Mass digitization and its impact on interlending and document supply

Perry Willett, Digital Preservation Project Manager, California Digital Library

Abstract

Purpose: The paper describes the complex landscape of mass digitization projects and their impact on interlending and document supply. The amount of library material available online is staggering, but still at very early stages in terms of discovery tools. Without a centralized source to search digitized collections from these projects, it is important to understand their goals and distinctions between the services they provide.

Design: The paper analyzes the history, goals and achievements of the mass digitization projects. It also looks at the sites that aggregate digitized materials from these projects and their interfaces for providing access.

Findings: The amount of library collections now digitized has reached numbers unimaginable even just a few years ago. The impact on interlending and document supply will likely be limited in the near term, since much of interlibrary borrowing is for materials published in the past decade, and much of the digitized content from this period is restricted in Google Book Search, the only mass digitization project to undertake digital conversion of copyrighted material. End users will be able to discover materials of interest in new ways, but are likely to need or want to see the print source for recent publications.

Practical implications: Library staff members need to understand these projects and how they differ from one another in order to provide optimal service to their readers.

What is original/value of paper: The paper is a guide to negotiating the landscape of mass digitization, with an overview of the projects and their goals and accomplishments, with analysis of the impact on interlending and document supply.

Keywords: Google Book Search, Internet Archive, HathiTrust, Million Book Project

Paper type: Research, an analysis of mass digitization projects.

Introduction

The number of books and journals digitized in the past few years and now available online has grown exponentially, with corporations such as Yahoo!, Microsoft, and of course Google, funding projects to digitize hundreds of thousands, even millions, of volumes. The idea of having millions of digitized books online was unthinkable even 7-8 years ago, yet that has happened and the amount of digitized materials will continue to grow as a result of these mass digitization projects. The impact of these projects on libraries and interlending and document supply (ILDS) is still very unclear, although

some predictions are possible based upon an understanding of the scope of these projects and trends in ILDS, requiring a close look at the mass digitization projects themselves.

1 Mass Digitization Projects

1.1. Google Book Search (GBS)

In October 2004, Google announced its partnership with several large publishers to digitize works (Holman 2004). Known initially as Google Print, it later changed to Google Book Search. In December 2004, Google announced a partnership with five research libraries in the US and UK to digitize millions of books and journals (O'Sullivan and Smith, 2004). Since then, the library program within Google Book Search has grown to include more than 30 research libraries around the world ("Library partners."). This is a different database from Google Scholar, and as discussed below, the two are not yet integrated.

1.2. Open Content Alliance (OCA)

In October 2005, Yahoo!, the Internet Archive, and several research libraries and archives including the University of California and University of Toronto, announced a project called the Open Content Alliance to digitize hundreds of thousands of books in the public domain (Young, 2005). Microsoft also joined this project, although they have since pulled out. Over 80 research libraries and archives have since joined this effort. It is not clear how many volumes have been digitized , but the Internet Archive claims to have over 1.2 million volumes online "posted under OCA principles" although they host collections digitized through other projects.

1.3. The Million Book Project

The Million Book Project, or Universal Library, was the first mass digitization project. Founded by Carnegie Mellon University in 2001 and funded by the (US) National Science Foundation along with government support from India and China, this project has digitized over 1.5 million volumes, including large numbers from partner libraries in China, India, Egypt, and other countries. The majority of books are in Chinese languages, with almost one million, and over 350,000 books in English. Books from this project are available through the Internet Archive as well as its own website, the Universal Digital Library, although the latter has not been updated since 2007.

1.4. Other projects

Other mass digitization projects include Gallica, from the Bibliothèque Nationale de France, and an effort by the US Government Printing Office to digitize the complete run of US government documents. This latter project, announced in August 2008, is still in the planning stages.

2. The Aggregators

In some cases, digitized content from these projects is available from multiple sites, with different access tools.

