

<http://chronicle.com/weekly/v51/i18/18a03701.htm>

From the issue dated January 7, 2005

Google Will Digitize and Search Millions of Books From 5 Top Research Libraries

By SCOTT CARLSON and JEFFREY R. YOUNG

Five of the world's largest libraries have joined Google in a herculean effort to digitize millions of books and make every sentence searchable.

The project involves libraries at Harvard and Stanford Universities, the University of Michigan at Ann Arbor, and the University of Oxford, in England, as well as the New York Public Library.

It could turn Google into the single largest holder of digitized published material, while providing researchers and students with an unprecedented tool for finding information.

Once finished, the digital repository could help researchers identify links among materials or discover books they would never have found by traditional methods. And the project could greatly increase access to books, since people will be able to call up many books online rather than having to make a trip to a library to read them.

The trickiest issue is copyright. The company will begin by scanning works that are in the public domain, and the full texts of those books will be accessible online through the popular Google search engine.

But the company also plans to scan copyrighted books in some of the libraries. The search engine will not give users the full texts of those volumes, but will provide up to three short excerpts, each consisting of only a few lines of text in which a search term appears.

Google officials and librarians hope that the excerpts will be sufficient to let researchers determine whether they want to check out or purchase the book. Google will include links to online booksellers and local library catalogs with the search results.

The number of volumes that could be scanned is astounding: The New York Public Library holds 20 million books, Harvard has some 15 million, Stanford more than 7.6 million, and Michigan 7.8 million. Oxford's main library alone has more than 6.5 million books.

Harvard, Stanford, and the New York Public Library have agreed only to pilot projects with the company. Google will initially scan subsets of their collections, and decisions about whether to proceed with the rest will come later. Oxford has signed an agreement with the company to let Google scan all of its 19th-century books, some 1.5 million volumes. Michigan officials, however, have agreed to allow all of their books to be scanned, and the effort has been quietly under way for months. The scanning is expected to take years.

Google will pay for the scanning and will dispatch a small group of employees to each library to do the job. The company would not reveal the cost of the project, although some experts estimate that the scanning will cost an average of \$10 per book. Library officials will decide in what order to scan the books, and which ones are too fragile to be handled.

Huge Benefits Foreseen

Some librarians see the deals as a major boon for libraries and patrons -- and a way to raise the public's awareness of the materials that can be found in the stacks.

"At a fundamental level, this is a very important move forward for the public's ability to access scholarly information," said Duane E. Webster, executive director of the Association of Research Libraries. "This enrichment of resources will entice even more users to those libraries that see themselves as learning commons."

Google officials said the effort is an expansion of the company's [Google Print](#) project, which searches the texts of books. Google Print, which started in October, initially invited only publishers, rather than libraries, to join.

Susan Wojcicki, director of product management for Google, said the digitization would lead to an increase in book sales because it would show readers what the works contain. "For publishers, we believe that this will be beneficial," she said.

Many publishers support Google's ambitious new project.

Allan Adler, vice president for legal and government affairs at the Association of American Publishers, said company officials "have made the attempt to recognize that with respect to copyrighted works, they can't simply offer the libraries or the public the ability to have full-text access to these works without the permission of the copyright holders."

Just how much access to copyrighted works will be allowed is still open to question, he said. "It's going to begin a dialogue between Google and the publishers, and the libraries and the publishers. We don't know the outcome of those discussions yet."

So far Google Print is separate from the company's recently introduced [Google Scholar](#) search engine, which lets users search academic materials. "But the products may be potentially integrated in a variety of interesting ways," said Ms. Wojcicki.

Some librarians, however, are ambivalent about Google's ambitious new effort.

"In some ways, it's a good thing," said Steven J. Bell, library director at Philadelphia University. Because Google is such a popular search tool -- among the first employed by almost anyone doing research -- "it's going to help people find high-quality sources of information," he said.

But he worries about what effect Google Print will have on library patrons' perceptions of electronic searching. Most library databases allow users to refine their searches more than they can by using Google's search engine, he noted. "This will add pressure to make things more like Google, and it will only serve to weaken the ability to get good information," he said. "It's going to be that much harder to convince people to use a more complex search tool."

He added that librarians and others should have a dose of "healthy skepticism" about the project. "Google is probably not going to do anything that doesn't have a profit return on it," Mr. Bell said. "What does that mean? Are people going to be getting a book out of Stanford's collection, and will they be prompted to buy something?"

Advocates for disabled people, meanwhile, are worried that Google's archive will not be accessible to the screen-reading software that blind people and others use to surf the Web. Google Print makes book pages available as image files rather than as the standard text that such software can process. Google officials did not reply to questions about the accessibility of their online books.

Varying Arrangements

The University of Michigan's library was the first to strike a deal with Google.

"We have been working on this for a couple of years, and it's amazing that we've been able to keep it under wraps," said John P. Wilkin, an associate university librarian. After more than a year of negotiations, he said, the company and the university finally agreed to start digitizing books this past spring.

Since then thousands of books have been digitized at the university with a machine owned and operated by Google. Mr. Wilkin would not describe the device other than to say that it works very quickly: "I've seen them whip through the book as fast as turning the pages." Digitizing all of the volumes in Michigan's collection will take about six years, he estimated.

Michigan will also store a copy of the digitized collection -- which takes up "hundreds of terabytes," Mr. Wilkin said -- for its own use.

