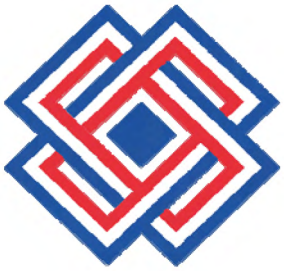

Mass Digitization: Implications for Information Policy

Report from
"Scholarship and Libraries in Transition:
A Dialogue about the Impacts of Mass Digitization Projects"
Symposium held on March 10-11, 2006
University of Michigan, Ann Arbor MI

U.S. National Commission on Libraries and Information Science (NCLIS)

May 9, 2006

NCLIS
U.S. National Commission on Libraries and Information Science
Phone: (202) 606-9200 • Fax: (202) 606-9203
E-Mail: info@nclis.gov • Web: www.nclis.gov
1800 M Street, NW • Suite 350 North Tower
Washington, DC 20036-5841

The National Commission on Libraries and Information Science is a permanent, independent agency of the federal government, established in 1970 with the enactment of Public Law 91-345. The Commission is charged with:

- Advising the President and the Congress on the implementation of policy
- Conducting studies, surveys, and analyses of the library and informational needs of the nation
- Appraising the adequacies and deficiencies of current library and information resources and services
- Developing overall plans for meeting national library and informational needs.

The Commission also advises Federal, state, and local governments, and other public and private organizations, regarding library and information sciences, including consultations on relevant treaties, international agreements and implementing legislation, and it promotes research and development activities that will extend and improve the national's library and information handling capability as essential links in the national and international networks.

Executive Summary

The project announced in December 2004 for a partnership between Google, Inc. and five major research libraries (the "G5") to digitize over 10 million unique titles launched a new era of large-scale digitization heretofore not imagined feasible or affordable. In the year or so since that announcement, many stakeholders have raised issues about the legal, social, economic, and other impacts of this and similar projects that will inevitably follow Google's lead. The project and the reactions to it inspired the idea of a public forum at the University of Michigan to allow scholars, librarians, publishers, government leaders, and others an opportunity to come together and discuss their concerns and issues. The symposium was held March 10-11, 2006. The Webcast of the entire symposium may be found on the symposium Web page: http://www.lib.umich.edu/mdp/symposium/.

The symposium was organized with a keynote and several other individual presentations, as well as five panels, each focusing on particular stakeholders or targets in digitization: libraries; research, teaching and learning; publishing; economics; and public policy. The individual speakers and panelists are listed at the end of this report.

Because of its responsibility to address the information and learning needs of the American people, NCLIS not only co-sponsored the symposium with the University of Michigan, it held its own meeting in conjunction with the symposium. At the end, the Commissioners summed up nine major issues that have information policy implications and connected them to key points made during the symposium.

The nine issues identified to have potential impact on national information policy are:

1. ***How should important aspects of copyright—fair use, orphan works, opt-in vs. opt-out models—be handled in digitization projects?***
   Copyright issues in digitization emerged as a major theme, with general consensus that laws need to be updated for the digital world. Google and the G5 libraries affirmed their intention to stay within the copyright law. However, publishers are concerned about the requirement in Google's program for publishers to opt-out if they do not want their materials included. Google believes that opt-out is allowable under fair use, and that the alternative model, opt-in, has large transaction costs that include search and negotiation, which are particularly time-consuming with orphan works.

2. ***Quality: When is the quality of OCR good enough? What about quality of content and authentication?***
   Because OCR (optical character recognition) introduces errors into the text, large-scale digitization may reverse the progress through the centuries toward increasingly accurate and high-quality printing. However, some feel that quality is improving—and in any case it will be *good enough.* They suggest that since we have the technology and now the resources from Google and others, we cannot slow down to

make things perfect. We need to "just do it," learn from mistakes, iterate the process, and make it better. The issue of authentication is of particular concern to government agencies. An electronic "watermark" is needed so that users will be able to trust the digital documents that GPO and NARA maintain.

3. ***What are the roles and priorities for libraries in the digital age?***
   Mass digitization will increase usage of libraries; the more information that is readily available about a collection, the more usage will increase. There is enormous opportunity for more digitization by libraries. They need to cooperate; it makes no sense to digitize the same thing more than once. Librarians need to develop new services and to transform collections space into new kinds of space—intellectual crossroads for working, learning, and teaching.

4. ***Who will assume long-term ownership of books and journals and other media? Who will take responsibility for long-term preservation of books and journals and other media, and preserving the public record?***
   Libraries have the commitment and are the only trusted agents for long-term preservation of digitized materials. Books are best "insured" by digitizing. The Federal government has a critical role in preserving government documents in perpetuity.

5. ***Standardization and interoperability: How can the silos of digital initiatives communicate with each other?***
   The rush to large-scale digitization may result in many individual and unique projects that have no way to communicate or search among them. Libraries are creating "silos" of data in digital repositories. Without standards for interoperability, searching many silos may be expensive and time-consuming, or even impossible.

6. ***What are the roles of publishers and booksellers in the digital age?***
   Despite the promise by proponents of mass digitization that it will drive additional usage of libraries and increase sales for publishers and bookstores, some in the book industry distrust Google. They believe that the "Dark Archive" may not be dark forever. Some publishers have no problem with the Google vision; an advantage of the project for them is making their backlist ("the long tail") more widely accessible, netting them more money from increased availability than they would get otherwise.

