Google’s Digitization Project – What Difference Will it Make?

Barbara Fister

Originally published in *Library Issues,* March 2005

Last December, for the first time in memory, research libraries made the front page of The New York Times — above the fold, yet! What could possibly make such traditional institutions newsworthy? “Google is Adding Major Libraries to its Database” read the headline. Ah, that explains it. Though the popular search engine’s project is just getting off the ground, far behind many other book digitization projects, Google’s announcement that it was partnering with major research libraries to put their collections online could be a tipping point in public awareness of virtual libraries. Whatever its ultimate outcome, it was part of a shrewd PR campaign by Google to retain the leadership position among competing search engines. First came the long-awaited IPO, accompanied by a novel prospectus. Most SEC filings don’t include multimedia files and the text of *Playboy* interviews. Then, the announcement of Google Scholar, a specialized search engine for research materials with a citation search built in. And finally, the most dramatic development: everyone, everywhere would soon have the contents of the libraries of Oxford, Stanford, Harvard, the University of Michigan, and the research collections of New York Public Library available at the click of a mouse. Students, who some worry use the Internet as their first and only research tool, would now be able to browse the world’s great research collections. Once Google included the most traditional and respected format of recorded knowledge, Google could finally live up to its grandiose mission “to organize the world’s information and make it universally accessible and useful.” Or is it just hype?

The Good News Is …

Many initial respondents were excited that so much knowledge would be available so democratically, that books “locked up” in libraries would finally be free for all to use. Librarians, whose mission is essentially identical to Google’s, were naturally dismayed that anyone would consider them corrections officers in a Big House for books, but participating libraries were pleased they could share their collections with geographically remote patrons. They will also receive a digital copy of books from their collection to use as they choose. Though many libraries are digitizing materials selectively, only a company with deep pockets could fund a project with such an ambitious scale. Scholars were excited by the dream of tracking down fugitive fragments of printed text with a few keystrokes. And many bibliophiles approved of the notion that books matching a search would appear before other Web results. Better yet, the resulting Web pages would include a link that would let searchers locate a copy in local library collections, thanks to an earlier project developed with OCLC, a vast international library cooperative. In short, the Web would be enriched by the full contents of millions of books, an enormous improvement on the hodgepodge of self-published and sometimes spurious information students so often find on the Web.

So What’s Not To Like?

Ask Michael Gorman, president-elect of the American Library Association. In an opinion piece in the *Los Angeles Times* he declared it “a solution in search of a problem.” Because books aren’t meant to be consumed in bits and bytes, but require sustained and contextual reading, he considers large digital book projects “expensive exercises in futility.” Searches that locate a snippet of text are inefficient and misleading compared to the more holistic and descriptive traditions of cataloging and classification. Or ask Rory Litwin, editor of *Library Juice,* who raises serious concerns about privacy. Google has been criticized for retaining massive amounts of information gathered through “cookies,” using it to target advertising according to search terms, the sort of over-the-shoulder monitoring that is anathema to librarians. John Wilkin, Associate University Librarian at the University of Michigan, acknowledges it is an issue, but says that Google’s privacy policy is well within the mainstream of Internet business practices. “They’ve developed a good track record of creativity and responding to market pressures.” Litwin further argues this alliance with a single for-profit company bent on dominating the search engine market threatens the democratic nature of libraries and their critical role in sustaining the information commons for the public good. He considers this partnership, kept a closely-guarded secret until the right media moment arrived, a sell-out to big business and a betrayal of core library values. Then there is the copyright problem. Three of the five libraries participating are only digitizing books clearly in the public domain. That means if students do encounter books through a Google search, chances are they will be so out of date as to be useless. Even classics will only appear in inferior editions, without the benefit of contemporary introductions and annotations, careful editing or the inclusion of recently-discovered material.

Stanford and the University of Michigan, willing to have in-copyright books scanned, are pushing the legal envelope. Google plans to show searchers only limited selections of copyright-protected books supplied by libraries. To see the rest of it, searchers will follow links to either buy the book or see if it’s in a local library. But in order to make the book searchable, Google will have to make a digital copy of the entire text, which is almost certainly a violation of the law. Obviously, there are still some kinks to work out.

