How Will Google Print Library Project Affect the Use of Books?

Lisa Zhao
Daley Library, University of Illinois at Chicago
Email: chzh@uic.edu

Keywords(關鍵詞): Google Print Library Project; Book Search, Web Site Search; Online Books; Electronic Books; Library Users

【摘要】

GOOGLE 作為一個商業巨人進入了圖書館領域，於 2004 年 12 月宣佈其圖書館計劃，將數百萬冊圖書館的藏書數位化後存入其數據庫，供公眾網上查詢。這一計劃引起了各界的廣泛討論。本文從三方面討論：(1)從對傳統印刷書籍應用的分析，探討 GOOGLE 圖書館計劃是否可能明顯改變對這些書籍的應用；(2)從分析存在於網上搜索中的問題和使用者的搜索行為，探討 GOOGLE 圖書館計劃是否將為用戶提供更好的搜索方法和準確的查詢結果；(3)最後，本文探討商業公司提供免費信息和金錢效益發生衝突時的局限性。

【Abstract】

Since its announcement in December 2004, Google Print Library Project has attracted many reactions, optimistic and pessimistic. Google as a corporation enters the library domain by digitizing millions of library books and providing them online. Many questions have been raised from various aspects since. This essay aims to join and enrich the discussion by tackling the follow three questions: (1) Based on the analysis of the usage of print collections in libraries, it questions how much this million-dollar project will change the use of the scanned books. (2) By looking at the problems in searching methods, ranking of retrievals and users' search behavior on the Web, it examines how the project will provide better search results to users. (3) It examines the impact and limitations on the project, when a corporation's provision of free information conflicts with profit pressures.

INTRODUCTION

Five major libraries of Harvard University Library, Oxford University Library, Stanford University Library, University of Michigan Library, and the New York Public Library join Google to digitize millions of books and make contents of the books readable and searchable online via Google's interface. The Project is called Google Print Library Project (will be referred as the Project below). The announcement of the Project has splashed into the library and information field and caused endless ripples. Since December 2004, reports and comments on this Project have continued in almost every major medium. From *Wall Street Journal* to *The Chronicle Higher Education*, from *Business Week* to *American Libraries*, from CNN to CBS news, all have reported this as an "ambitious plan," a "Herculean," "massive," and "exciting" project. Meanwhile, the Project prompts arguments about the library's future. One side contends that "it's the beginning of the end for

libraries: if all the books are on Google, why would anyone need a library?" "It's the beginning of the end for print books and traditional catalogs." (Crawford, 2005). "Do people still need physical library?"(Smith, 2005). This consequently brings up some librarians' concern "whether the five participating libraries are acting in the best interest of libraries and users in genera,l" which is the so called "less well publicized question" by Vaidhyanathan in her recent article (2005). On the other side, people see hope for libraries. The Project will be "a rising tide [that] lifts all boats, rather than the tsunami image" (Quint, 2004). "Google brings libraries into [the] cyber-age" (CNN, 2004). A year later, the Project is beginning to take shape—Google Book Search beta version, as the embryo of the Project, is in use now—while the debate on some major issues of the Project such as copyright is still in full swing.

The current essay does not intend to argue that Google is a popular tool for searching information, but aims to confer how much actually this million-dollar project will provide better retrievals to book searchers and improve the overall use of the scanned books through discussing the following three issues based on reviewing some of previous researches and experiences. It may still be too soon to see the clear answer but useful to bring worthy issues to consider in the upsurge of digitization. Through such discussion from every possible angle we will be able to approach the probable answers to our questions and find a way to use technology intelligently to make a real difference for the development of libraries in providing information and services to their users.

CHALLENGE FROM AN OLD LIBRARY ISSUE: USAGE OF COLLECTIONS

Circulation analysis is a long standing method to evaluate the usage and value of library collections. One benchmark used quite often in the library collection analysis is Italian economist Vilfredo Pareto's 80/20 rule. According to Pareto's principle, also called the "vital few and trivial many rule" (Hafner, 2001), among library collections, in general, twenty percent of items in a collection produce eighty percent of its library usage. Conversely, eighty percent of a collection generates about twenty percent of all other usage. In other words, eighty percent of collections are rarely used and in some cases perhaps never checked.