7. ***What business models are needed in the era of mass digitization? How will the open access movement affect the economics of digitization?***
   The business model for access to valuable information that has evolved is not "pay-per-view"—what has evolved instead is either free or advertiser-supported information. This model appears to be continuing with the Google and other mass digitization projects. Open access is another model promoted by some, but others question the sustainability of that model.

8. ***Information literacy: What should be done about information illiteracy?***

If students limit their research to only what they can retrieve though simple Web searches, they are not only missing key information, they are not learning vital advanced searching skills. Librarians, publishers, and authors need to raise information literacy in our society, especially the skills of students and scholars. Search engines should be "tuned" for different needs of users.

9. ***Assessment: What types of assessment are being used? How will we know if digitization and electronic access are meeting people's needs?***
MIT's survey of their campus in 2005 revealed that e-resources are very heavily used. People know that some high-value information may not be freely available, and they want help sorting through the chaos. They also want integration across sources. Ongoing market research will be necessary, including developing standard questions and time series and running the right experiments.

Under each of the nine areas, this report synthesizes the relevant comments made by the speakers at the symposium. The report concludes that the challenges ahead are many and finding workable solutions will be like fitting together puzzle pieces. The pieces include authors, scholars, publishers, libraries, associations, and government agencies. The solutions will involve education and awareness, policies, responsibility, standards, quality, cooperation, rights, sustainability, technology, and assessment.

Mass Digitization: Implications for Information Policy

Report from:
"Scholarship and Libraries in Transition:
A Dialogue about the Impacts of Mass Digitization Projects"

Symposium presented by
The University of Michigan Library and
The U.S. National Commission on Libraries and Information Science (NCLIS)
May 10-11, 2006
Ann Arbor, MI

INTRODUCTION

Mass digitization of books and other materials for the purposes of preserving them for future generations and making them available to a much wider audience than could ever access the physical objects in libraries, archives, and museums is not new. Ever since the Internet enabled graphics in 1995, libraries have been scanning their older documents and pictures to both preserve and provide access to them. Thousands of libraries of all sizes have scanned images, cataloged them, and made them available on the Web. Not just libraries, but also archives, museums, and publishers have all been involved with digitization projects for many years. According to the closing speaker, Clifford Lynch, some of these should be described as "large scale" rather than "mass" digitization, because projects of widely varying sizes have been done and will be done in the future.

The project announced in December 2004, however, for a partnership between Google, Inc. and five major research libraries—the University of Michigan, Harvard University, Stanford University, Oxford University, and the New York Public Library (commonly referred to as the "Google 5" or "G5")—to digitize over 10 million unique titles launched a new era of large-scale digitization heretofore not imagined feasible or affordable. In the year or so since that announcement, many stakeholders have expressed concerns and raised issues about the legal, social, economic, and other impacts of this and similar projects that will inevitably follow Google's lead.

The project and the reactions to it inspired the idea of a public forum at the University of Michigan, where the staff of the Library took the leadership in organizing, planning, running, and funding it. Nearly 500 people attended the symposium March 10-11, 2006 in Ann Arbor, Michigan. In an open and congenial forum, scholars, librarians, publishers, government leaders, and others discussed their concerns and issues. During the symposium, participants were able to participate in a blog. The blog and also the Webcast of the entire symposium (which was watched live by over 150 people) may be found on the symposium Web page: http://www.lib.umich.edu/mdp/symposium/.

Because of its responsibility to address the information and learning needs of the American people, NCLIS not only co-sponsored the symposium with the University of Michigan, it held its own meeting in conjunction with the symposium and most of the Commissioners were present throughout.

The symposium was organized with a keynote and several other individual presentations, as well as five panels, each focusing on particular stakeholders or targets in digitization: libraries; research, teaching and learning; publishing; economics; and public policy.

ISSUES IN DIGITIZATION FOR INFORMATION POLICY

Many topics raised significant issues that have information policy implications. A panelist, Barbara Allen, stated, "We have a window of opportunity before public policy is set on issues surrounding mass digitization." Therefore, following the symposium, the Commissioners summed up nine major issues and connected them to key points that speakers and participants made throughout the symposium.

> **"We have a window of opportunity before public policy is set on issues surrounding mass digitization."**

The nine issues and the key points related to each issue made by the speakers at the symposium are summarized here. Attribution to individual speakers is not given unless it is a direct quote or represents the views or policies of a particular individual or institution. The individual speakers and panelists are listed at the end of this report. Readers wishing to examine the forum content more closely may view the Webcast, which is available from the symposium Web page: http://www.lib.umich.edu/mdp/symposium/.

1. **How should important aspects of copyright—fair use, orphan works, opt-in vs. opt-out models—be handled in digitization projects?**

   Copyright issues in digitization emerged as a major theme of the symposium, with general consensus that many improvements in the copyright laws are needed. The problem is not that we have insufficient property protection; the problem is that we are deploying new protections at an accelerating pace—more and more protections around smaller and smaller things.

   Representatives of both the University of Michigan (President Mary Sue Coleman) and Google (Adam Smith, Senior Business Product Manager) emphasized strongly their intention to stay within the copyright law. Smith reiterated Google's goal: to create a comprehensive, searchable, virtual card

catalog of all books in all languages, while respecting copyright. Google is using the works to produce an index, not to produce a new product to compete with the original work. The index should increase the market potential for the digitized works.