2.1. Google Book Search

GBS provides access to the materials Google has scanned through the GBS website and has several important features:

- the ability to download entire public domain books in a single PDF file
- links to OpenWorldCat for end users to discover the closest library holding a copy of the book
- the owning library of the digitized copy
- links to new booksellers such as Amazon and Barnes & Noble, and sellers or used books such as AbeBooks and Alibris

There are some difficulties in using GBS, particularly for multi-volume sets and journals, discussed below.

2.2. HathiTrust

The HathiTrust, launched in October 2008, is a consortium of 24 US research libraries (Albanese, 2008). These libraries are all participants in GBS, and several also participate in OCA. They have agreed to store copies of their digitized collections in the HathiTrust, which acts as both a preservation digital archive for these files, as well as an access point. The HathiTrust uses the Michigan OPAC, known as "Mirlyn," as the only discovery tool for information about availability. There is no full-text searching of the entire digitized corpus available, although one is in the works; only bibliographic searches through Mirlyn are possible. End users can create and store personal collections, and are able to conduct full-text searches across these collections.

It is important to note that items may be freely available in HathiTrust that are restricted in GBS, and vice versa. The University of Michigan has taken the lead in determining copyright status of the items in HathiTrust, for instance researching copyright status for US publications. They have determined that thousands of volumes published after 1922 are actually in the public domain. HathiTrust provides access to these materials, while many are restricted in GBS. US Government documents are another body of publications for which HathiTrust is providing open access, while GBS is at present, in many cases, restricting access. The only way to discover this is to search in both systems.

2.3. Internet Archive (IA)

The Internet Archive provides access to a wide range of materials, from archived websites, to audio and video, to books scanned through the OCA, Million Book Project, and other digitization efforts. For books, IA provides several formats in which to view any given volume, including PDF, DJVU, text-only, and their own "flip book" format, which has both two-up and one-up page displays. End users can order print-on-demand copies through Lulu, a print on demand service.

2.4. Other sites

- Europeana is aggregating content from digitization projects in European libraries, museums, and archives. There are no statistics on the site, but the impression from searching is that the majority of the collections are images, with only a few texts.

- The Library of Congress has been digitizing collections for over a decade, and they have a number of significant collections including “American Memory,” and “Chronicling America” for newspapers.
- JSTOR contains the back files of hundreds of important journals, but is a subscription service.

These millions of online volumes give unprecedented access to books and journals. These projects have attracted a great deal of attention in the popular press. There are a wide range of issues surrounding these projects, particularly concerning rights, but this article will focus on the impact they will have on interlending and document supply. In order to understand the potential impact, we first must understand what is available.

3. Availability

3.1. Books.

It is important to understand what is, and what is not, available as a result of these projects. OCA is the simpler of the two main mass digitization efforts to describe. Its goal is to digitize materials in the public domain. Therefore, the books that are available were published prior to 1923 for US publications. For non-US publications, information other than the date of publication is used to determine copyright, generally involving a period of time after the author’s death, with 70 years the most widely adopted. This means that the dates for non-US publications in the public domain can vary widely depending on the author’s lifetime. End users can view several different formats of any given book, can read it on the IA website or download and save a copy of a PDF or a full-text version from uncorrected OCR [1]. As these books are in the public domain, there are few restrictions on what end users can do with them.

The GBS collection has much more complexity concerning rights, as it includes both public domain and copyrighted material. While many library partners have agreed to only allow Google to digitize public domain materials from their collections, the University of Michigan and the University of California libraries are allowing Google to digitize their entire collections. In addition, Google is working with many publishers to digitize their backlists. Google allows end users to view and read the entire volume for items they determine to be in the public domain. As of February 2009, 1.5 million volumes were available as full-text in GBS, out of a total of seven million digitized (Ratnakar *et al.*, 2009). For copyrighted materials, Google at present only allows access to “snippets,” two- to three-line excerpts. Under the proposed legal agreement between Google and the Authors Guild and the American Association of Publishers, end users will be able to view up to 20 percent of copyrighted materials while end users at subscribing libraries will be able to view entire books still in copyright; however some items will be restricted to the several pages allowed for access (Drummond, 2008; “Google Book Search Settlement Agreement”). Public libraries in the US will be allowed access through a single workstation. It is unknown at the present how many libraries will subscribe to this service, or whether this service will be available outside the US.