“Books have to be created before they can be digitized and though online books have been available on the Web for many years, they haven’t destroyed the public’s love of books - or the book publishing industry.”

Google’s partnership with libraries was a strategic move. Not only have they “co-branded” their name with those of prestigious libraries, they have scored an end-run around reluctant publishers. Over a year ago Google rolled out a plan to engage willing publishers in a “Google Print” program, arguing better visibility of the content of books would lead to sales. But Amazon had already signed agreements with publishers to make the full text of tens of thousands of current consumer-oriented books searchable at Amazon's site, allowing customers to search full text and browse a limited number of pages without copying or printing pages. Many trade publishers who willingly participated in the Amazon project were hesitant to work with Google. After all, Amazon is a bookseller; selling books is not Google's main mission. As of this writing, the country's largest publishing group, Random House, has yet to come to an agreement with Google, though much of their list is already in full text at Amazon. (Incidentally, books that publishers willingly contribute to Google only have links to booksellers included on their pages; a Google staffer told me in an e-mail "while we currently are not featuring the library links on publisher-submitted titles, we may expand this option in the future.") When publishers baulked, Google brought aboard allies willing to put the copyright issue firmly on the table — libraries. This is not just a matter of commercially viable books. Much material under copyright is out of print and not available at any price. Tracking down the rights-holders is difficult and sometimes impossible — and those in the content industries have an interest in keeping it that way. Stanford law professor and copyright activist Laurence Lessig hopes this bold project may play a role in much-needed copyright reform.

Academic Questions

Interestingly, University presses are more likely to participate in Google's venture than Amazon's, according to Douglas Armato, Director of the University of Minnesota Press and current president of the Association of American university presses. An informal survey of the AAUP Board found the majority of presses were interested in submitting at least some of their list to Google. The University of Minnesota Press has submitted approximately 70 percent of their list, leaving out those that have significant material belonging to multiple rights holders. Armato points out that the majority of scholars buy their scholarly books online. The potential for dissemination and discovery through the Web, while not without risk, is too promising for university presses to "stand on the sidelines." Evidence suggests people will pay for the convenience of reading sustained texts offline. After all, the 9/11 Commission's report became a bestseller even though it was available for free on the Web. The National Academies Press, which makes the full text of their books available online, has conducted a study that supports the argument that free, online browsing does not hurt sales; it may even create new opportunities to "unbundle" book content and create new revenue streams.

Abundance: Dream or Nightmare?

Technological change tends to be met with utopian optimism or dystopian gloom, and Google's partnership with libraries is no exception. One common leap of logic made in responses to this announcement is that digitizing millions of books will relieve us of the need to travel to physical libraries and hunt painstakingly through miles of shelves to find what we need. Once the easily searched contents of research libraries are universally available, regardless of geographic or economic barriers, schoolchildren in poor countries will be on equal footing with Harvard students. We will enter a new era of free, global access to information, and the world will be a much better place. Either that, or it's the end of civilization as we know it. We will drown in a sea of undifferentiated information. Students will be even more convinced "it's all on the Web" and will cobble together undigested and decontextualized bits of information, mistaking information for knowledge. Their attention span will further deteriorate.

The skill of sustained, contemplative reading will be lost in the clamor of the Web, and the tactile pleasure of books will be erased forever by the flickering, transitory glow of the computer screen. Of course, neither scenario is remotely plausible. The project is at least a decade away from being completed. Books have to be created before they can be digitized and though online books have been available on the Web for many years, they haven't destroyed the public's love of books — or the book publishing industry.

