Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Million Book Project vs Google™ Print

ST. CLAIR Gloriana
(*University Libraries, Carnegie Mellon University, Pittsburgh, PA 15213, USA*)
E-mail: gstclair@andrew.cmu.edu
Received Aug. 5, 2005; revision accepted Sept. 10, 2005

Abstract: Google's announcement that it intended to digitize all the books in several major research libraries was met with mixed reactions. John Wilkin at the University of Michigan declared "This is the day the world changes," while Rory Litwin said in *Library Juice* that the move would "commercialize the great research libraries with a handshake, suddenly and epochally." The four directors of the Universal Library and Million Book Project have received many questions about the comparative aspects of our work and Google Print. My purpose is to compare the two, talking about their genesis, the realities of collections and logistics, and the worries that arise from these realities.

Key words: Google, Million Book Project, Universal Library
doi:10.1631/jzus.2005.A1195 **Document code:** A **CLC number:** TP391

INTRODUCTION

Google's announcement that it intended to digitize all the books in several major research libraries was met with mixed reactions. John Wilkin at the University of Michigan declared "This is the day the world changes," while Rory Litwin said in *Library Juice* that the move would "commercialize the great research libraries with a handshake, suddenly and epochally." The four directors of the Universal Library and Million Book Project have received many questions about the comparative aspects of our work and Google™ Print. The purpose of this article is to compare the two, talking about their genesis, the realities of collections and logistics, and the worries that arise from those realities.

GENESIS

The Million Book Project is a part of a larger universal library initiative directed by Dr. Raj Reddy, former head of Computer Science at Carnegie Mellon and former chair of the President's (Clinton and Bush) Information Technology Advisory Committee; Dr. Jaime Carbonell, head of Carnegie Mellon's Language Technologies Institute and expert in example based language translation; Dr. Michael Shamos, noteworthy computer scientist and intellectual property attorney; Dr. Gloriana St. Clair, Dean of University Libraries and long time editor of several library journals.

The Universal Library seeks to bring all content in a variety of forms to the Web free-to-read (http://www.ulib.org, http://www.dliernet.in/, http://www.ulib.org.cn/).

Funded by the National Science Foundation for equipment and travel, the Million Book Project plans to create a test bed for pursuing computer science research areas, such as (1) machine translation; (2) massive distributed databases; (3) storage formats; (4) use of digital libraries; (5) distribution and sustainability; (6) security; (7) search engines; (8) image processing; (9) Optical Character Recognition (OCR); (10) language processing; (11) copyright laws.

Researchers in India, China, and other countries are engaged in different aspects of these problems and look to the content as a place to try out research solutions. For instance, recently, researchers developing e-book interfaces selected this content to work with.

The Million Book Project is an international effort with research and scanning centers in several countries, especially India and China, which have about seventeen scanning centers each. Partners in China include: Chinese Academy of Sciences; Chinese Ministry of Education; Fudan University; Nanjing University; Peking University; Tsinghau University; Zhejiang University; and others.

Partners in India include: Anna University; Arulmigu Kalasalingam College of Engineering; Goa University; Indian Institute of Information Technology; Indian Institute of Science; International Institute of Information Technology; Maharashtra Industrial Development Corporation; Mysore University; Shanmugha Arts, Science, Technology & Research Academy; Tirumala Tirupati Devasthanams; and others.

The President of India, Dr. A.P.J. Abdul Kalam, supports the project and offers suggestions. In his book *Ignited Minds*, President Kalam calls knowledge the prime mover of prosperity and power. For him, knowledge is associated with education but also with skills of artists, craftsmen, philosophers, saints and housewives. In his view, academic learning coexists with the earthy wisdom of villages and their hidden knowledge of the environment. President Kalam also recounts an anecdote about an 80-year old industrialist and academician planning his next research project on Tamil scripts produced in the first Sangam, some 5 000 years ago (Kalam, 2002). These broad visions of content and audience are shared among participants in the Million Book Project.

The project has several other partnerships in the U.S. both with researchers and with institutions. Several American universities are sending content. OCLC is giving access to its databases for the creation of metadata and will be a sustainer of the content. The Internet Archive has been involved from the inception and also archives content.

To date, the project has scanned about 200 000 books in China, in modern Chinese, ancient Chinese, and English. The Ministry of Education sponsors the project and has provided funding for it in China. About the same number of books has been scanned in India, in a broader group of languages. India has eighteen official languages and has funded a broad set of language initiatives, with this project as one of them.

Google™ Print project partners include: Google, Inc.; University of Michigan; Stanford University; Harvard University; The University of Oxford, and The New York Public Library.

While Google itself began as a research project at Stanford, the Google™ Print project focuses on making Google content searchable so that the knowledge within print books can be easily accessed. Google™ Print is a commercial enterprise.

