

Running Head: BOON AND BURDEN

Boon and Burden: Digitization in America

Jeffrey E. Collins

The Information Landscape: Emerging Perspectives

2nd Annual Graduate Student Symposium

School of Information Resources and Library Science

The University of Arizona

18 November 2006

Abstract

Digitization is a necessary burden that librarians and other information professionals must bear in order to prove their relevance and maintain a critical link with society. Two recent examples of digitization projects, Google's Print Library Project and Yahoo!'s Open Content Alliance, are utilized as case studies to illuminate the complex issues associated with digitization.

Boon and Burden: Digitization in America [SLIDE]

The prevalence of computers and the Internet have led to an influx of digitization projects throughout the United States and elsewhere in recent years. These endeavors, comprised of formats including text, image, and sound, can be very beneficial to society. However, they also raise important questions in regard to intellectual property, preservation, and the library science field. What effects will the digitization of text have on the role of the library in American society? Will digitized items supplant access to the cultural inheritance of America or will it foster a renaissance of knowledge? While there are no obvious answers, it is clear that librarians and other information professionals have a responsibility to provide universal access to knowledge. It is imperative for libraries to prove their relevance and maintain a critical link with society by understanding and embracing digitization. Further, by collaborating with diverse entities on digital projects, libraries may most effectively preserve access to the cultural record.

The Process of Digitization [SLIDE]

In order to understand the current trends and ramifications of digital projects in America, one must understand the digitization process. Digitization is the method of “converting data to digital format for processing by a computer.”¹ This is currently accomplished by utilizing a machine to convert an item, such as text or image, into binary signals for subsequent use on a computer. Essentially, the digitization process takes information from one format (analog) and turns it into another format (digital). This development also involves the storage and maintenance of the newly digitized item, as well as the consequent migration or transfer from

one medium of technology to another. Also of note, in the digital environment, items can be transmitted and manipulated in a multitude of unique ways. [SLIDE] Prominent examples of current book digitization programs in the United States include Project Gutenberg, and more recently, Google's Print Library Project, Yahoo's Open Content Alliance, and Microsoft's MSN Book Search.

Digital Project Challenges [SLIDE]

There are many advantages and disadvantages to converting analog items into digital format. Despite the deficiencies inherent in the process, the switch from analog to digital is not only beneficial to information seekers and providers, but is inevitable. As a result, librarians and information professionals need to embrace this change, as well as position themselves to be the purveyors of information and protect access to the cultural heritage of America for the future.

Information professionals must contend with several challenges that arise due to the digitization process. These include legal considerations, policy and guidelines, and technological issues, among others. It is important to briefly analyze these issues in order to gain insight into the complexities of the role the digital format will play in libraries, archives, and elsewhere.

Technology is currently the most problematic issue facing these programs. Chief among technological concerns is the method by which the artifact or item is digitally obtained. The choice in equipment and software are both of paramount interest for the digitizer.² The cost and functionality of the software, as well as the physical capacity of the equipment, are critical components. The virtual size of the item is another important concern. Not only does it take more space to store a larger image on a computer or server, but it also takes the user longer to

download the object. Another technological challenge with the digital format deals with the subsequent storage of the artifact. The question of where the item is physically stored and who has access to the item is essential. Finally, there will invariably be technological advances, and as a result, storage mediums change on a regular basis. Accordingly, it is imperative to plan for the migration or transfer of files from one digital format to another over the ensuing years.³

Guidelines, policies, and procedures are issues currently facing digitization projects in the United States. Specifically, creating a digital library collection can be a very daunting task due to the inconsistencies in format, the high cost involved, and the fear of the unknown.

Regrettably, there is very little uniformity among projects in America, and confusion abounds. As a result, policy and guidelines are entirely dependent upon the specific project and institution. Consequently, interoperability and resource sharing is often disregarded in the short-term. This, in turn, will create future problems and will do little to benefit the information seeker or the library professional.⁴

Another current issue critical to the understanding of digitization in America is preservation. Digital preservation can briefly be defined as ensuring the longevity of a digital item. Essentially, it is the process of maintaining a useable form of the item in the future. This includes items that are both born digital as well as items that are converted from the analog format. The largest challenge to digital preservation involves the frailty of the technological format. Hardware and software usually have a life cycle of about three years. Thus, thoughtful planning is absolutely necessary in order to effectively preserve the digital item.⁵

Types of Digitization Programs [SLIDE]

Modern digitization projects in the United States are diverse and vary largely upon the type of institution. Academic libraries are presently the most prominent type of library involved in the process. These libraries tend to have access to more resources, such as historical artifacts and rare books, receive federal funding, and also serve the scholarly community. As a result, the demand for access to information in a convenient format is often higher in an academic library than in other types of libraries. In stark contrast, school libraries and media centers have largely avoided these types of projects due to limited resources. Instead, the focus has been on providing useful links to digital resources for students, while building school library collections in-house, if at all. Public libraries have a very diverse user base, and as a result, generally only focus on the preservation of local materials for digital projects. Specialized libraries also digitize material, but depending on the type of institution, there is often little to no allocation of funds for the digitization process. In many ways, archival institutions have led the way for digitization projects in America. During the last decade, many archival repositories have begun to make finding aids and collections available online.