A few studies have tested collection usages before and after the electronic form become available and their results fit the 80/20 rule. Justin Littman and Lynn Silipigni Connaway (2004) compared the usage of 7,880 titles that were available in both print and electronic formats at Duke University Libraries. They found "Seventy-one percent of titles that did not circulate in print were not accessed in e-book format. This suggests that the same titles that were unpopular in print were also unpopular in e-book format." Lynn Sutton (2003) in her article "Collaborating with Our Patrons: Letting the Users Select" describes how Wayne State University (WSU) Libraries in Detroit, Michigan signed an agreement for the Patron Driven Access (PDA) model with the NetLibrary in April 2002. Under this model, over 16,000 bibliographic records for all academic titles offered by NetLibrary were loaded into the WSU online catalog. "Under the 'two click' PDA model, the second time that a WSU patron opens a NetLibrary book (via either browse or checkout) a single copy of that title is purchased and charged against a previously established deposit account." Between May 1 and December 9, 2002, electronic books that were purchased through the PDA were accessed a mean 4.12 times per title. By comparison, electronic books purchased in the traditional manner, in the same testing period received only 0.43 accesses per title. This experience illustrated the difference between users' choices and the library collection purchased through the regular way.

Books the Project will put online are from the five participating libraries and generally published before 1923. The usage pattern of

these books in print format has already been defined during the years of use in those libraries. Will Google's worldwide clientele take advantage of its extraordinary search engine and break the 80/20 rule in using these books of the Project? Based on the earlier studies, the chance is on a shaky ground.

Besides relatively low use of most books in an electronic collection reported by the previous studies, we will see another curb of using the online books by looking at the interests of web surfers. According to Jeffrey I. Cole and his colleagues' "UCLA internet report" (Cole et al., 2003), the five most popular Internet activities are: 1) E-mail and instant messaging; 2) Web surfing or browsing; 3) Reading news; 4) Accessing entertainment information; 5) Shopping and buying online. Resite Information Technology (2005) has similarly reported that "bill payment, travel planning, email, and other tasks are the most common activities." Scott Kessler (2004) has also found that "approximately two-thirds use a search engine to find information about products and services, to locate Web pages." According to a recent OCLC (Online Computer Library Center) report "Perceptions of Libraries and Information Resources" (OCLC, 2005), its Market Research Team surveyed over 3,300 information consumers in Australia, Canada, India, Singapore, the United Kingdom and the United States. One question asked is about the awareness and usage of sixteen electronic resources, which include the use of electronic books (digital) even if they only were used once. The result shows that fifteen percent of respondents indicated they used such books, which ranked at the 12th of the 16 electronic resources surveyed. From these researches, reading books is obviously not listed at the top of either popular Internet activities or the most familiar usage among existing electronic resources.

Going over the previous studies about traditional patterns of using print collections and experience on the use of the collection available in both print and electronic form, the current essay suggests that there will be little improvement on the use of the vast majority of books the Project has digitized and planned to.

CHALLENGE FROM EXISTING ISSUES IN WEB SEARCH: SEARCHING AND RANKING

The Project will eventually make the text of more than ten million books searchable on Google. In its Beta version, the Project is currently so called Goolge Book Search (http://books.google.com/). One presumes in the Project each word scanned will be retrieved by keyword search. According to Hiawatha (2004) that billions of pages of the ten million books in the Project will be indexed, word by word, and made available for searching on the Internet. Increasing the amount of information by adding billions of pages of books without coordinated searching methodology can only cause more frustration on web searching. By looking through the amount of online information, online user search behavior, and two sample tests, this section examines how the Project will provide better search results for book hunters.

How much information is on the web? As of February 2005, Google has already indexed 8.05 billion web pages, more than one billion images, and 845 million Usenet messages—in total, over 9.5 billion items (Wikipedia, 2005). According to Peter Lyman and Hal R. Varian's research (2003), "the World Wide Web contains about 170 terabytes [Terabyte=10^{12} bytes] of information on its surface; in volume this is seventeen times the size of the Library of Congress print collections." The size of collection on the web is still growing. Lyman and Varian estimated that newly stored information in total grew about thirty percent a year between 1999 and 2002, and it has been growing every day since.