Given the volume of materials they are processing, GBS is relying on bibliographic metadata from partner libraries to determine copyright status. Also, given the differences

in copyright law, end users in the US may have access to items that are restricted for end users outside the US. Without going into the complex formulas and changes in copyright laws involved, suffice it to say that there are likely to be many materials restricted in GBS that are actually in the public domain. It would however require research to determine the true copyright status of many of the books. Even with the legal settlement that is currently proposed, the situation is likely to remain unsettled for some time.

As mentioned above, the University of Michigan is undertaking its own copyright research for materials in HathiTrust. In 2008, they were awarded a grant from the Institute of Museum and Library Services to develop a copyright review management system for copyright evidence that could widen this effort to other libraries (IMLS News and Events, 2008). In some cases, HathiTrust provides access to materials that are restricted in GBS.

3.2. Journals and Newspapers

Google Book Search announced in December 2008 an effort to include magazines in GBS. It appears from the descriptions and press releases that these are largely popular magazines such as *Ebony*, *Life*, *New York*, *Popular Science*, although it does include some research journals such as the *Bulletin of Atomic Scientists* (Foulser, 2008). It is not clear which other titles are included in this effort. Of course, Google is also digitizing journals from library collections. Google has also digitized newspapers, but they are part of Google News Archive, not GBS. It is also not clear from the Google News Archive interface which newspapers are available, or in which date ranges.

3.3. Media

Audio and video digitization is not dominated by a few large projects as text digitization is, and the rights issues surrounding items in these formats are more complex. IA contains a large collection of audio and video, collected from other sources, but these typically do not include the commercially-produced films and recordings available in media collections in many libraries. Interlibrary loan for media is controversial already, and it is unlikely that digitization efforts in this area will have any impact on interlibrary borrowing.

4. A complex landscape

This is a complex landscape for users and librarians, in which much remains unsettled and services are still in nascent stages. There is no central service or discovery tool for an end user to discover whether a particular book or journal has been digitized by one of these projects. The GBS interface for known-item searches, particularly multi-volume sets or journals, is particularly unhelpful. The strength of GBS, as with Google in general, is of course in keyword searches of the entire corpus.

Another complexity has to do with the differences in copyright law, particularly with differences in interpretations between US and non-US copyright law. Both Google and HathiTrust take the location of the end user as well as date of publication into account in determining whether to grant access. Without going into the details, there will be books, particularly those published in the 19th century through the early 20th century, that will be

accessible by end users in the U.S. but not accessible by those outside the US. This will provide complications for library staff in interlending and reference services trying to help end users in different countries.

Until a service such as OCLC or some other utility has access to all the records from GBS, OCA, and other mass digitization projects, there will be no central location for discovery even by simple bibliographic searches for author or title. Library staff will need to be mindful of the existence of these projects and the likelihood that any given title requested might be included, by searching at these sites directly.

Unfortunately, the discovery tools available at Google Book Search are less than optimal. The main discovery tool for Google is keyword search. Other types of search are available, such as author or title search, but overall precision is lacking. Google favors keyword searches, perhaps in the belief that this is how the majority of people want to search, and that known item searches are far less frequent; and in this they may be correct (Wildemuth and O'Neill, 1995). However, one time-honored research strategy is to follow footnotes from one article to the next, requiring known-item searches. More importantly for the scope of this article, known-item searches remain important for library staff, particularly those trying to fulfill interlibrary loan requests.

An example will address the complex searching environment: try searching for the *Journal of the Royal Geographical Society of London*, vol. 23, 1853 which exists in Google Book Search:
http://books.google.com/books?vid=uom39015010945783

Searching, even using the advanced interface by title, returns links to only two "full-view" volumes; the first is mislabeled as 1880, but is really v. 10, 1841, with the second from 1860. Many other volumes of this journal are available online in "full-view," but it is not clear why they are not retrieved by this search.