Each of these projects varies in scope, depending on the type of institution, but all share similar characteristics of the digital process.⁶ Many programs are currently in use or in the planning stages in the United States. These programs are diverse and multifaceted, often dealing with historical artifacts, image, and text. An established and well-documented example of these types of programs is the Library of Congress American Memory Project.⁷

While archival materials and limited text-based digitization has been the norm, a more recent development, the full-fledged digitization of books, will have lasting ramifications on the

library profession. For the first time, entire books have been digitized on a massive scale. The breadth and scope of these projects has also dramatically increased during the last several years. Although Project Gutenberg was one of the first projects to place the full text version of books on the Internet, and now has more than 15,000 books online, the floodgates have been opened. Recent examples of book digitization programs include the On-line Books Page, the Internet Public Library, the Google Print Project, the Open Content Alliance, and the MSN Book Search program. While the majority of these projects have been conducted in the United States, other institutions and countries from around the world have recently begun the book digitization process as well. A response to the perceived Americanization and control of digitized text, the European Library, is one such project.⁸

Google Book Search [SLIDE]

In October 2004, Google, a for-profit corporation based in Mountain View, California, announced a controversial plan known as the Google Print Project. In essence, the company proposed to create a digital and searchable database of copyrighted and non-copyrighted books available on the Internet. The project has serious implications for the organization, dissemination, and use of information in the United States and elsewhere.

Google is currently digitizing massive amounts of books for use in a searchable full-text database on the Internet. This process includes materials in the public and private domain and raises significant questions about copyright laws and the future role of libraries. The Google Print Project is composed of two distinct sections: the Print Publisher Program and the Google Book Search. [SLIDE] The Print Publisher Program is an effort by Google to digitally scan

material by first obtaining permission from the copyright holder.⁹ This project has been largely unopposed, because it is done in an entirely legal manner.

Google Book Search is another story. As part of this endeavor, Google is scanning books from six libraries in the United States and two international libraries. [SLIDE] Google is then making the text of the books searchable online through a proprietary search tool only accessible through Google's Internet Web site. More importantly, this process is being done without the permission of the copyright owner. According to Google, the resulting search will reveal the entire text [SLIDE] a limited view [SLIDE] a "snippet" of the text [SLIDE] or information about the book similar to a library catalog.¹⁰ [SLIDE]

[SLIDE] Not surprisingly, this project is highly controversial and has attracted the ire of many in the publishing industry. For example, the Authors Guild and the Association of American Publishers filed a federal lawsuit in October 2005 charging Google with copyright infringement.¹¹ This legislation is still pending. Several significant issues arise from the Google Book Search program. These include copyright law, the fair use doctrine, and implied license. More importantly, this program raises questions as to the future role of libraries and archives.

The success of the Google Book Search rests largely on the precept of the implied license and fair use doctrine.¹² Many institutions of learning utilize fair use in order to legally duplicate limited amounts of copyrighted material, but it has previously not come under close scrutiny at such a high level. Another contention by Google, as well as other Internet search engines, is that they have implied license to copy materials that are posted to the Internet because the creators of Web sites want the Web site to be found by search engines.¹³ To their credit, Google does offer copyright owners an option to opt out of their program. However, rather than assuming that copyright owners want their property to be included in Google's project and giving them the

opportunity to opt out, Google should have to establish an opt-in policy for those that actually want Google to digitize their copyrighted work. In any case, Google must prove that the purpose of this project is to promote arts and science for the good of all under existing copyright laws in the United States.

Open Content Alliance [SLIDE]

Another example of a recent book digitization program is the Open Content Alliance. Shortly after Google announced its plans, Yahoo publicized a related program known as the Open Content Alliance (OCA).¹⁴ The Open Content Alliance is a nonprofit consortium of a large number of library, publishing, and Internet entities.¹⁵ This group is comprised of more than thirty institutions and is administered by the Internet Archive. [SLIDE] In effect, the Open Content Alliance scans text materials that are currently in the public domain, and subsequently makes it available on the Internet.