Facing the explosion of information, online information seekers are not satisfied with the responding speed of computers and the Internet and become less patient. They are not patient enough to look through even a fraction of their search results. Bernardo A. Huberman observed that "the average number of pages surfed at a site was almost three, users typically requested only

one page" (Huberman, et al., 1998). In its top-10 tips for building web site, MarketScapes told its audience that "the average users' patience lasts between four to ten seconds." Similarly, Grokdotcom (Future Now, 2000) advised its readers if you want your web site visitors come back, you have to know that "your average prospect will view two to three pages before moving on." Another characteristic of information searchers' behavior is the randomness of their way of search. They may retrieve in every way they think the most appropriate to their needs which means one size does not fit all.

To see how the Google Project brings books up to searchers and fits users' searching behaviors, the author did two tests in Google Book Search both in December, 2005. The test results show that the Project rarely brings up the target book from the collection to the higher rank of the search results unless the exact title is keyed in.

The author's first test is to search the book "The Changing Chinese" by Edward A. Ross published in 1911. The book came to the top of the search results after inputting the exact title into the Google Book Search. However, it disappeared among the retrievals when searched by the subject. The subject headings assigned to this book by the Library of Congress (LC) are "China, social conditions" and "China, social life and customs." Using the "Advanced Book Search" function of the Google Book Search, the author received zero match after inputting "China, Social conditions" into the window of "with the exact phrase" and set publication date between 1900 and 1922. Then the author separated the subject phrase to "China" and "social conditions", put them into the window "with all of the words" and "with the exact phrase" respectively, and set "Publication date" between 1900 and 1922. Twenty one results were returned but not the title "The Changing Chinese" by Edward A. Ross. All results have either the word "China" or the phrase "social conditions" in their title or content. There was no match when inputting "China, social life and customs" as search words no matter in what way. Regardless of general or advanced search in the Google Book Search, the word(s) used for searching has to be somewhere in the retrievals either title or content. Otherwise, it will not come into the result. Many results may contain the keywords but nothing to do with the content about "China, social conditions" and "China, social life and customs." Searchers have to figure out from the retrievals what is really useful for them.

Free online books are not new. Among many existing free online book databases, Project Gutenberg, which is in the .org domain, is the first running such services founded by Michael Hart in 1971 (Price, 2004), with over 17,000 books available in its online book catalog as this article is being written. The second test the author did was to search the book "Indian Frontier Policy" by Sir John Adye published in 1897, in both Gutenberg Project and the Google Book Search. On the Gutenberg site, searching either by the title from its Online Catalog's alphabetic title list or by the subject heading "Afghan wars", the title came out easily. In Google Book Search, the title came to the top of the results when the exact title was input. Surprisingly, there was no 1897 edition available on Google when this test was done. It only had a 2004 edition provided by Kessinger Publishing and was a "copyrighted material." The searcher was allowed to log in to read the first three pages of the book. However, even in the advanced search function, the title disappeared from the top forty results returned by searching the subject "Afghan wars," which actually was a key phrase in the limited content shown of the book.

The above two sample tests betoken the unpredictability of searching books in the Project for future users. Retrievals from the Google Project do not return the books that are about what users are searching for but only those containing users' search keyword(s) in either the book's title or content that may not relate to the subject in which searchers are really interested. This exposes a weakness of the Project that it

only supports keyword search. Thereby, even the wanted book is in the result, it may or may not show up on the top list of retrievals when searching by keywords, since millions of books are indexed word by word and there are possibly hundreds of books containing the same word. Considering the rapid growth of information on the web and users search behavior, how can Google's method of searching and ranking ensure that the books relevant to user's interests will not be buried in the fourth or fifth screen retrieved? Users, as aforementioned, seldom dig far beyond the second or third screen. The chance for a searcher to read a book listed on the fourth screen or beyond is almost zero. Printed books untouched on library shelves for years become dusty. They may also get "dusty" online in the Project.

LIMITATION OF .COMS BEING A FREE INFORMATION PROVIDER

Although placing the full text of books on the web is not new, the Project has attracted the greatest attention because of Google, a giant of commerce with the most used search engine. Google's mission, as it stated, is to "organize the world's information and make it universally accessible and useful" (Google, 2005). Similarly, the mission of libraries, in general, is "to acquire, to make accessible and to preserve information which a user may need" (Friend, 1998). The two missions are seemingly very comparable, but the typical library and Google live in two different domains, which mark the primary difference.