In regard to the types of works involved in the Google 5 project, about 15% are out of copyright, in the public domain. For the 85% that are in copyright, about 20% are in print and available for sale via normal retail channels, and about 65% are out of print and available via used book sellers, libraries, document delivery and print-on-demand. It is this last group—those that are still under copyright but not in print—that will be most impacted by the Google Library Project. Nearly every book in America goes out of print within five years. Mass digitization will mean that nothing will ever go out of print. We need public policies to bring it into line with patent protection.

The least controversial area of copyright in regard to mass digitization is for works in the public domain. Early in the symposium, a member of the audience challenged the use of the phrase "falling into the public domain" and suggested that a better phrase would be "rising into the public domain." Later symposium speakers adopted that alternative expression.

The concept of public domain should cover not only books, but also such materials as sound recordings, manuscripts, special collections, and the huge number of images now available—for example, 50-year old amateur photographs. Google views books different ways. One is Full View, for public domain books, which represent about 20% of books. For books not in the public domain, Snippet View shows a limited number of pages. A snippet is mostly a pointer, to show the book's availability. No ads appear when snippets are displayed.

Although public domain means that someone cannot be sued for making a copy, it does not mean that the person who owns a copy of a work in the public domain has to make it easy for others to get it. Access and delivery are still issues for public domain materials. For example, museums set restrictions on photographing their artifacts that are in the public domain because they have to be *stewards* of the materials.

Fair use. An important concept in understanding how copyright relates to digitization is the "fair use" exclusion in U.S. copyright law. Fair use depends on the purpose and character of the use—whether the use is commercial; whether the use is for criticism, comment, news reporting, teaching, scholarship, or research; and whether the use is transformative vs. consumptive. It also depends on the nature of the copyrighted work (published vs. unpublished; factual vs. fictional), the amount and substantiality of the portion used (only a tiny section of content is allowed), and the effect of the potential market for/value of the work

Some legal cases that set precedent for the Google case were reviewed at the symposium. One is Kelly v Arriba Soft, 9th circuit 2003. Arriba Soft was an image search engine that made thumbnails of Kelly's photographs. The purpose of the use was commercial, but Arriba did not try to sell the works; the use was transformative. The amount of work copied was only a thumbnail, not a substitute for the work, and it was necessary to copy entire image to produce the thumbnail. The effect on the potential market was positive: it guided users to Kelly's photographs.

Orphan works. Another key concept in the copyright discussion is orphan works—copyrighted works whose owners may be impossible to identify and located. In a recent report (*Report on Orphan Works; a Report of the Register of Copyrights*, January 2006, available at: http://www.copyright.gov/orphan/), the U.S. Copyright Office stated that "there is good evidence that the orphan works problem is real and warrants attention." The orphan works problem is so huge because only 4% of books are in print, and 75% or more are in a "twilight zone"—they may be in print but they are not for sale because the rights have reverted to the author. Or, they may be in the public domain, but we do not know for certain—only 20% are known for certain to be in the public domain. The orphan works problem applies to all kinds of copyrightable things, not just books.

Tim O'Reilly, the keynote speaker suggested that digital rights management (DRM) should be done with a light touch—"more like a cat than a dog. When you take dogs to the vet you hold them tightly. When you take cats to the vet, you hold them *loosely*." DRM requires only delicate pressure on all parties.

In addition to other institutions and the information industry, Congress has a role in regard to orphan works—particularly in regard to revising the copyright law. Orphan works law will make it easier to discover which books are not under copyright and to read them online—"lifting all boats."

Solving the orphan works problem will also help collections in all parts of the world. For example, a conference was held recently in Alexandria Egypt about digitizing Arabic works. Libraries in the Arabic world are small and not well supported, and not accessible to most native speakers of Arabic. By digitizing the collections, many more Arabic-speaking Muslims will be able to discover their own heritage.

Opt-in; opt-out. An important distinction exists between the "partners program" and the "library program." In the former, the publishers (the rights holders) can *opt-in*: they can submit books that are in print, and then Google shows a few pages in response to queries and offers links to booksellers. In the "library program" on the other hand, Google scans in a complete collection of library books (which may be in or out of print), and a rights holder that does not wish to be included can *opt-out*. The controversy is about the library program. Publishers say that it violates copyright; they would prefer the opt-in option.

Google says that what they are doing is allowable under fair use; they prefer opt-out.

Copyright used to be entirely opt-in. This was changed in the 1970s when the default was reversed. In the Copyright Act of 1976, which still remains the primary basis of copyright law in the United States, the term of copyright was changed from a fixed period requiring renewal to an extended period based on the date of the creator's death. Everything that was "fixed" was protected by copyright. It was still a good idea to register, but an author did not have to register in order to be protected. When the law was changed, a lot of information was lost about copyright owners who no longer had to renew their copyrights after 28 years.

Opt-in and opt-out models have different transaction costs. The transaction costs with opt-in are huge, especially for orphan works. The costs include search costs to find the rights holder and then negotiation costs with the rights holder. Finding the rights holder can be difficult—especially if the publisher is out of business, has moved, has been acquired, or changed its name, or if the rights have been assigned to the author. If a publisher cannot find the copyright holder, he may go ahead and use the material, but at some risk. An enormous amount of time is spent clearing copyright, and the further one reaches back in time, the greater the transaction costs.