A Look Back

We've had mixed emotions about having too much information since Biblical time; the Book of Ecclesiastes complains "of the making many books there is no end." With the advent of the printing press, much uneasiness was caused by the availability of so much unregulated textual production. "One of the diseases of the age is the multiplicity of books," Elizabethan-era writer Barnabe Rich grumbled. "They doth so overcharge to [the] world that it is not able to digest the abundance of idle matter that is every day hatched and brought into the world." In 1937 H.G. Wells published an essay "World Brain" in which he proposed a radical new way to build a shared and always growing archive of human knowledge, a vast and dynamic encyclopedia to which experts would constantly contribute — essentially the same concept as today's *Wikipedia.* "The whole human memory can be, and probably in a short time will be, made accessible to every individual" thanks to the miracle of microfilm. He was hopeful this sharing of knowledge would lead to the unification of humankind through sharing a singular body of common wisdom — sorely needed in those troubled times. A few years later, toward the end of the war, Vannevar Bush fretted about the growth of information in a new age of federally-funded science. With opportunity came a problem of controlling and containing knowledge as it spun off into new and ever-narrowing specialties. "There is a growing mountain of research," he wrote in 1945. "But there is increased evidence that we are being bogged down today as specialization increases. The investigator is staggered by the findings and conclusions of thousands of other workers — conclusions which he cannot find time to grasp, much less remember as they appear." His solution? The hypothetical "memex," a desk-sized storage and retrieval machine that would call up microfilmed material indexed by trails of association. This technological dream for extending human memory and improving recall of text is often considered the first expression of the hypertext concept. Perhaps Jorge Luis Borges expressed the paradox of abundance best in his short story, "The Library of Babel" (1941). He describes a library so vast it has no exact center and no circumference. It contains an infinite number of books, a scholar's dream. "When it was announced that the Library contained all books, the first reaction was unbounded joy . . . There was no personal problem, no world problem, whose eloquent solution did not exist." This hopeful state of affairs, however, was not to last. "This unbridled hopefulness was succeeded, naturally enough, by a similarly disproportionate depression. The certainty that some bookshelf in some hexagon contained precious books, yet that those precious books were forever out of reach, was almost unbearable"— an emotion any researcher can identify with.

Impact on Libraries and Learning

If within the next decade Google does manage to digitize research collections, what will be the impact on our campuses? Will our libraries become obsolete, or at least less likely to be used by the students for whom we build selective collections, design catalogs, and provide service to help them find what they need? Copyright issues aside, searching the full text of books can be helpful for the scholar in search of a very specific phrase, but it's nearly useless for novice researchers. In fact, all researchers, including the undergraduate, need at least three different approaches to finding information in books. No single approach is the ultimate solution.

— *Subject cataloging.* Catalogers describe books with a small number of headings that describe the book *as a whole*. When you find books listed by subject, you can be sure it isn't just a throwaway line on page 312, it's going to be covered in depth. Of course, it's hardly a fool-proof system. It can take several years for a new concept to be named as an official subject heading.

— *Classification*. By placing books on similar topics near each other on the shelves, browsing can enable a researcher to discover varieties of approaches to the same idea. Because the system is designed for expansion, new concepts can find their place on the shelves before they are named in subject headings. One drawback is that every classification system betrays the assumptions of its creators. The Library of Congress system used by most libraries implies the study of women is subordinate to the study of families. And interdisciplinary fields such as Classics or Environmental Studies are scattered through several sections.

— *The citation network.* One of the best ways to locate selected, related materials is to follow the leads provided by other scholars in the form of footnotes. Of course, in books, these links are all to earlier publications, but databases with citation indexing can connect researchers to more recent work. Unfortunately to many students, footnotes are only so much fine print; they often fail to appreciate how useful they can be.

Even if Google were able to build such functionality into their search, students would still find it no substitute for traditionally printed books. Students dislike reading sustained texts on the computer and are reluctant to rely on sources that they cannot hold in their hands. Though full-text articles are popular, students almost always print them out so they can sort, compare, and annotate them, consulting them multiple times as they compose their own texts. Finally, cyberspace is no substitute for the physical libraries. Their uniquely human social space is important to students for study, group work, browsing and selecting resources, all while checking e-mail and using Instant Messenger or chance encounters to chat with friends.