Providing good contents/products to make more money is a general norm of .coms and nothing to be ashamed of. Yet, it is this norm that makes the .coms' statement of equal entitlement to information and communication resources remain only a normative standard. There is always determined priority for .coms when there is a conflict between making money and providing equal access universally. For example, as Knight reported (2004), Dynamic Internet Technology (DIT), a US company that provides technology for circumventing Internet restrictions in China, has discovered that "the recently-launched Chinese version of Google News omits censored news sources from its results." DIT's chief executive Bill Xia told New Scientist, Google reinforced Chinese Internet restrictions by leaving some sites off its list. "When people do a search they will get the wrong impression that the whole world is saying the same thing." Corroborating reports come from other sources. One saying is that Google, intending to move into the Chinese market, acquainted itself with China's self-censorship policies. Google sought a competitive edge by aligning itself with China's norms (ICE, 2004). Clive Thompson recently reported on the *New York Times Magazine* (2006) that to obey China's censorship laws, "Google had agreed to purge its search results of any web sites disapproved of by the Chinese government." Do the same search inside China on google.cn, and most, if not all, the disapproved links will be gone. "Google will have erased them completely." This sort of manners seems omnipresent in the .com arena but not in the library field. Hereby, to provide information services, .coms and libraries are essentially different. Google's filtering is fundamentally unlike the library's policy of acquiring and weeding in its collection development.

In other cases, Google used filtering at the technical level on the basis of threatened or implied legal liability or responsibility, such as Google.fr (Google in France) and Google.de (Google in Germany). According to Zittrain and Edelman's research (2004), Google's counterparts, Google.fr and Google.de, intended for French and German audience, are screened search results corresponding to sites with content that might be sensitive or illegal in the respective countries. The initiative of filtering in Google.fr and Google.de is fully different from the case in China though. No matter in which situation, Google in China or in Germany and France, the filter technology can censor some words and bar contents out of the search results. Therefore, it is reasonable to suspect a part of books in the Project

will be blocked away by the Google's adaptation to certain requests or demands for exclusion in order to make profits or compete with others.

In sum, this essay concludes that the Google Print Library Project does not show the apparent change in the usage of the scanned books nor provides the improved search method to users. Then why is this million-dollar Project? As a giant in the .com world, Google will gain more publicity and be the biggest beneficiary from the Project through scanning thousands of library books into its database, more visitors to its service, and more advertising revenue. As Hansell (2005) pointed out, "now that Google is a publicly traded company, its advertising network will become more important to its business than its search engine… It is an advertising business that has nothing particularly to do with search."

Each issue discussed above can be extended and studied further. The essay presents only a trial of matters that may affect the consequence of the Project. If the challenging questions are not satisfactorily resolved, the conclusion of the Project may not be as significant as reported at its beginning and the benefit the Project will bring to libraries and users may not as much as to Google itself.

The Project, at present, may not threaten libraries nor protect them, although many say that Google and libraries are complementary in the Project. For decades, the issues of the General Agreement on Trade in Services (GATS) under the World Trade Organization (WTO) have led to extensive discussions about the future of libraries and concerns about the possibility of corporations taking over libraries. Now Google as a corporation walks into the library world. Are there the same concerns as with GATS when libraries confront the Project now?! The Project can be seen as another spur to the library world from other domains if the GATS has not been stimulating enough. Libraries need to re-examine, develop, and distinguish their role in the face of challenges from a new crowd of information providers in today's rapidly changing environment.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the great help of Dr. Stephen Wiberley. The author also thanks the journal editors and reviewers for comments.