[SLIDE] Although similar in many respects, the Open Content Alliance differs from the Google Book Search program in several key areas. In a marked difference, the Open Content Alliance includes copyrighted material from publishers who opt-in to the program, unlike Google which copies the item regardless of potential copyright law infringement. Subsequently, this enables the OCA to circumvent much of the controversy surrounding the Google Book Search program. Additionally, full-text materials accessed from the Google Book Search are only viewable in pdf format, whereas files from the OCA include everything from raw images to metadata associated with the file. Also, the Open Content Alliance does not have proprietary search methods like Google Book Search. In fact, any common Internet search engine will be able to access the digitized text, even Google.¹⁶ Most importantly, the Open Content Alliance

utilizes an open-source software development model. This emulates the success of the Internet Archive and is important because nearly anyone with a computer and Internet access can utilize the system. The Open Content Alliance is truly a remarkable coalition and may revolutionize the book digitization process.¹⁷

Ramifications [SLIDE]

The current trends and future ramifications of digitization in America are numerous and multifaceted. In many cases, the digitization of various forms of content has changed the way individuals seek and obtain information. It has also increased the demand for ubiquitous access to knowledge. For centuries, physical books have been the most accessible and prevalent form of information technology. However, the advent of computers and the Internet, as well as the resulting digitization of text and image-based resources have raised serious questions regarding the existence of the library profession. Most prominently, these include the lasting value of the book, the way libraries store, obtain, and preserve information, and more broadly, the purpose of the library as a cultural institution.

As a case study, Google Book Search and the Open Content Alliance demonstrate many of the serious issues currently facing libraries in the United States. The rising use of digital technologies during the last decade has forced librarians to reevaluate the role they play in American society. Due, in part, to the digitization of information resources as well as increased access to the Internet, many individuals now argue that libraries are becoming obsolete. Therefore, libraries need to ensure they maintain relevance to their patrons by reacting to the changing environment of digital projects in innovative and unique ways.

Librarians have a responsibility to provide universal access to information. Libraries, public and government research institutions, museums, and archives should work together to perpetuate access to America's cultural heritage. The digitization process can be good for humanity, as it further contributes to public awareness and education. Rather than a single private corporation undertaking to digitize copyrighted text regardless of the legal ramifications, or even a single library undertaking to digitize text without the necessary financial support, much can be gained by partnering together to digitize materials. Utilizing innovative partnerships, such as the Open Content Alliance, is an important step toward librarians maintaining this critical link with society while ensuring that everyone has unfettered access to information.

Librarians must decide if outsourcing information resources to corporations such as Google in the short-term is worth the potential long-term costs. Quite simply, there is no guarantee that an eight-year old corporation known as Google will exist in the future. On the other hand, libraries are vital cultural institutions that will stay alive and thrive in the future. There are contentions from within and out of the library field that portend the end of the library sometime in the next ten years.¹⁸ It is our responsibility to ensure that does not happen. [SLIDE]

References

- American Memory*. 26 Apr. 2006 <<http://memory.loc.gov/ammem/index.html>>.
- Band, Jonathan. "The Google Print Library Project: A Copyright Analysis." *E-Commerce Law and Policy* Aug. 2005: 7.8.
- Barnsley, Victoria. "Publishing in the Digital Age." *Bookseller* 20 Jan. 2006: 5213, 21.
- Bartenbach, Bill. "Scenes from U.K.'s Past." *Information Today* Jan. 2006: 23.1, 36.
- Bone, Allison. "Google Faces Euro Challenge." *Bookseller* 25 Nov. 2005, issue 5206: 6.
- Creative Commons*. 14 Apr. 2006 <<http://creativecommons.org/>>.
- Dames, Matthew K. "Library Organizations Should Support Google Book Search." *Online* Mar./Apr. 2006, 30.2: 18-19.
- Dye, Jessica. "Scanning the Stacks." *EContent* January/February 2006, 29.1: 32-37.
- Elias, Stephen, and Richard Stim. *Patent, Copyright and Trademark: An Intellectual Property Desk Reference*. Berkeley: Nolo, 2004.
- Rose, Mark. "Literary Property Determined." *The Book History Reader*. David Finkelstein and Alistair McCleery, eds. London: Routledge, 2005. 231-240.
- Fishman, Stephen. *The Copyright Handbook: How to Protect and Use Written Works*. Berkeley: Nolo, 2004.
- Goldsborough, Reid. "The Brave New World of Book Research." *Teacher Librarian* Feb. 2006, 33.3: 49.
- Google*. 21 Apr. 2006 <<http://www.google.com>>.
- Helm, Burt. "A New Page in Google's Books Fight." *Business Week Online* Jun. 2005, 22.24.
- Internet Archive*. 18 Apr. 2006 <<http://www.archive.org/>>.
- Labi, Aisha. "Online." *Chronicle of Higher Education* May 2005, 51.36.

