asis&t
The Information Society for the Information Age

sign in | contact us | site index

About ASIS&T Membership Conferences Publications SIGS & Chapters Careers E-Mail Lists

Go

Home > Publications > Bulletin > October/November 2006

ARIST
Bulletin
JASIST
Conference Proceedings
Digital Library
Online Bookstore

***Bulletin*, October/November 2006**

Impacts of Mass Digitization Projects on Libraries and Information Policy

by Trudi Bellardo Hahn

Trudi Bellardo Hahn is affiliated with the University of Maryland, College of Information Studies. She can be reached at thahn<at>umd.edu.

This article summarizes highlights from a symposium presented by the University of Michigan Library and the U.S. National Commission on Libraries and Information Science (NCLIS). The title of the symposium was "Scholarship and Libraries in Transition: A Dialogue about the Impacts of Mass Digitization Projects." The author, former executive director of NCLIS, prepared and NCLIS published a 24-page report that is available at www.nclis.gov. – Editor

Digitization – whether "mass" or just "large-scale" – of books and other materials is not new. Whether for the purpose of preserving them for future generations or making them available to a much wider audience than could ever access the physical objects, some libraries, archives, museums and publishers have been scanning their older documents and pictures for many years. Thousands of libraries of all sizes have scanned images, cataloged them and made them available on the Web.

Nonetheless, the project announced in December 2004 for a partnership between Google, Inc., and five research libraries (the "Google 5") – the University of Michigan, Harvard University, Stanford University, Oxford University and the New York Public Library – to digitize millions of unique titles launched a new era of large-scale digitization heretofore not imagined feasible or affordable. Since that announcement, many stakeholders have expressed concerns and raised issues about the legal, social, economic and other impacts of this and similar projects.

At a March 2006 symposium at the University of Michigan, scholars, librarians, publishers, government leaders and others discussed these concerns and issues, addressing five stakeholders or targets in digitization: libraries; research, teaching and learning; publishing; economics; and public policy. The Webcast of the symposium is at www.lib.umich.edu/mdp/symposium/.

Digitization Issues for Information Policy

Many topics raised significant issues that have information policy implications. Panelist Barbara Allen stated, "We have a window of opportunity before public policy is set on issues surrounding mass digitization." The key questions for information policy clustered into four areas of discussion.

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.

Mary Sue Coleman, president of the University of Michigan, and Adam Smith, senior business product manager at Google, emphasized their

Articles in this Issue

An Ecological Approach to the Design of Information Systems

We Don't Do That Here: Using Cognitive Work Analysis to Learn About Organizational Culture

Information Sharing of an Informatics Student Design Team

Using CWA to Frame an Investigation into the Use of Corporate Body Publications

Contextual Analysis for the Design of Controlled Vocabularies

Impacts of Mass Digitization Projects on Libraries and Information Policy

Changing Approaches to Metadata at bbc.co.uk: From Chaos to Control and Then Letting Go Again

Visibility of Student Member Involvement in ASIS&T Leadership Roles

Selected Abstracts from *JASIST*

Editor's Desktop

Inside ASIS&T

intention to stay within copyright law. Google's stated goal is 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 create a new product to compete with the original. 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 booksellers, 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. Nearly every book in America goes out of print within five years. Mass digitization will mean that nothing will ever go out of print.

The least controversial area of copyright in regard to mass digitization is for works in the public domain. At the symposium, it was suggested that the phrase "falling into the public domain" be changed to "rising into the public domain." 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 these 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 it is commercial; whether it 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, the amount and of the portion used and the effect of the potential market for or value of the work.

One legal case that set precedent for the Google case was 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 an 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 locate. In a recent report (*Report on Orphan Works: A Report of the Register of Copyrights*, January 2006, available at 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 huge because only 4% of books are in print, and more than 75% 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. 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.

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," 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. In the Copyright Act of 1976, however, the copyright term 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; 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 searching to find the rights holder and then negotiating with the rights holder. Finding the rights holder can be difficult if the publisher is out of business, has moved, has been acquired, changed its name or if the rights have been assigned to the author.

Google believes that the costs associated with opting-out are comparatively small – basically the costs to the publisher to notify Google with book identification. For publishers and authors, however, opting-out can be difficult if many organizations – not just Google – are digitizing. They may not even know that someone is digitizing their works.

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. Copyright laws need to be updated for the digital world – you cannot have a market that works well if rights are not established.