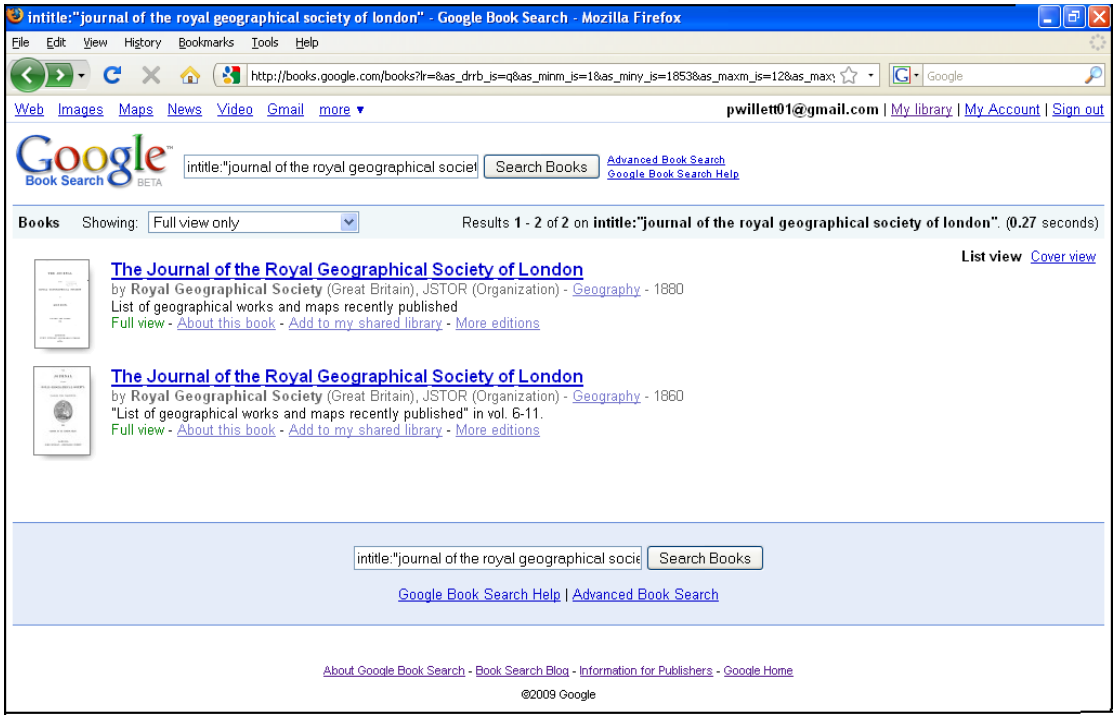

Figure 1

Searching in the HathiTrust via Mirlyn works much better for known-item searches. Of course, this is an online library catalog, so it was designed for this purpose. The University of Michigan library has linked each volume for journals and multi-volume sets from the holding screen for its titles to online versions in both GBS and HathiTrust. Seeking the same volume of the *Journal of the Geographical Society of London*, the user will note that this title is also available from JSTOR. [Figure 2] However, clicking on the link labeled “Online links to individual volumes” brings the end user to a list of holdings. Those volumes that are online have links to the copy at HathiTrust and GBS. [Figure 3]