Google feels that the costs associated with the opt-out model are relatively small; they are basically the costs to the publisher to send an e-mail to Google or call an 800 number with book identification. However, from publishers' and authors' perspectives, opting-out can be difficult if many organizations—not just Google—are digitizing. Publishers and authors may not even know that someone is digitizing their works.

The Web has become a de facto "opt-out" – you can tell Web spiders not to crawl in a protected space. Google's Library Project uses the same model as the Web: fair use and opt-out. However, even though it is fair use, Google picked a Web model—where you sometimes cannot copy pages.

What will be the economic impact of the Google Library Project? It is difficult to predict, but publishers and authors should be asking: Will there be fewer books? Lower quality? Lower profits? Readers should ask: Will it be easier to find relevant books? Will there be a better search experience?

Publishers feel that the opt-out model would be very costly and benefit no one. Opt-in and opt-out are not legal concepts, they are economic models. What is needed is a rights clearinghouse to reduce the costs for everybody.

We can also reduce costs for everybody in the future if we minimize human intervention and have computers scan or copy works. If institutions are required

to do prior search and negotiation, huge transaction costs will be placed on the cataloging industry.

Copyright laws need to be updated for the digital world—you cannot have a market that works well if the rights are not established. We need to articulate the need for having the valued works available where people can find them…or they will not find them. What matters most is that there is a digital version.

2. Quality: When is the quality of OCR good enough? What about quality of content and authentication?

A general concern about large-scale digitization is that progress through the centuries toward increasingly accurate and high-quality printing may be reversed. Jean-Claude Guédon, Université de Montréal, noted that before printing became a visible mass phenomenon, the quality of a manuscript was tied to its genealogy—the quality and accuracy depended on who created it. When printing came in, printers grabbed whatever text they could get their hands on. They realized they needed to establish trustworthiness so they tried to grab three or four versions of the manuscript and resolve a single authoritative version. Out of that came the modern version of the reliable, authoritative text—this was the method to deal with the fact that documents do change in nature over time.

**We cannot slow down to make things perfect.
The rising tide will lift all boats.**

Optical character recognition (OCR), however, introduces errors into the text and so may be considered a step backward. Some feel that the quality is improving; Karin Wittenborg, University of Virginia, said it will be nearly perfect. "At least it will *be good enough."*

She also noted that there is pressure to rush digitization projects before everything is perfected and all the problems solved. We have the technology and now the resources from Google and others. Others are jumping in quickly—e.g., the European effort to digitize six million volumes. Students are also putting pressure on us. We cannot slow down to make things perfect. The rising tide will lift all boats.

Google's stance is similar: it is still in the early days and Google has limited resources too, and thus has to set priorities. Google wants to "get it out there"—to "just do it," learn from mistakes, iterate the process, and make it better. Adam Smith from Google said, "Do not let perfection be the enemy of the good."

Wikipedia (the online encyclopedia that is created collaboratively) is another example of a digitized resource that has omissions and mistakes. In Wikipedia, how does one tell which version is the good one? The one written last month? The current one? The one created ten years from now? Wikipedia is only a snapshot of the moment. Actually, Wikipedia is not a product or a thing, it is a *process*; it is an aggregating tool to create a community. It decreases the distance between text and people, and is surprisingly self-correcting over time.

Issues in authentication fall particularly on government agencies such as the Government Printing Office. According to Bruce James, the Public Printer, GPO and the National Archives and Records Administration (NARA) need some form of electronic "watermark" so that users will be able to trust the digital documents that GPO and NARA maintain.

3. What are the roles and priorities for libraries in the digital age?

Librarians know what scholars want and need and they know something about the magnitude of the information available. However, librarians are not going to be without competition—others will provide services and perhaps will do it better.

Nowadays, fewer people are coming into libraries, and as a result, reference and circulation statistics have dropped. However, interlibrary loan has increased 148% since 1991. Overall, it must be remembered that the campus library adds enormous value; it is a point of competition among universities. In the past, faculty and students demanded the books and the buildings. Libraries were funded as a public good. Mass digitization makes this local public good a mass public good. Once you pay for digitizing a local copy, it is "free" to additional users around the world. This changes the business model. Libraries no longer all have to have a copy of everything—it no longer matters who actually owns the book.

There already has been mass digitization of journals. A significant amount of content has appeared in digital formats since late 1990s and the perceived value is high. Very old journals are still used nowadays, but mainly in digital form. In fact, evidence is mounting that any material that is not available in digital form does not get used. Digitized information is how students work today—they hardly know any other way.

Our mental model of large-scale digitization has been shaped—perhaps unfortunately—by journal articles and article-finding tools (databases), which were a clear and significant improvement over print indexes. E-journals, however, could be absorbed by library practice more easily than e-books. To date, e-books have not achieved a similar status as e-journals.

Students, scholars, and researchers all over the world are relying on the ease and speed of digital access and are unearthing many new and rare treasures they never would have known about or found in print collections. Mary Sue Coleman, University of Michigan, called it "instant gratification of a one in a billion need." Even for material that is readily available, people are really *annoyed* if they have to go find a paper book, photocopy it, retype the relevant passage or quote. As library users' behaviors have changed, so have library expenditures.

Mass digitization will likely drive additional usage of libraries; the more information that is readily available about libraries' collections, the more collection use will increase.

The Google project will create an index; their efforts will complement not compete with libraries. In fact, mass digitization will likely drive additional usage of libraries; the more information that is readily available about libraries' collections, the more collection use will increase. Google-type projects will bring *big* resources to libraries, which never have enough funds.