- LaGuardia, Cheryl. "The World in a Database." *Library Journal* Apr. 2005, 1306.
- Liu, Yan Quan. "Best Practices, Standards and Techniques for Digitizing Library Materials: A Snapshot of Library Digitization Practices in the USA." *Online Information Review* Oct. 2004, 38: 338-345.
- Lottman, Herbert. "Google: Man the Barricades!" *Bookseller* Mar. 2005, 5171.
- Kniffel, Leonard, and Gordon Flagg. "Library of Congress Gets \$3 Million from Google." *American Libraries* Jan. 2006, 37.1: 17.
- Kupferschmid, Keith. "Are Authors and Publishers Getting Scroogled?" *Information Today* Dec. 2005: 22.11.
- Marcum, Deanna B., et al. "Google at the Gate." *American Libraries* Mar. 2005, 36.3: 40-43.
- Miller, Arthur R., and Michael H. Davis. *Intellectual Property: Patents, Trademarks, and Copyright in a Nutshell*. St. Paul: West, 2000.
- Oder, Norman. "Google Launches Librarian Newsletter." *Library Journal* Feb. 2006, 131.2: 22. *Online Dictionary for Library and Information Science*. 13 Mar. 2006 <<http://lu.com/odlis/index.cfm>>.
- Open Content Alliance*. 21 Apr. 2006 <<http://opencontentalliance.org>>.
- Quint, Barbara. "Tick, Tock." *Searcher* Feb. 2005, 13.2: 5-6.
- Rogers, Michael. "European Digital Library in 2010?" *Library Journal* 1 Apr. 2006, 131.6: 28.
- St. Lifer, Evan. "Guiding the Googlers." *School Library Journal* Jan. 2005, 51.1: 11.
- Seltzer, Leon E. *Exemptions and Fair Use in Copyright*. Cambridge: Harvard U P, 1979.
- Stielow, Frederick. *Building Digital Archives, Descriptions, and Displays: A How-to-Do-it Manual for Archivists and Librarians*. New York: Neal-Schuman, 2003.
- United States Copyright Office. 2 May 2006 <<http://www.copyright.gov>>.

United States Congress of the United States Congressional Budget Office. *Copyright Issues in Digital Media*. Washington: GPO, 2004.

Warner, Julian. "Information Society or Cash Nexus? A Study of the United States as a Copyright Haven." *Journal of the American Society for Information Science* 1999, 50: 461-470.

Zeitchik, Steven, and Jim Milliot. "Google Draws Fire, Creates Book Page." *Publishers Weekly* 30 May 2005, 252.22.

¹ Online Dictionary for Library and Information Science (ODLIS), available from <<http://lu.com/odlis/index.cfm>>.

² *Ibid.*, 341.

³ Frederick Stielow, *Building Digital Archives, Descriptions, and Displays: A How-to-Do-it Manual for Archivists and Librarians* (New York: Neal-Schuman Publishers, 2003).

⁴ Liu, 340.

⁵ Stielow, 83.

⁶ Liu, 339.

⁷ The Library of Congress American Memory, available from <<http://memory.loc.gov/ammem/index.html>>.

⁸ Allison Bone, "Google Faces Euro Challenge," *Bookseller*, Nov. 2005, 5206, 6.

⁹ Keith Kupferschmid, "Are Authors and Publishers Getting Scroogled?" *Information Today*, December 2005: 22.11.

¹⁰ Matthew K. Dames, "Library Organizations Should Support Google Book Search," *Online* Mar. /Apr. 2006, 30.2: 18.

¹¹ George H. Pike, "Google Print and the Fair Use Doctrine," *Information Today*, December 2005:22.10.

¹² ODLIS available from <<http://lu.com/odlis/index.cfm>>.

¹³ Jonathan Band, "The Google Print Library Project: A Copyright Analysis," *E- Commerce Law and Policy*, August 2005: 7.8.

¹⁴ Reid Goldsborough, "The Brave New World of Book Research," *Teacher Librarian*, Feb. 2006: 33.3, 49.

¹⁵ Jessica Dye, "Scanning the Stacks," *EContent*, Jan. /Feb. 2006, 29.1, 36.

¹⁶ *Ibid.*, 33.

¹⁷ While beyond the scope of this paper, another excellent example of a book digitization program that bears mention is the European Digital Library. See Michael Rogers, "European Digital Library in 2010?" *Library Journal*, 1 Apr. 2006, 131.6, 28, among others.

¹⁸ Barbara Quint, "Tick, Tock," *Searcher*, Feb. 2005, 13.2, 5.