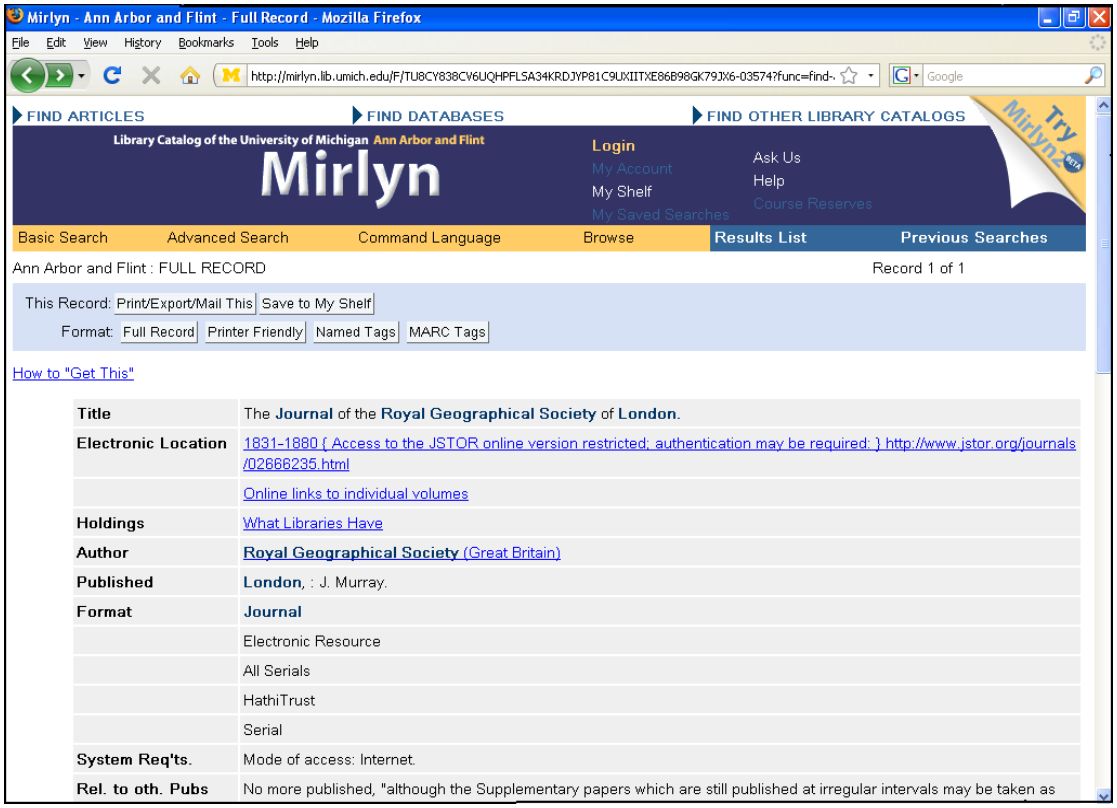

Figure 2

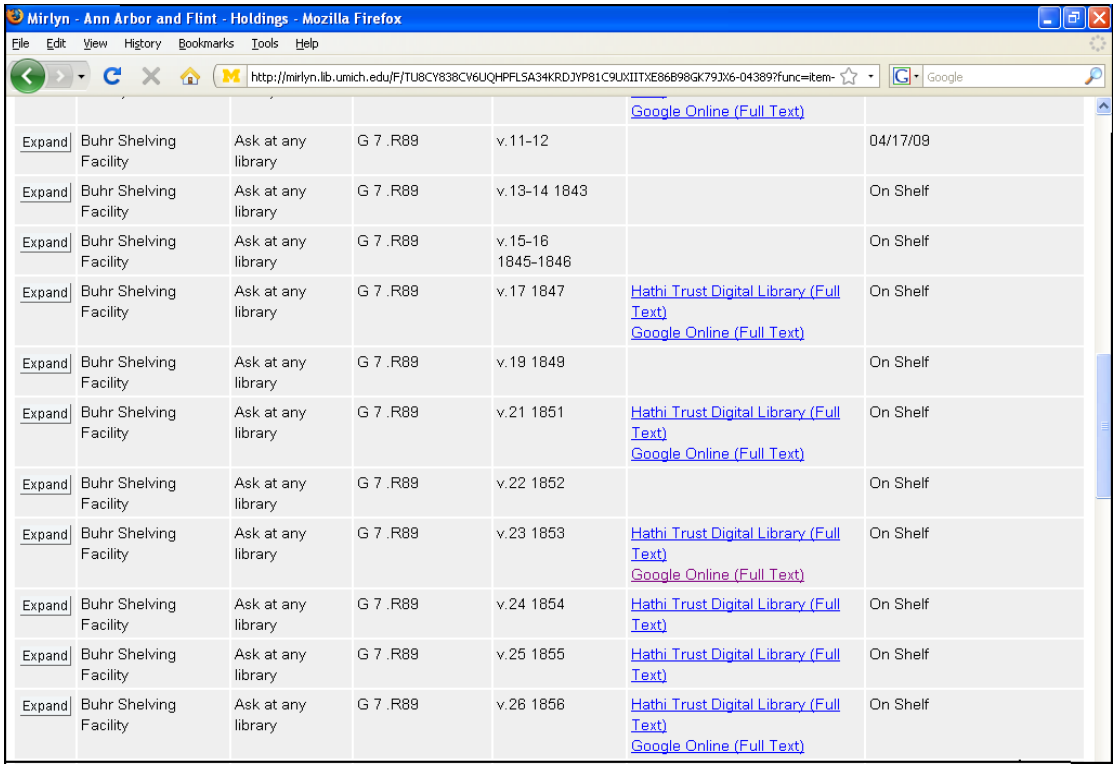

Figure 3

In addition, Google has not fully exploited Google Scholar in this regard. For instance, the following article is available through GBS:

Blumberg, H. (1919), "On Certain Saltus Equations," *American Journal of Mathematics.* Vol. 41 no. 3, pp.183-190. <http://books.google.com/books?id=UNtKfhY5wKwC>.

Google Scholar contains a reference to this article, but the citation links to copies in JSTOR and OCA, but not to GBS. One assumes that they plan to improve the integration of Google Scholar and GBS. The possibility of linking from footnotes in one book or journal to other volumes available in GBS, while technically complex, would make GBS a more conducive environment for research, and overcome at least partially the deficiencies of known-item search that currently exist.

5. Impact on ILDS

Some recent research has indicated that interlibrary borrowing has dropped as a result of increased electronic availability (Echeverria and Barredo, 2005; Goodier and Dean, 2004; Egan, 2005). However, a study of mean lending and borrowing statistics of the approximately 112 members of the Association of Research Libraries show a continued increase over the past 20 years through 2006 (ARL Statistics Interactive Edition). An ARL white paper states that the increase is largely in "returnables" (e.g. books) rather than "non-returnables" (e.g. photocopies of articles) (Beaubien, 2007). The availability of public domain items through these three mass digitization projects and a percentage of copyrighted materials through Google Book Search is unlikely to change this trend. First, the vast majority of requests to borrow materials is for items published after 1974, with a majority for items published in 2000 and later (Benn, 2009). These items, if they have been digitized, will not be in OCA at all and will not be viewable in the HathiTrust. Before the