Another myth is that we are going to digitize all of knowledge; there is a vast amount of out-of-copyright material that needs attention from librarians. There is enormous opportunity for more digitization by libraries. The OCLC database represents aggregate holdings of 32 million records worldwide. Nearly 40% are held uniquely by single institution and half were published before 1977. As libraries nowadays are buying fewer books, there are likely to be even more unique holdings. Librarians need to cooperate in digitization; it makes no sense to digitize the same thing more than once. Librarians should focus instead on increasing access to rich and deep unique treasure troves.

Librarians can also help by working with authors, publishers and users to agree on guidelines for digitization—how it is done matters very much to everyone. In addition, librarians can work with rights holders to improve rights management practices. They should also avoid duplication of effort, because there is a vast amount of out-of-copyright material that needs attention from librarians.

Stanford University Library intends to provide new services for readers, for example, to index the books taxonomically (by the ideas represented in the book), link citations from the footnotes in the books, and highlight places and names with links to short bios and related information. These projects are not harmful to commercial interests. They are, however, beneficial to the readers.

As far as space in libraries that may be freed up by digitization, Stanford is planning a bookless engineering library. They want collaborative spaces, group study space, more information services space, and an information commons to do

digital research. They envision fewer paraprofessionals and more professional staff.

Libraries need to be less warehouses and need to transform collections space into new kinds of space—*intellectual crossroads* for working, learning, teaching, and new types of programs.

Physical libraries are not going away; virtual and physical libraries will coexist. However, libraries need to be less warehouses and need to transform collections space into new kinds of space—*intellectual crossroads* for working, learning, teaching, and new types of programs.

Jean-Claude Guédon, Université de Montréal, offered two innovative ideas for libraries. The first is based on the fact that most doctoral theses have a chapter that is a review of the literature. If librarians would extract those chapters and make the links to references, it would produce a map of knowledge: "applied epistemology." His second idea is to create concordances of the books that are being digitized, identify the 100 rarest words in the books, and cluster similar books to create pathways of knowledge—real pathways through knowledge, to create something like H.G. Wells' *The World Brain.*

Research libraries have collected only 3 to 5% of books published—because librarians *select* them. The academy is based on selecting what to learn, winnowing from all the information/knowledge out there what is considered by credentialed people to be important. Librarians need to make sure that their readers know that libraries are more than bricks and mortar; libraries are an "ethereal ideal" that people carry in their minds.

4. Who will assume long-term ownership of books and journals and other media? Who will take responsibility for long-term preservation of books and journals and other media, and preserving the public record?

Mary Sue Coleman, President of the University of Michigan, using a reference from the local Detroit area, contrasted the fact that General Motors does not have to be able to manufacture a 1957 Chevy, or even maintain parts, but libraries have to preserve books from all periods. This is a serious challenge considering that over 20% of the University of Michigan's collection is brittle books and about half is on acidic paper. In stating Michigan's commitment to protecting the written word for all time, she quoted John Wilkin: "We believed in this forever." Michigan was digitizing *before* the Google project and will be doing it *after*.

"Only libraries have kept a long-term commitment to preservation."

The keynote speaker, Tim O'Reilly, noted that the World Wide Web developers did not think about preservation, and that the Internet Archive (Brewster Kahle's "Wayback Machine") does not go back far enough. Only libraries have kept a long-term commitment to preservation. Paul Courant, University of Michigan, agreed that Internet archiving is really sketchy compared with libraries. Courant asked, "Who is the trusted agent for the digital repository?" Then he answered his own question: "Librarians are the only ones who care."

Digitization is one of the best forms of insurance we have; it is not a replacement for the physical objects, but increasingly a good (albeit not perfect) surrogate.

Another question raised was how libraries set the value and insure collections of unique treasures (books and other materials). Clifford Lynch, in his wrap-up, suggested that instead of insurance, libraries should buy better environmental controls. Digitization is one of the best forms of insurance we have; it is not a replacement for the physical objects, but increasingly a good (albeit not perfect) surrogate that at least preserves the content.

All of the information created by the Federal government belongs to the people. The ability to sustain a democratic republic over time depends on knowing what your government is doing. The Federal government has had a proactive duty to make sure the information is widely available throughout the U.S. since 1813 when the Government Depository System was established. It now has 1250 partners, and 50 of them get a copy of everything produced by the Federal government. This system worked very well up until 1993. Then Congress ordered GPO to put all government documents on the Internet for free. GPO could also put on any other publications and charge whatever they wanted. But they quickly discovered that it cost too much to collect the payments. So they decided to put up almost everything (92%) for free (8% is not free—for example, maps).

Last year 50% of government documents were born digital and put on agencies' Web sites. As Public Printer, Bruce James' job is to save these documents in perpetuity. "In perpetuity" actually means the time that the United States will exist as a country. That could be 100 years or 500 years, but probably not 1000 years. James asked, "How many companies have been around 100 years?" James concluded that we must trust Federal government agencies to preserve

government documents rather than a company that likely will not be here in 100 years.

5. Standardization and interoperability: How can the silos of digital initiatives communicate with each other?

A widespread concern is that the rush to large-scale digitization may be creating a Tower of Babel, with too many individual and unique projects that have no way to communicate or search among them.