REALITIES: COLLECTIONS AND LOGISTICS

Collections for Google™ Print

The initial news stories about this project suggested that the entire collection of each of these libraries would be scanned. A closer reading of stories and Web postings tells a different story. The University of Michigan has, in fact, agreed to scan its entire seven million volume collection, with the university receiving and owning high quality digital copies of their books. Their plan was to provide access to their campus constituents for these books. Harvard and Stanford each agreed to a pilot project of about 40 000 volumes. The New York Public Library intended to scan public domain books from its collection and to make those books available on the Web to its clientele and presumably the world as well. They were planning on selecting materials that might be interesting and not too fragile.

The publishers have interrupted the work on the Google™ Print project with concerns about copyright. Google announced that it would halt work on copyrighted materials until November 1, 2005, so that publishers could decide whether they wanted to opt out of the project. Publishers feared that Google would begin to sell advertising to the results of searches of copyrighted materials without sharing revenues with publishers (Wyatt, 2005).

Collections for the Million Book Project

The original collection strategy for the Million Book Project was to scan out-of-copyright book materials, i.e. books published before 1923 and books not renewed between 1923 and 1963; government documents; and academic press books whose publishers had granted permission. Books in many foreign languages, especially Chinese and Indian lan-

guages, were welcome because example-based translation works best with a parallel corpora of about 10 000 volumes.

The Universal Library directors believe that all books in libraries are worthy of scanning. Books on a typical library shelf have not only been selected for the collection but have also often been selected for continued inclusion. Any process for choosing among eligible books is considered to be biased on the interests of the selector—one person's trash is another's treasure. The philosophy is that all books will be digital eventually and that the pace of change is such that expending resources on prioritizing content for inclusion is an unnecessary expense.

The copyright permissions work derived from an interest in finding the best books for the project. ALA's authoritative Books for College Libraries was used to select academic presses and scholarly societies whose content was cited in Book for College Libraries. These publishers were then approached to give content to the project.

In 2005, the Food and Agriculture Organization of the United Nations invited directors of the Universal Library to participate in a workshop to think through the issues around creating and using a database of agricultural information to help the rural poor. The FAO then began to send some of its content to be scanned. Other agriculture libraries joined the project, including the National Agriculture Library and some of the land grant libraries.

Logistics for the Million Book Project

The logistical challenges that face the Million Book Project are substantial. Because all materials scanned must be out-of-copyright, the status of books published between 1923 and 1963 must be checked. Dr. Michael Lesk, a professor at Rutgers University, has been helpful in comparing records from a library catalog with those in the scanned version of the copyright renewal records. The list yielded by this process must then be pulled from the shelves, boxed with a packing list, loaded into containers, and shipped to scanning centers in India and China.

Scanning centers are staffed to have a capacity of one million pages per day in India and China. Those figures, however, are dependent on having a suitable supply of books to scan. One scanning center in China is in a free trade zone so that books will not have to go through Chinese customs. Dr. Anthony Ferguson at Hong Kong University is currently supplying some materials to be worked on there while other materials from the U.S. are being prepared. Because air freight is quite expensive, books are packed and sent by container ship.

Transferring data from India and China has proven difficult because of bandwidth issues and the difficulties caused by compression. However, many files have been transferred by carrying hard drives from continent to continent. Currently, Internet2 node-to-node transfer from China to the U.S. is being explored as the research on this aspect of the project continues.

The best practice developed for metadata is to use the OCLC MARC record if it is available. Chinese scanning centers are using the METS wrapper to keep the bibliographic and administrative metadata together with the TIFFs and OCRd versions of the text. Many books being scanned in Indian languages do not have MARC records. For all books without a good MARC record, a Dublin Core record is being created.

Logistics for Google™ Print

Neither the Million Book Project nor Google™ Print is harming the books scanned; both are using preservation friendly scanners. Google is removing books and journals from library shelves, taking them to the scanning center, and then returning them to the shelves. Michigan estimates that it will take six years to scan its collection. Their original intent seems to have been to do the content in call number sequence, but publishers' reactions may force a different approach. Overall in the project, the throughput is expected to be 2.25 books per minute. Sample books are available to be viewed at http://www.googleprint.com. Google will track individuals' use of their materials in order to ensure that the copyright laws are being followed.

While Google has made some disparaging remarks about metadata in various meetings on Google™ Print, they will have OCLC MARC records easily available to them and will probably begin to use them for this project. In order to satisfy publishers, Google will have to evolve a system that allows tracking of copyright status.