legal settlement, GBS will display only three-line snippets of text from copyrighted works. Should the legal settlement become enacted, GBS will display up to 20% of any copyrighted item. Thus, while researchers may get lucky and find particular quotations or information they need, they are just as likely to have their appetites whetted to see the entire volume, which they would need to borrow if it is not in their local library. GBS allows end users to search in ways via keyword otherwise impossible and discover books that traditional research methods would not have uncovered. While some end users will be satisfied with the limited access and find what they need, just as many others are likely to want to see the entire volume. Second, neither GBS nor OCA provides very good systems for known-item searching, particularly for journals and multi-volume sets. There is no Open URL linking to materials in these services. Items in these sites are not well connected to indexes and footnotes, and generally are not cited in references themselves. End users will need to know about and remember these projects when conducting research, because otherwise they are likely to be unaware of the availability of materials of interest through these projects.

6. Conclusion

Until the projects described above are much better integrated into the research infrastructure, it is likely that interlibrary borrowing will continue to rise, because people

will discover more sources of interest without having full access to many of them. The landscape is changing rapidly, mostly due to the Google's unprecedented efforts in digitizing millions of books and forging partnerships with publishers and rightsholders. This turbulence will stabilize, and one can foresee a day when researchers, students, and common readers will have online access to a wide array of recent publications, either provided at no cost or through subscription services from their local libraries.

