteen million books and other documents.[6]

There's some fine print to note. Google is up against the same obstacles that confront digitizing libraries themselves, such as copyright. Google says its system will work as follows:

> *Users searching with Google will see links in their search results page when there are books relevant to their query. Clicking on a title delivers a Google Print page where users can browse the full text of public domain works and brief excerpts and/or bibliographic data of copyright material.*[7]

That is, unless publishers agree to inclusion of copyrighted materials, Google will not be able to offer access to entire libraries such as Harvard's or Stanford's. So the new development is still a long way from what an enthusiastic media reporter called "Google's goal" to "have everything at your fingertips, all the world's information digitized and instantly available . . . ."[8] But Google does plan to digitize much more of the participating libraries' collections than the libraries, themselves, have been able to do. Apparently Google is even working on ways to digitize and search collections of hand-written manuscripts.[9]

One must remember that Google is a business that must profitably finance all this. Those who search Google's new digital library will find advertisements along the way as well as what Google calls "buy-this-book links" to publishers who sell the books being searched or related titles. A reviewer for *Newsweek* has raised, quote, "the very big issue of how much we want the world's information transformed into a giant ad environment."[10]

But the participating libraries find ads of little if any concern in light of the great benefits that massive library digitization could have. As the president of the University of Michigan said, "We believe passionately that such universal access to the world's printed treasures is mission-critical for today's great public university."[11] And Michael Keller, Stanford University Librarian, explained as follows:

> *This is a great leap forward. We have been digitizing texts for years now to make them more accessible and searchable, but with books, as opposed to journals, such efforts have been severely limited in scope for both technical and financial reasons. The Google arrangement catapults our effective digital output from the boutique scale to the truly industrial.*"[12]

Now, think about all this in relation to cataloging. Such mass digitization, with its word-level indexing, forces us even more than previous digital developments have done to ask questions about the future of library cataloging as traditionally performed. When Google co-founder Larry Page says that users will be able to "browse the full text of public domain works and brief excerpts and/or bibliographic data," the latter may sound like catalog records, and the need for classification of works may continue. Indeed, a *New York Times* account of the libraries' Google agreement, which it calls a "major stride toward" the "long-predicted global virtual library," says Google plans to create a "digital card catalog" along with its "searchable library."[13] But what's happening now may well go beyond current cataloging for bibliographic control. Google has declared its "mission" to be nothing less than, quote, "to organize the world's information."[14]

Libraries have a long, proud history of cataloging. Our previous systems for classifying books and other forms of information, and for recording information about

them, including our creation of CIP and of MARC records—all undertaken for collections control and the benefit of our patrons—have been ingenious and effective. As Arlene Taylor of the Library and Information Science faculty at the University of Pittsburgh has said,

> *The objective of cataloging and indexing operations is the production of an intermediate-level summary description of the intellectual and physical characteristics of an information artifact. In the absence of digital full-length texts, these descriptions serve as the primary device by which people gain access to the contents of libraries."*[15]

But now, digital full-length texts are available. And thousands if not millions more of them are in prospect. Potentially, people will be able to search every word from a book's dust jacket to its back-of-the-book index. The need for intermediate-level descriptions will come under serious scrutiny. When library schools began at the end of the nineteenth century, cataloging had a central part in the curriculum. But then, so did handwriting.

My staff at the Library of Congress believes that, in providing access, there already has been a major shift. Cataloging now involves identifying metadata that already exist and taking advantage of existing description and access points. Different approaches are needed depending on whether resources are archived or linked and how long they will last. New hybrid systems take advantage of traditional library catalog information along with abstracting and indexing tools and online reference tools.

Staff members at the Library of Congress are experimenting with an "access level" record focused on access by subject, which is more useful for digital resources that

are constantly changing. Also within LC, we are looking at ways to take advantage of self-describing metadata in digital resources and to make more use of computer systems to capture bibliographic information from others. And we are working with publishers and with software vendors on the development of more useful metadata.

Additionally, we are rethinking who does what in cataloging. For example, with the advent of ever more automated sophistication, the detailed attention that we have been paying to descriptive cataloging may no longer be justified. If the task of descriptive cataloging could be assumed by technicians, then retooled catalogers could give more time to authority control, subject analysis, resource identification and evaluation, and collaboration with information technology units on automated applications and digitization projects. This coming spring, LC will host a small, informal gathering of managers of bibliographic operations at national libraries to review the assumptions underlying cataloging, and, I hope, to develop more cost-effective ways to meet future needs for bibliographic control.