Campus-based digital repositories are a powerful opportunity for libraries, working with scholars, to create digital "containers" for scholars to "dump" their data. However, a problem is that libraries are creating "silos" of data in digital repositories, and scholars may have to search many silos to find what they need. Without standards for interoperability, the search may be expensive and time-consuming, or even impossible.

What is needed is an honest broker, an arbiter with authority to create some sort of clearinghouse.

6. What are the roles of publishers and booksellers in the digital age?

"Book publishing is very high-risk, labor-intensive—it is a broken system." This strong statement by Ann Wolpert, MIT Libraries, underscored the observation that of all the players in mass digitization, publishers and booksellers appear to have the most anxiety about their future roles—or even their future existence in the digital world.

Mark Sandler, University of Michigan Library, said that the term "publish" originally meant "to make public." Now, anyone can be a publisher: bloggers, librarians, nine-year old kids, or even Tim O'Reilly's cat. Truman Capote once put down Jack Kerouac by saying: "That's not writing, that's typing." Now we can say: "That's not publishing, that's scanning." Sandler asked, "What does it really mean to be a publisher?"

Writer and historian of technology Ed Tenner opined that the very problems with publishing are what make it so attractive. The problems that make it so crazy make it *irresistible*. Tenner said, "The heart's desire of every blogger is to have a book—it shows that you're loved by somebody." John King, Dean of the School of Information, University of Michigan, countered by saying that since he became a dean, the wall of books in his office has become a decorating element only.

Publishers are extremely concerned about the Dark Archive—it may not be dark forever.

Some hold a pessimistic view of publishing in the digital age despite the assertion by proponents of mass digitization that it will drive additional usage of libraries and additional sales for publishers and bookstores. At the symposium, the G5 libraries and Google affirmed their commitment to follow the law. They will not give access to copyrighted materials; they will keep them in a Dark Archive. However, some publishers have a fundamental distrust of Google; they believe there will be leaks of full digital copies and the digital copies will become more valuable than Google says they will. Publishers are extremely concerned about the Dark Archive—it may not be dark forever. Google could eventually become a competitor. Google could drive down prices below what would be sustainable. In a word, mass digitization may kill publishers. Publishers ask, "Without publishers who will service the real information needs of scholars?"

Hal Varian, University of California, Berkeley, responded by saying that Google tends not to own anything—they are in the business of indexing the world's knowledge. There will not be a new security problem that does not already exist on the Internet—even now you can find a digital copy of Harry Potter if you look hard enough.

An apparent advantage of the Google project for publishers is that their backlist will be more widely accessible. So why are some publishers suing Google for making their backlists available and getting publishers more money from increased availability than they ever would get otherwise? These publishers believe that they could make their backlist more accessible through "print on demand." Print on demand is obviously becoming a valuable service, and it is becoming exploited in the marketplace—more and more content is becoming widely accessible. As far as "bind on demand," Karl Pohrt, owner of Shaman Drum Bookshop in Ann Arbor, said that he is part of a consortium of booksellers who have a machine to print books in the store. The technology, however, is not fully developed yet. Nonetheless, he would love to do it.

Some publishers have no problems in principle with the Google vision. For them, Google Print is a good thing as publishers can negotiate terms. Mechanisms exist for licensing the long tail (the backlist) of publishers through one-stop shops, but Google has not yet availed itself of these. However, the concern of Alicia Wise, Publishers Licensing Society (UK), is about concentrating so much content in the hands of one player. She questions how well understood is the complexity of the market in which Google operates. Google has competitors in indexing, search, advertising, etc. Publishers have relationships with them all—it is a very complex market. Wise asserted that publishers are not Luddites; they understand the potential for new readers to find out about published works, and that Google must

index the full text for them to be found. However, Google does not have the right to store publishers' works and make them accessible if publishers choose to publish them in other ways.

Defenders of the critical contributions that publishers make to scholarship emphasized that publishers are not the bad people here. Publishers try to fill needs. However, publishers need to make some money back so they can cover their costs.

For general needs, the "good enough" online search is *good enough* to answer simple questions. In the era of mass digitization, however, publishers will be working with targeted audiences, with more complex needs. "Good enough" will not be good enough.

What do publishers feel that they do best? For general needs, the "good enough" online search is *good enough*—to answer simple questions. In the era of mass digitization, however, publishers will be working with targeted audiences, with more complex needs. "Good enough" will not be good enough. Suzanne BeDell, ProQuest Information and Learning, cited a ProQuest product: *Historical Documents*. She said that it is extremely tedious to work with fragile documents such as parliamentary papers and old newspapers that are fragile, dense, dirty, complex, and often have no headlines. An enormous amount of clean-up is needed; "It's a job for humans." Publishers have to be close to their market and assess what is needed. BeDell believes that publishers will not undertake mass digitization; they will do *targeted* digitization; they will have to provide real value to specific users; they will fill specific gaps. For publishers, it is not about what they *can* digitize, but what they sh*ould* digitize. Librarians have never been about what is easy to find, but what is important to find, and publishers will support them in this.

Wise reviewed some aspects of digitization that are old, including the *vision* of getting everyone access to what they want to read; the *legal framework* (digitization falls under the two spheres of copyright law and contract law); and the *real costs* (creativity, distribution, marketing are not without costs and these must be met in some way).

On the other hand, she said, some aspects are new: the *technology* (we can digitize faster and realistically put everything online, and more works are born digital); the *funding* (increasingly fragmented, more now from commercial sources than from foundations or government); and the *stakeholders* (diverse and

including everyone involved in the information/entertainment 'value chains' from creator to user).