Endnotes

[1] Optical Character Recognition, or OCR, is the process by which an image of a page is converted into searchable text. This process is not without flaws, and the resulting text files generally include errors.

Projects and Aggregators

Google Book Search: http://books.google.com
Internet Archive: http://www.archive.net
HathiTrust: http://www.hathitrust.org
Mirlyn: http://mirlyn.lib.umich.edu
Million Book Project (Universal Digital Library): http://www.ulib.org
Europeana: http://www.europeana.eu
American Memory (Library of Congress): http://memory.loc.gov
Chronicling America (Library of Congress): http://www.loc.gov/chroniclingamerica/

References

Albanese, A. (2008), "HathiTrust is launched." *Library Journal*, November 1, http://www.libraryjournal.com/article/CA6606527.html?q=hathitrust

ARL Statistics Interactive Edition. Geostat Center. Alderman Library University of Virginia. http://fisher.lib.virginia.edu/arl/index.html

Beaubien, A. (2007), "ARL White Paper on Interlibrary Loan." Association of Research Libraries. http://www.arl.org/bm%7Edoc/ARL_white_paper_ILL_june07.pdf

Benn, J. (2009), "Who uses the interlibrary loan and document delivery service and what do they request? A case study at the University of Western Australia." *Interlending & Document Supply*, Vol. 37 No. 1, pp.41-45.

Drummond, D (2008), "New chapter for Google Book Search." http://googleblog.blogspot.com/2008/10/new-chapter-for-google-book-search.html

Echeverria, M. and Barredo, P. (2005), "Online journals: their impact on document delivery", *Interlending & Document Supply*, Vol. 33 No.3, pp.145-149.

Egan, N. (2005), "The impact of electronic full-text resources on inter-library loan: a ten year study at John Jay College of Criminal Justice", *Journal of Interlibrary Loan, Document Delivery and Electronic Reserve*, Vol. 15 No.3, pp.23-41.

Foulser, D. (2008), "Search and find magazines on Google Book Search."

http://googleblog.blogspot.com/2008/12/search-and-find-magazines-on-google.html

Goodier, R and Dean E. (2004), "Changing patterns in interlibrary loan and document supply," *Interlending & Document Supply*, Vol. 32 No. 4, pp. 206-214.

"Google Book Search Settlement Agreement," http://books.google.com/googlebooks/agreement/

Holman, T (2004), "Google brings books to net," *Bookseller*, October 8, Issue 5149, p. 6.

IMLS News and Events (2008). Press Releases http://www.imls.gov/news/2008/091008a_list.shtm#MI

"Library partners." http://books.google.com/googlebooks/partners.html

O'Sullivan, J. and Smith, A. (2004) "All booked up." http://googleblog.blogspot.com/2004/12/all-booked-up.html

Ratnakar, V., Poncin, G., Badger, B., and Haugen, F. (2009), "1.5 million books in your pocket,"http://booksearch.blogspot.com/2009/02/15-million-books-in-your-pocket.html

Wildemuth, B. and O'Neill, A. (1995) "The 'Known' in Known-Item Searches: Empirical Support for User-Centered Design." *College and Research Libraries*, Vol 56. No. 3, pp. 265-281.

Young, J. (2005), "Microsoft, Joining Growing Digital-Library Effort, Will Pay for Scanning of 150,000 Books." *Chronicle of Higher Education*. October 27. **http://chronicle.com/free/2005/10/2005102701t.htm**

Author details

Perry Willett is the Digital Preservation Project Manager at the California Digital Library since 2008. He was previously the Head of the Digital Library Production Service at the University of Michigan.

Perry Willett
California Digital Library
415 20th St., 4th Floor
Oakland CA 94612-2901
perry.willett@ucop.edu