Paul N. Courant, provost and executive vice president for academic affairs, said the digital collection at Michigan would be used "to the maximum extent permitted by law." He envisions students' and researchers' getting access to works in the public domain from their home computers. He also sees the university library setting up a catalog in which the entire collection is searchable down to the level of words and phrases.

A project like this is worth "hundreds of millions" of dollars to the university, he said. "This is an important moment in the history of libraries, and an important moment in the history of scholarship."

Other participants are proceeding cautiously. Harvard has agreed to let Google scan only 40,000 books during the pilot phase of the project. The books will be selected randomly from the five million volumes in the Harvard Depository, where seldom-requested books are stored, said Peter Kosewski, director of publications and communications at the university's library.

Sidney Verba, the library's director, said that librarians at Harvard were "very optimistic" that the pilot phase would succeed, and that they would then go forward with scanning the entire collection, which could take many years.

"It is so big that we just wanted to be sure that our hopes and expectations really pay off," he said. During the test period, library officials will be watching the process closely. "We want to make sure the books don't get damaged," said Mr. Verba, "and we want to make sure that we have a work flow such that the books don't get lost or are unavailable to our users."

Researchers would benefit enormously if whole libraries could be searched by Google's software, he said. "Everybody that's got a teenage kid knows that that's how people find information. By making the existence of the world's books available online through Google, in a way we're trying to take advantage of the fact that people go there for information."

No Longer in the Dark

In February *The New York Times* made a fleeting mention of an ambitious digitization effort by Stanford and Google that was code-named Project Ocean. For months, officials at both the university and the company refused to elaborate, and librarians across the country wondered what was up.

But Stanford did not sign its agreement with Google until December. Andrew C. Herkovic, director of foundation relations and strategic projects at the Stanford University Libraries, said the university had taken time to clarify copyright concerns and ensure its rights to the digital files.

Stanford will at first offer Google "hundreds of thousands" of items that are in the public domain, he said, and only later might make the university's entire collection available for scanning.

Officials of the New York Public Library said the project fits well into the library's mission to make information available free to the public. "This is the first time that the public is able to search the full content of any of our holdings electronically," said Nancy Donner, vice president for communications and marketing. "Frankly, without Google's assistance the cost of digitizing our books, in both time and dollars, would really be prohibitive."

During the pilot phase of the project, the library has agreed to let Google scan "more than 10,000 and less than 100,000 books," said Ms. Donner, declining to reveal the exact number. Only public-domain books will be scanned; librarians will choose those that they believe will be of the widest interest.

Paul LeClerc, president of the library, said the project would be a "huge" benefit to researchers because it would make the process of finding materials more efficient. "The search engine, in effect, is reading all the books for you" and helping decide which are the most promising, he said. "People don't have to spend an extraordinary amount of time looking for things that are contained in the volumes physically."

Google's deal with Oxford allows the company to scan nearly all of the library's 19th-century books. The only ones to be excluded are those that Oxford librarians deem too fragile to scan, said Ronald R. Milne, acting director of library services and librarian of the university's main library.

The initial agreement is "for two or three years," he said, after which officials would consider letting Google scan the rest of the university's holdings that are not protected by copyright.

Google's venture is not the first to try to assemble a digital library of books.

Last month the Internet Archive, a nonprofit group dedicated to building online archives, announced that major libraries in six countries -- Canada, China, Egypt, India, the Netherlands, and the United States -- have agreed to put their digitized books into the group's collection and in other free open-access libraries.

The group already has some 27,000 books online, with 50,000 more expected by early this year.

Brewster Kahle, librarian for the Internet Archive, appreciates Google's commitment to putting books online. "I think Google will do a great job of that," he said, "as well as other search companies" that will very likely follow suit.

But Mr. Kahle said he hoped the libraries involved would also place copies of their scanned books in open-access archives. Otherwise, he said, the result will effectively "commercialize the public domain."

"We would lose tragically if we did not make sure that there is great noncommercial access as well, that unifies and provides a critical mass of materials to researchers and scholars," Mr. Kahle said. One reason such noncommercial archives are important, he said, is so researchers can test new search techniques that companies like Google may not be willing to try.

It is unclear whether there will be such open access to the books that Google scans, however.

Under the terms of the libraries' deals with Google, each university will be given a digital copy of every book scanned and will be able to use those copies in almost any way it wants. One restriction, however, is that the libraries cannot "give it all to Yahoo or the other search companies," noted Oxford's Mr. Milne.

Mr. Milne said he would be "happy to talk to anybody about any sensible idea" to add the university's digital copies to open archives, as long as doing so stayed within the bounds of the university's agreement with Google. "We as librarians are quite used to cooperating with our peers at other institutions," he added.

Daniel Greenstein, director of the California Digital Library, a project of the University of California system, said that if universities are able to cooperate in building a digital collection, they could focus less on keeping printed books in main libraries and dedicate library space to other things: special collections, places for collaborative learning, facilities that support high-tech computer research. "If concerns about preservation and archiving are all hammered out," he said, "it enables libraries to invest resources elsewhere."

<http://chronicle.com>

Section: Information Technology

Volume 51, Issue 18, Page A37

[Copyright © 2005 by The Chronicle of Higher Education](#)

[Subscribe](#) | [About The Chronicle](#) | [Contact us](#) | [Terms of use](#) | [Privacy policy](#) | [Help](#)