Some aspects of digitization can be borrowed, even though they are in their infancy. These include *business models* (e.g., iTunes with inexpensive per track downloads) and *digitization standards* (although it is not clear where 15 years' worth of experience in digitizing books is captured).

Keeping on the same theme of something old, something new, something borrowed, Wise next discussed what makes publishers blue. Copyright is complex, roles and responsibilities are changing, costs are high, economies are sluggish, and technology does not often work as well as it should. We need standards for content, rights, metadata, and access management, and they do not yet exist. Her final worries: too little funding and so many books!

Wise's vision for digitization includes convenient and affordable access to resources whenever and wherever individuals want it; the drive for convenience will fuel innovation in online services; a powerful yet invisible infrastructure will be in place to enable personalization and ensure security and privacy online; information and literature will be freely accessible (but not free of charge) and accessible in socially responsible ways, taking account of freedom of expression, ability to pay, and the environment. Wise said that most publishers try to be socially responsible and take the moral high ground. The HINARI (Health InterNetwork Access to Research Initiative) project is one example: through the combined contributions of over 70 publishers, HINARI enables local, not-for-profit institutions in developing countries to get free or very low cost online access to the major journals in biomedicine.

In the future publishers will, as they do now: listen to authors, readers and customers; work in partnerships; digitize their own content and professionally manage their own assets; ensure that they use copyright to stimulate creativity and innovation; protect authors' rights and their own rights to choose how works are published; and negotiate robust terms to ensure sustainability.

7. What business models are needed in the era of mass digitization? How will the open access movement affect the economics of digitization?

In early days of Internet, it was assumed that access to valuable information would be a "pay per drink" or "pay-per-view" model, even though that would make access to information unaffordable for some. What has evolved instead is either free or advertiser-supported information. This model appears to be continuing with the Google and other mass digitization projects.

Google has taken the advertising model and evolved it in an extremely efficient way. Google's business model has been to reduce barriers, and its success is in

marrying content with advertisers, which creates synergy for customers. Google's business model is simple: create a lot of value to users.

> Google is stepping forward to do it and to take the risks. "This is why the Google Library Project *matters.*"

Viable and sustainable technological innovations such as iPods do not spring forth suddenly without a period of experimentation during which an economic model is developed. For example, iPods and the selling of billions of songs would not exist today without Napster. However, the economic model is much harder to develop for books because users are not helping to build the ecology as they did with music. Google is stepping forward to do it and to take the risks. According to Tim O'Reilly, "This is why the Google Library Project *matters.*"

Open access. Many inside and outside the publishing field think that open access sounds exactly like publishing, and question the sustainability of that model. If all of this is becoming a public good, who is going to pay for it? Brian Kahin, Computer & Communications Industry Association, said that initiatives such as the Open Content Alliance consortium sound "like a big playpen with no rules. I don't see how it will work."

Supporters of the Open Content Alliance say that it fits in the digital world in a variety of ways. It is building a collection of openly accessible information. The University of California is trying to scale up to digitizing 5000 books a month, of largely out of copyright materials. The materials need to be held by a trusted third party in an archive—so it will be held in perpetuity. An open services definition is needed—to learn how the materials are made available so tools can be built on top of it. Collection support tools are also needed—what is in the collection and how can I use it? Ultimately it is about building trust in the collection.

On the other hand, a lot of the value in Google is its vast amount of content, which is not true for the Open Content Alliance.

In any case, faculties are become increasingly aware about open access. According to Daniel Greenstein, California Digital Library, "That genie is out of the bottle!" Professors are becoming interested in managing their own copyright, and in the economics of scholarly publishing. They are asking librarians for details on the cost of information.

Senators Lieberman and Cochrane are legislating for NIH and CDC and other agencies to provide open public access to research funded by the government. Publishers are trying to figure out how to exploit these developments by building a business case for added value.

Information is becoming a commodity and a utility. Ultimately, we must consider: Who can do this most cheaply?

Google Print is a new kind of structure—a proprietary infrastructure. Google has a strong positive image, unlike Microsoft. Right now it does not display advertising on pages that show Snippets. However, publishers are worried that Google will change and start adding advertising. The deepest fear of publishers is that Google has the attention of the economy. Unlike publishers, Google does not have to compete with its backlist (the long tail). It makes money in the current flow. Google has managed to find a way to make money off the long tail. Publishers want a piece of that. It is a revenue stream that nobody ever saw coming.

8. Information literacy: What should be done about information illiteracy?

Ed Tenner, author, editor, and historian, began his presentation by reading a headline from a London newspaper: "Tutors despair at illiterate freshers." When TV watching is down, and Web use is up, why are students so information illiterate? Why can't they read, synthesize, and evaluate complex text? Tenner sought to illuminate this puzzle by describing his search on the term "world history" in Google. He retrieved articles on specific topics, but none that were an overview of what the field of World History is all about. A Wikipedia entry proved to be skewed and missing some key information; it was not a good introduction to the field. There were no entries for the topic at all in *Online Britannica*.

If students are limiting their research to only what they can retrieve though simple, "good enough" searches on the Web, they are not learning vital advanced searching skills. In the 21st century, "good enough" isn't.

If students are limiting their research to only what they can retrieve though simple, "good enough" searches on the Web, they are not only missing key information, they are not learning vital advanced searching skills. In the 21st century, "good enough" isn't.