WORRIES

Duplicates

According to a Stanford librarian's Web posting, "De-duplication is NOT part of the [Google™ Print] process. NOTE Stanford is interested in having multiple copies of the same materials across various partners" (Misseli, 2005). Both projects face the potential both of internal duplication and of duplication between the projects. The issue of duplication of existing resources (from such projects as the Making of America and Virginia's scanning efforts) also exists. Duplication may be expensive to avoid, as it will involve checking. If checking is desirable, then OCLC's Digital Registry needs to be populated as a central source and machine routines need to be developed to accomplish the check. Human checking title by title would be prohibitively expensive. Stanford, as the originator of LOCKS (Lots of Copies Keeps Stuff Safe), seems justified in its interest in multiple copies. Many anticipated problems, such as missing pages and degraded files, would be ameliorated by having another copy available.

Printing

Both Google™ Print and the Million Book Project discourage individuals from ad hoc printing of whole documents. The central rationale for this decision involves working with publishers for permission to include copyrighted materials. Publishers want to have Buy buttons associated with copyrighted full-text, so that individuals who have discovered books by searching online will be able to purchase those books from the publishers, as copyright holders. Print-on-demand at local facilities also offers publishers a method for continuing to gain some revenue from their out-of-print materials. Finally, print-on-demand is a technique that publishers might use to print specialized books that have only a small market. A scholar of medieval monastic history once remarked that the audience for his highly regarded monographs consisted of only a half a dozen colleagues.

Litwin's Litany

In "On Google's Monetization of Libraries," Rory Litwin notes four concerns, discussed below:

(1) Privacy. Libraries have a strong and continuing commitment to the privacy of readers in their facilities. In Google™ Print, Litwin argues that readers may be treated as customers and data sources for marketing. While Litwin toys with the idea of political repression, the realistic concern is targeting for advertising based on reading preferences. The good of having materials available in a convenient online manner must be weighed against this encroachment into the realm of privacy.

(2) Introduction of commercial bias. Litwin argues that "The aim of research, scholarship and education is truth, and people sense correctly that commercial interests have the potential to distort the discovery and spread of truth" (Litwin, 2004). He sees the academy as being largely protected from the compromises of advertising. Numerous reports suggest that commercial funding does impact results of research projects. At the same time, U.S. audiences consume a great deal of commercial television, spending about 30% of that time watching commercials. The nature of the commercial bias, as Google™ Print develops, will determine its tolerability.

(3) Democratization and equity of access. The vision of the Million Book Project is to create a free-to-read resource so that individuals worldwide can have access to information. When stored in physical libraries only, this information is not available to citizens worldwide.

Nevertheless, the creation, preservation, and dissemination of knowledge are not free. The Million Book Project relies primarily on government funding from the government of India, the Ministry of Education in the government of China, and the National Science Foundation in the U.S. Universities contribute the talents of project workers because of the importance of the research projects being undertaken. These funding sources introduce their own biases, but these biases are familiar.

Litwin argues that Google™ Print will not democratize knowledge because, very quickly, individuals will be asked to pay for in-copyright information. He sees transferring knowledge from research libraries to commercial enterprises being superficially democratizing "but deeply contrary to democracy's need for information in the public sphere" (Litwin, 2004). However, the barrier of having to make some small payment to make information quite accessible seems lower than the cultural, geogr-

aphical, and class barriers of trying to obtain that same information from a premier university library.

(4) Disintermediation and decontextualization of knowledge. Litwin describes disintermediation as the substitution of software solutions for human librarian services. Human librarian services can be excellent, but are not available, worldwide, in multiple languages, in rural locations. The Million Book Project's vision specifically seeks to create databases and texts for machine searching. Similarly, Google™ Print is about bringing the power of computer search engines inside the covers of monographs. Metadata of various kinds has been the strength of human librarian excellence with monograph full texts being known only through reviews and the relatively few volumes each librarian has read. What machine searching may lose in its lack of spark and inspiration, it well makes up for by its unceasing effort.

Decontextualization is an enormous challenge in the online environment of both projects. Coming to a paragraph or a three-line snippet through the task of identifying a book through an index, locating it on the shelf with others of its kind, testing its validity through an inspection of its physical appearance, and seeing the surrounding text supplies the viewer with an understanding that can be replicated only with a great deal of programming. Litwin calls this "a major loss of value" (Litwin, 2004). The search box does reduce results to a common, anonymous format.

How this fundamental change in the habits of students and researchers will shape their processes of ingesting information, analyzing it, and reaching conclusions will be another research agenda for scholars. The speed and precision of the new information technologies offer advantages but require the sacrifice of rich contextual information.

Sustainability

A frequently-asked question about the Million Book Project is about the plan for sustainability—after the digitization is done, who will maintain the collection and migrate it from platform to platform as needed? Carnegie Mellon University's School of Computer Science and University Libraries have made a commitment to maintain a free-to-read version of the project. Government sponsors in India and China have made similar commitment. The result will be a set of mirrored sites serving the material from different locations. The operation of that mirroring system is another of the research agendas of the project. In addition, OCLC, a partner in the project, will host a copy of the contents to be served through the WorldCat database. If Google™ Print is commercially successful, that will ensure its sustainability.