REFERENCES

CNN report (2004). http://www.cnn.com/2004/TECH/internet/12/14/google/index.html (accessed on 4/14/05)

Cole, J. I., et al. (2003). The UCLA Internet Report Surveying the Digital Future Year Three. Feb. http://www.uiowa.edu/~commstud/resources/digitalmedia/digitalnets.html (accessed on 4/27/05)

Crawford, W. (2005). Google Print: Prototypical Reactions. Cites & Insights. http://cites.boisestate.edu/civ5i6.pdf (accessed on 4/14/05)

Fried, F.J. (1998). New mission? Or old mission with a new face? *Proceedings of the International Conference on New Missions of Academic Libraries*, Oct.25-28 1998, Beijing, 17-20. Beijing, Peking University Press, 1998. http://www.ucl.ac.uk/scholarly-communication/articles/beij.htm (accessed 4/4/05)

Future Now (2000). KISS Your Visitors if You Want Them Back. http://www.grokdotcom.com/kiss.htm (accessed on 5/5/05)

Google. Comapy overview. http://www.google.com/corporate/ (accessed on 4/16/05)

Hafner, A. W. (2001). Pareto's Principle: The 80-20 Rule. http://www.bsu.edu/libraries/ahafner/awh-th-math-pareto.htm (accessed on 4/19/05)

Hansell, S. (2005). Google to sell ads not related to searches. *New York Times.* (April 25, Section C).

Hiawatha, B. (2004). Google to index works at Harvard, other major libraries. Boston.com news, December 14. http://www.boston.com/news/education/higher/articles/2004/12/14/google_to_index_works_at_harvard_other_major_libraries/ (accessed on 4/13/05)

Huberman, B.A., et al. (1998). Strong Regularities in World Wide Web Surfing. Science 280, (Apr. 3), 96.

ICE (nternet Censorship Explore) (2004). Google and China, Jun. 16. http://ice.citizenlab.org/?p=26 (accessed on 4/2/05)

Kessler, S. (2004) Search Users Weigh In on Google. BusinessWeek online, June 11. http://businessweek.com/investor/content/jun2004/pi20040611_4085_pi076.htm (accessed on 4/12/05)

Knight, W. (2004) Google omits controversial news stories in China, September 21. NewScientist.com news service. http://www.newscientist.com/article.ns?id=dn6426 (accessed on Apr. 5, 2005)

Littman, J. & Connaway, L. S. (2004) A circulation analysis of print books and e-books in an academic research library. http://www.oclc.org/research/publications/archive/2004/littman-connaway-duke.pdf (accessed on 4/13/05)

Lyman, P. and Varian, H. R. (2003). How much information? http://www.sims.berkeley.edu/research/projects/how- much-info-2003/ (accessed on 4/14/05)

MarketScapes. Top-10 Tips for Building Your Website. http://www.marketscapes.com/news/ms_story5.htm (accessed on 5/8/05)

OCLC Report: Perceptions of Libraries and Information Resources (2005), http://www.oclc.org/reports/2005perceptions.htm (accessed on 12/11/2005)

Quint, B. (2004). Google's Library Project: Questions, Questions, Questions. http://www.infotoday.com/newsbreaks/nb041227-2.shtml (accessed on 4/11/05)

Resite Information Technology (2005). The Internet & Apartment Management. http://resiteit.com/internettoday1.php (accessed on 4/26/05)

Smith, Kathlin (2005). The Value of Library as Place. *Council on Library and Information Resources*, no.43, Jan/Feb. http://www.clir.org/pubs/issues/issues43.html (accessed on 4/15/05)

Sutton, L. (2003) Collaborating with Our Patrons: Letting the Users Select, Learning to Make a Difference. Proceedings of the Eleventh National Conference of the Association of College and Research Libraries (Chicago: Association of College and Research Libraries, 2003), 211-215. http://www.ala.org/ala/acrl/acrlevents/lsutton.PDF (4/14/05)

Thompson, Clive (2006) Google's China Problem (and China's Google Problem). *New York Times Magazine*, April 23, 2006. http://www.nytimes.com/2006/04/23/magazine/23google.html (accessed on Apr. 27, 2006)

Wikipedia (2005) Google (search engine). http://en.wikipedia.org/wiki/Google_%28search_engine% 29 (accessed 4/5/2005)

Vaidhyanathan, Siva (2005) A risky gamble with Google. The Chronicle Review, December 2, 2005, B7.

Zittrain, J. & Edelman, B. (2004). Localized Google search result exclusions. http://cyber.law.harvard.edu/filtering/google/ (accessed on 4/5/05)