What can librarians, publishers, and authors do about improving information literacy in our society, especially the skills of students and scholars? Two

suggestions were offered: academic sites should appear before other sites in results lists, and search engines should be "tuned" for different needs of users.

9. Assessment: What types of assessment are being used? How will we know if digitization and electronic access are meeting people's needs?

Ann J. Wolpert reported that MIT Libraries has been working with Google Scholar to develop and assess it. Can you create an environment with these large-scale products that will serve the scholarly community? How do they compare to proprietary databases? (At MIT, Google Scholar gives hot links to citations within the proprietary databases that MIT subscribes to, so that students do not have to go in and out of them).

MIT Libraries surveyed faculty, students, and researchers in late 2005. They learned that where e-resources are available, people vote with their *mice*: 85% regularly use online resources. However, the digitized resources themselves rank lower than the digital finding tools in importance.

What respondents want next are (1) Single interface to search across a variety of information sources; (2) Expanded online content, especially for older materials; (3) More access to all library material via commercial search engines; and (4) A "wizard" to help choose the best tools for a topic.

We can learn from these responses that people want help sorting through the chaos, people know that some high-value information may not be freely available, and integration across sources is a priority. Ongoing market research will be necessary, including developing standard questions and time series and running the right experiments.

OTHER ISSUES

Other issues were touched on at the symposium, but not developed very far. For example, the digital divide is still very much a reality. Policy needs to be developed in regard to access. Who will get access to digital materials? Will everybody take part in the digital revolution? What about the underserved or unserved? What will happen to those who are left out? Will the Open Content Alliance make much of a difference? Karin Wittenborg, University of Virginia, was optimistic on this score; she believes that inexpensive portable devices will be developed to read and interact with books and access will be affordable to everyone.

Another area related to access focused on people in developing countries. Even in areas where it is necessary to drive 20 miles to get access to information in digital form, people will do it, because it means at least they can get access—without digitized information,

they have no access at all. Michael Keller, Stanford University, said, "We need to support bright young researchers from developing countries. They will bring bright ideas. How can we not support them?"

CONCLUSIONS

Areas where work is needed to develop policies and practices for the 21st century world of mass digitization include:

- Copyright law needs to be updated for the digital world.
- At the same time that digitization projects are moving forward, quality and authentication need to be improved and preserved.
- Libraries have enormous opportunities for more digitization, but need to cooperate to focus their projects on unique and rare materials.
- Libraries have the responsibility to keep collections and preserve them for the long term; governments have the responsibility to preserve public documents in perpetuity.
- Standards for interoperability and cross-searching of digital repositories are needed to avoid "silos" that cannot communicate with each other.
- The value added by publishers and booksellers needs to be preserved, especially where they are addressing the needs of targeted audiences.
- Alternatives to the advertiser model, such as the open access model, need to be explored, especially in regard to sustainability.
- Students' and scholars' information literacy skills need to be improved.
- Coordinated ongoing assessment and market research is needed in order to understand changing user needs and preferences.

Overall, the challenges ahead are many and finding workable solutions will be like fitting together pieces of a puzzle. The pieces include authors, scholars, publishers, libraries, associations, and government agencies. The larger puzzle has elements of education and awareness, policies, responsibility, standards, quality, cooperation, rights, sustainability, technology, and assessment.

Appendix: Speakers and affiliations (in order of appearance on the symposium schedule)

Welcome
Beth Fitzsimmons, Ann Arbor, Michigan (NCLIS Chairman)
Brenda Johnson & ***John Wilkin***, Associate University Librarians, University of Michigan

Opening Remarks
Mary Sue Coleman, President, University of Michigan

Keynote Speaker
Tim O'Reilly, Founder & CEO, O'Reilly Media

Panel Session: Libraries
Josie Parker (moderator), Director, Ann Arbor District Library
Barbara Allen, Director, Committee on Institutional Cooperation
Michael Keller, University Librarian, Stanford University
Karin Wittenborg, University Librarian, University of Virginia

Panel Session: Research, Teaching & Learning
John King (moderator), Dean, School of Information, University of Michigan
Jean-Claude Guédon, University of Montreal
Ed Tenner, Professor & Author, Princeton University
Ann Wolpert, Director of the Libraries, MIT

Panel Session: Publishing
Mark Sandler (moderator), Collection Development Officer, University of Michigan
Suzanne BeDell, Vice President, ProQuest Information and Learning
Daniel Greenstein, University Librarian and Executive Director, California Digital Library
Alicia Wise, Chief Executive, Publishers Licensing Society

Special Presentation
Adam Smith, Google

Panel Session: Economics
Ron Milne (moderator), Acting Director of University Library Services & Bodley's Librarian, Oxford
Paul Courant, Professor, University of Michigan
Karl Pohrt, Owner, Shaman Drum Bookshop
Hal Varian, Professor, University of California, Berkeley

Panel Session: Public Policy
Nancy Davenport (moderator), President, Council on Library and Information Resources
James Hilton, Associate Provost for Academic, Information and Instructional Technology Affairs and Interim University Librarian, University of Michigan
Bruce James, Chief Executive Officer, U.S. Government Printing Office
Brian Kahin, Senior Fellow, Computer & Communications Industry Association & Adjunct Professor, University of Michigan

Closing Remarks
Clifford Lynch, President, Coalition for Networked Information