Copyright, copyright, copyright

Randall Stross's New York Times article "Google Anything, So Long as it's not Google," mentions that Eric Schmidt, the CEO of Google, describes himself as "a political junkie who never tires of debating the great issues of our day" (Stross, 2005). The ongoing debate around copyright is one of those issues and has been brought to the public's attention by the existence of digitization projects. The breadth of the Google™ Print project has made for much newspaper coverage.

Google's original plan was to digitize both in-copyright and out-of-copyright books but to display only "a snippet of text" for copyrighted materials. 'Snippet' was defined as three lines, with a list of the number of times the search terms appear in the book, and a limit of three snippets per book. A Buy button would be available so that searchers could purchase the full text from the publisher. Google's attorneys thought that this approach was within the scope of Fair Use. In August 2005, Google announced that the project was delayed while publishers reviewed the idea (Wyatt, 2005).

The Million Book Project's collection focus is on out-of-print, public domain, and publisher-permitted material. A project to gain permission to digitize books focused on the academic and scholarly society presses in *Books for College Libraries*. Having discovered that seeking copyright permission to digitize and provide open access to books using a per-title approach was too expensive to pursue on a large scale, Carnegie Mellon University Libraries changed to a per-publisher approach for the Million Book Project. The new strategy of asking permission for all of a publisher's out-of-print titles or lists of titles they designate yielded roughly 53 000 copyrighted titles for the Million Book collection and reduced the cost of acquiring permission from $78.00 to $0.69 per title.

University presses were the most likely to re-

spond to digitization requests, but the least likely to grant permission because copyright for their out-of-print books had often reverted to the authors. Most participating publishers granted permission for designated titles. Limited printing and the potential for a Buy button were important issues to these publishers (Troll Covey, 2005a; 2005b).

In the U.S., pressures around copyright issues continue to mount. However, the agendas are being driven not so much by the book publishing industry as by the movie and music industries. In the movie industry, the loss to piracy is estimated at three billion dollars annually, with DVDs accounting for 55.6 billion in revenues—about two thirds of total industry revenue (O'Brien, 2005). While part of the difficulty around movies is the cost, part is the lag time between release to theatres and appearance of the DVDs.

In colleges and universities, the reward for producing new knowledge comes through increased reputation and increments in salary. Scholarly monographs rarely produce revenue for their authors. Thus, those authors are better served by having their works available on the Web so that their ideas can receive the greatest amount of attention possible. Treating intellectual property for scholarship differently from intellectual property for entertainment would seem desirable, if somewhat complex.

CONCLUSION

The Million Book Project and Google™ Print compliment each other. Together they focus on the core issue of bringing content that is currently held inside of books to the Web where it is available for human and machine searching. The two share a vision of access to materials currently stored in physical libraries. As this vision is realized, the information seeking habits of researchers will change. On balance, knowledge will become more accessible than it is now—and that should be good for humankind and their machine assistants.

References

Kalam, A.P.J.A., 2002. *Ignited Minds: Unleashing the Power within India*. Penguin, New Delhi, p.121-127.

Litwin, R., 2004. On Google's Monetization of Libraries. *Library Juice*, **7**(26) (December 17, 2004): 7. http://www.libr.org/Juice/issues/vol7/LJ_7.26.html#3.

Misseli, 2005. The Google Deal (Down on the Farm). Message Posted by a Stanford University Staff Member to Confessions of a Mad Librarian (January 7, 2005). http://edwards.orcas.net/~misseli/blog/archives/000220.html.

O'Brien, T.L., 2005. King Kong vs. the Pirates of the Multiplex. *New York Times* (August 28, 2005): Business Section 1, 7. http://www.nytimes.com/2005/08/28/business/media/28movie.html.

Stross, R., 2005. Quoting Google CEO Eric Schmidt in "Google Anything, So Long as It's Not Google." *New York Times* (August 28, 2005), A3. http://www.nytimes.com/2005/08/28/technology/28digi.html.

Troll Covey, D., 2005a. Acquiring Copyright Permission to Digitize and Provide Open Access to Books. Council on Library and Information Resources and Digital Library Federation, Washington, DC.

Troll Covey, D., 2005b. Copyright and the Universal Digital Library. Proceedings of the International Conference on the Universal Digital Library (ICUDL), Hangzhou, China.

Wyatt, E., 2005. Google Library Database is Delayed. *New York Times* (August 13, 2005), A15, 22. http://query.nytimes.com/gst/abstract.html?res=F50614FD3E5A0C708DDDA10894DD404482.