

Tuesday, December 14, 2004

<http://chronicle.com/free/2004/12/2004121401n.htm>

Google Will Digitize and Search Millions of Books From 5 Leading 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, which Google officials plan to announce today, involves libraries at Harvard and Stanford Universities, the University of Michigan at Ann Arbor, and the University of Oxford, as well as the New York Public Library. It could soon turn Google into the single largest holder of digitized published material, while also providing researchers and students with an unprecedented tool for finding information.

The trickiest issue is copyright. The company will begin by scanning works that are in the public domain, and the full contents 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 return the full texts of those volumes, but will instead 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 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 catalogues along with search results.

The number of volumes that could be scanned is astounding: Harvard holds some 15 million volumes; the New York Public Library has 20 million; Stanford has more than 7.6 million; and the University of Michigan has 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 will allow Google to scan only books published before 1900, according to Nathan Tyler, a spokesman for Google.

Officials at the University of Michigan, however, have agreed to allow all of their books to be scanned, and the effort has been quietly under way for months. All of the projects are expected to take years to complete.

Google will pay for the scanning, and it will dispatch a small group of employees to each library to do the job. Library officials will decide in what order to scan the books, and also which volumes are too fragile to be handled.

Huge Benefits Foreseen

Some librarians see Google's deal with the five institutions 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, the 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 that the Google Print project would lead to an increase in book sales because it would show readers what the volumes contain. "For publishers, we believe that this will be beneficial," she said.

So far, Google Print is separate from the company's [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 project.

"In some ways, it's a good thing," said Steven J. Bell, the 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 make more refined searches than they can using Google's search engine, he said. "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," he 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?"

Varied Deals

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, in an interview Monday. He said that after more than a year of negotiations, Google and the university had finally agreed to start digitizing books last spring.

Since then, thousands of books have been digitized at the university through 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," he said. He expects that it will take about six years to digitize all seven million volumes in Michigan's collection.

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

Paul N. Courant, the university's provost and executive vice president for academic affairs, said the digital collection 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 individual words and phrases.

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

Other participants are proceeding more cautiously. Harvard University, for example, 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, an off-site storage facility for seldom-requested books, said Peter Kosewski, director of publications and communications for the university's library.

Sidney Verba, the library's director, said librarians at Harvard were "very optimistic" that the project would succeed and that they would 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, officials will be watching the process closely. "We want to make sure the books don't get damaged, 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."

Mr. Verba said researchers would benefit enormously if whole libraries could be searched by Google's software. "Everybody that's got a teenage kid knows that that's how people find information," he said. "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 project by Stanford and Google that was code-named "Project Ocean." For months, officials at the university and the company refused to elaborate, and librarians across the country wondered what the project might entail.

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

Mr. Herkovic said that the university would first offer Google "hundreds of thousands" of items that are in the public domain, but that Google might eventually scan the university's entire collection.

Officials at 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 for the library. "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, though she would not reveal the exact number. Only public-domain books will be scanned, and librarians will select volumes 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."

[Front page](#) | [Career Network](#) | [Search](#) | [Site map](#) | [Help](#)

Copyright © 2004 by The Chronicle of Higher Education