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 mass phenomenon, the quality of a manuscript was tied to its genealogy – quality and accuracy depended on who created it. When printing came in, printers grabbed whatever text they could get their hands on. They needed to establish trustworthiness so they tried to find 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.

Optical character recognition (OCR), however, introduces errors into the text and so may be considered a step backward. On the other hand, 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 from European competitors, student users and others to do digitization now, before everything is perfected and all the problems solved. We have the technology and now the resources from Google and others. We cannot slow down to make things perfect.

Google also wants to “just do it” – learn from mistakes, iterate the process and make it better. Google’s Adam Smith said, “Do not let perfection be the enemy of the good.”

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

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.

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 Google 5 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. Publishers are extremely concerned that the Dark Archive may not be dark forever. Google might eventually become a competitor and 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 at Berkeley, says that Google tends not to own anything. Google is 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.

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 tedious to work with fragile documents such as parliamentary papers and old newspapers that are fragile, dense, dirty and complex and often have no headlines. An enormous amount of clean-up is needed. BeDell believes that publishers will not undertake mass digitization; they will do *targeted* digitization; they will have to provide value to specific users; they will fill specific gaps.

Alicia Wise, Publisher’s Licensing Society (UK), reviewed some old aspects of digitization, 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 and 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 put everything online, and more works are born digital); the funding (more 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.

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

In the Internet’s early days, 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.

Viable and sustainable technological innovations 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 they question the sustainability of that model. If all of this is becoming a public good, who is going to pay for it?

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 – largely out of copyright materials.

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

Digitization Issues for Libraries

Issues and concerns for libraries in the digital age fell under three areas.

1. 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. Fewer people are coming into libraries nowadays, however, and reference and circulation statistics have dropped. Nonetheless, overall 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. Libraries no longer have to have a copy of everything – it no longer matters who actually owns the book.

A significant number of journals have already appeared in digital formats 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. 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. Even for material that is readily available, people are annoyed if they have to go find a book, photocopy it, retype the relevant passage or quote. As library users' behaviors have changed, so have library expenditures.

The Google project will create an index; their efforts will complement, not compete with libraries. In fact, mass digitization will drive more usage of libraries; the more information that is readily available about collections, the more collection use will increase.

A vast amount of out-of-copyright material needs attention from librarians. 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. Librarians must cooperate in digitization; it makes no sense to digitize the same thing more than once. Librarians should focus instead on increasing access to rich unique treasure troves.

2. 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 for preserving the public record?

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 a high percentage of library collections are brittle books or on acidic paper.

The symposium keynoter, Tim O'Reilly, noted that 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 sketchy compared with libraries. Librarians should be the trusted agents for digital repositories – they are the ones who care most.

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 imperfect) surrogate that at least preserves the content.

3. 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, 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.

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. Will everybody take part in the digital revolution? The problem is especially acute in developing countries. What about the underserved or unserved? What will happen to those who are left out?

Another area of concern was information literacy, especially among college students who seem blissfully unaware of materials and information that they cannot find easily on the Web. Students are limiting their searches to only what they can retrieve though simple, "good enough" searches. They are not only missing key information, they are not learning advanced searching skills. In the 21st century, "good enough," isn't.

Assessment was another area of concern. How will we know if digitization and electronic access are meeting people's needs? MIT Libraries surveyed their faculty, students and researchers and learned that where e-resources are available, people vote with their mice: 85% regularly use online resources. Respondents expressed desires for

- A single interface to search across a variety of information sources (a way to sort through the present chaos);
- Expanded online content, especially for older materials;
- More access to all library material via commercial search engines; and
- A "wizard" to help choose the best tools for a topic.

Ann J. Wolpert, director of MIT Libraries, called for ongoing market research, including developing standard questions and time series and running the right experiments.

Conclusions

Work is needed to develop policies and practices for the 21st century world of mass digitization in the areas of copyright law, quality control, deciding which works libraries should digitize, preservation of digital collections, standards for interoperability and cross-searching of digital repositories, clarification of the roles and value added by publishers and booksellers (especially where they are addressing the needs of targeted audiences) and sustainability of alternatives to the advertiser model, such as the open access model.

Overall, many challenges lie ahead 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 puzzle has elements of education and awareness, policies, responsibility, standards, quality, cooperation, rights, sustainability, technology and assessment.