But the future of cataloging is not something that the Library of Congress, or even the small library group with which we will meet, can or expects to resolve alone. We are eager to work with many relevant communities of librarians, publishers, and others to deal with cataloging issues. I hope that what I have said today will encourage you to join in an expanded discussion of that subject. In the discussion, the following seem to me the critical questions for all of us to face.

1. If the commonly available books and journals are accessible online,

   should we consider the search engines the primary means of access

   to them?

2. Massive digitization radically changes the nature of local libraries. Does it make sense to devote local efforts to the cataloging of unique materials only rather than the regular books and journals?
3. We have introduced our cataloging rules and the MARC format to libraries all over the world. How do we make massive changes without creating chaos?
4. And finally, a more specific question: Should we proceed with AACR 3 in light of a much-changed environment?

Whatever the answer to these questions, all of us in the library world must recognize that, in the future, the Internet is increasingly where people will go for information, whether from Google's library or to our own Web sites or both. Let me conclude with a light note about that point. When I, myself, was looking for something on the Web recently, I came across a press release from the Internet company Yahoo. The release reported the results of what it called an "Internet Deprivation Study"—I'm serious, or rather, they are!—designed to see how Web users would react to, quote, "life without the Internet." The study purported to find that users have such "emotional connections" to the Internet that "nearly half . . . indicated they could not go on without" it "for more than two weeks." And "the median time [that] respondents could go without being online," said the report, "is five days." "All participants," the study showed, "found living without the Internet more difficult than they expected, and in some cases impossible." They experienced what the report called "withdrawal and feelings of loss, frustration, and disconnectedness when cut off from the online world."[16]

Shall we dismiss all that as the self-serving, scientifically questionable hyperbole of an Internet company? Perhaps. But there's enough truth in it for us to get busy exploring and resolving questions such as those I have posed about the future of cataloging. The library, up until now, has been viewed as the place for reliable, authentic information. The catalog linked the users to vetted resources. Can we rethink cataloging to achieve something similar in the world of Google? I hope so.

Thank you for considering this issue with me.

[1] John B. Horrigan and Lee Rainie, Counting on the Internet (Washington, D.C.: Pew Internet & American Life Project, 2004), available at http://www.pewinternet.org/reports/toc.asp?Report=80.

[2] Karl V. Fast and D. Grant Campbell, "'I Still Prefer Google': University Student Perceptions of Searching OPACS and the Web," presentation to the ASIST 2004 Annual Meeting, Providence, R.I., Nov. 13-18, 2004, scheduled for publication in proceedings of the meeting; abstract available at http://www.asis.org/Conferences/AMO4/abstracts/137.html.

[3] "OCLC White Paper on the Information Habits of College Students," published electronically by the OCLC Online Computer Library Center, Inc., June 2002; available at http://www5.oclc.org/downloads/community/informationhabits.pdf.

[4] http://www.A9.com.

[5] "Google Checks Out Library Books," Google press release, http://www.google.com/intl/en/press/pressrel/print_library.html.

[6] John Markhoff and Edward Wyatt, "Technology; Google is Adding Major Libraries to its Database," *New York Times*, 14 Dec. 2004, Sec. A, p.1, col.6. Abstract available through http://query.nytimes.com/gst/abstract.html.

[7] Ibid.

[8] Steven Levy, "Google's Two Revolutions," *Newsweek*, available at http://www.msnbc.msn.com/id/6733225/site/newsweek/print/1/displaymode/1098/.

[9] Levy, Ibid.

[10] Levy, Ibid.

[11] Mary Sue Coleman, quoted in "Google Checks Out Library Books," press release, http://www.google.com/intl/en/press/pressrel/print_library.html.

[12] Michael Keller, quoted in "Stanford and Google to Make Library Books Available Online," news release, http://www.stanford.edu/dept/news/pr/2004/pr-google-011205.html.

[13] John Markhoff and Edward Wyatt, "Technology; Google is Adding Major Libraries to its Database," *New York Times*, 14 Dec. 2004, Sec. A, p.1, col.6. Abstract available through http://query.nytimes.com/gst/abstract.html.

[14] Google mission quoted in "Google Checks Out Library Books," press release, http://www.google.com/intl/en/press/pressrel/print_library.html.

[15] Arlene Taylor, [SOURCE?]

[16] "Yahoo! And OMD Reveal Study Depicting Life Without the Internet," Yahoo! Media Relations press release, New York, 22 Sept. 2004, available at http://docs.yahoo.com/docs/pr/release1183.html.