

December 14, 2004

## Google Is Adding Major Libraries to Its Database

By JOHN MARKOFF and EDWARD WYATT

**G**oogle, the operator of the world's most popular Internet search service, plans to announce an agreement today with some of the nation's leading research libraries and Oxford University to begin converting their holdings into digital files that would be freely searchable over the Web.

It may be only a step on a long road toward the long-predicted global virtual library. But the collaboration of Google and research institutions that also include Harvard, the University of Michigan, Stanford and the New York Public Library is a major stride in an ambitious Internet effort by various parties. The goal is to expand the Web beyond its current valuable, if eclectic, body of material and create a digital card catalog and searchable library for the world's books, scholarly papers and special collections.

Google - newly wealthy from its stock offering last summer - has agreed to underwrite the projects being announced today while also adding its own technical abilities to the task of scanning and digitizing tens of thousands of pages a day at each library.

Although Google executives declined to comment on its technology or the cost of the undertaking, others involved estimate the figure at \$10 for each of the more than 15 million books and other documents covered in the agreements. Librarians involved predict the project could take at least a decade.

Because the Google agreements are not exclusive, the pacts are almost certain to touch off a race with other major Internet search providers like [Amazon](#), [Microsoft](#) and [Yahoo](#). Like Google, they might seek the right to offer online access to library materials in return for selling advertising, while libraries would receive corporate help in digitizing their collections for their own institutional uses.

"Within two decades, most of the world's knowledge will be digitized and available, one hopes for free reading on the Internet, just as there is free reading in libraries today," said Michael A. Keller, Stanford University's head librarian.

The Google effort and others like it that are already under way, including projects by the Library of Congress to put selections of its best holdings online, are part of a trend to potentially democratize access to information that has long been available to only small,

select groups of students and scholars.

Last night the Library of Congress and a group of international libraries from the United States, Canada, Egypt, China and the Netherlands announced a plan to create a publicly available digital archive of one million books on the Internet. The group said it planned to have 70,000 volumes online by next April.

"Having the great libraries at your fingertips allows us to build on and create great works based on the work of others," said Brewster Kahle, founder and president of the Internet Archive, a San Francisco-based digital library that is also trying to digitize existing print information.

The agreements to be announced today will allow Google to publish the full text of only those library books old enough to no longer be under copyright. For copyrighted works, Google would scan in the entire text, but make only short excerpts available online.

Each agreement with a library is slightly different. Google plans to digitize nearly all the eight million books in Stanford's collection and the seven million at Michigan. The Harvard project will initially be limited to only about 40,000 volumes. The scanning at Bodleian Library at Oxford will be limited to an unspecified number of books published before 1900, while the New York Public Library project will involve fragile material not under copyright that library officials said would be of interest primarily to scholars.

The trend toward online libraries and virtual card catalogs is one that already has book publishers scrambling to respond.

At least a dozen major publishing companies, including some of the country's biggest producers of nonfiction books - the primary target for the online text-search efforts - have already entered ventures with Google and Amazon that allow users to search the text of copyrighted books online and read excerpts.

Publishers including HarperCollins, the Penguin Group, Houghton Mifflin and Scholastic have signed up for both the Google and Amazon programs. The largest American trade publisher, Random House, participates in Amazon's program but is still negotiating with Google, which calls its program Google Print.

The Amazon and Google programs work by restricting the access of users to only a few pages of a copyrighted book during each search, offering enough to help them decide whether the book meets their requirements enough to justify ordering the print version. Those features restrict a user's ability to copy, cut or print the copyrighted material, while limiting on-screen reading to a few pages at a time. Books still under copyright at the libraries involved in Google's new project are likely to be protected by similar restrictions.

The challenge for publishers in coming years will be to continue to have libraries serve as major influential buyers of their books, without letting the newly vast digital public reading rooms undermine the companies' ability to make money commissioning and publishing authors' work.

From the earliest days of the printing press, book publishers were wary of the development of libraries at all. In many instances, they opposed the idea of a central facility offering free access to books that people would otherwise be compelled to buy.

But as libraries developed and publishers became aware that they could be among their best customers, that opposition faded. Now publishers aggressively court librarians with advance copies of books, seeking positive reviews of books in library journals and otherwise trying to influence the opinion of the people who influence the reading habits of millions. Some of that promotional impulse may translate to the online world, publishing executives say.

But at least initially, the search services are likely to be most useful to publishers whose nonfiction backlists, or catalogs of previously published titles, are of interest to scholars but do not sell regularly enough to be carried in large quantities in retail stores, said David Steinberger, the president and chief executive the Perseus Books Group, which publishes mostly nonfiction books under the Basic Books, PublicAffairs, Da Capo and other imprints.

Based on his experiences with Amazon's and Google's commercial search services so far, Mr. Steinberger said, "I think there is minimal risk, or virtually no risk, of copyrighted material being misused." But he said he would object to a library's providing copyrighted material online without a license. "If you're talking about the instantaneous, free distribution of books, I think that would represent a problem," Mr. Steinberger said.

For their part, libraries themselves will have to rethink their central missions as storehouses of printed, indexed material.

"Our world is about to change in a big, big way," said Daniel Greenstein, university librarian for the California Digital Library of the University of California, which is a project to organize and retain existing digital materials.

Instead of expending considerable time and money to managing their collections of printed materials, Mr. Greenstein said, libraries in the future can devote more energy to gathering information and making it accessible - and more easily manageable - online.

But Paul LeClerc, the president and chief executive of the New York Public Library, sees Web access as an expansion of libraries' reach, not a replacement for physical collections. "Librarians will add a new dimension to their work," Mr. LeClerc said. "They will not abandon their mission of collecting printed material and keeping them for

decades and even centuries."

Google's founders, Sergey Brin and Larry Page, have long vowed to make all of the world's information accessible to anyone with a Web browser. The agreements to be announced today will put them a few steps closer to that goal - at least in terms of the English-language portion of the world's information. Mr. Page said yesterday that the project traced to the roots of Google, which he and Mr. Brin founded in 1998 after taking a leave from a graduate computer science program at Stanford where they worked on a "digital libraries" project. "What we first discussed at Stanford is now becoming practical," Mr. Page said.

At Stanford, Google hopes to be able to scan 50,000 pages a day within the month, eventually doubling that rate, according to a person involved in the project.

The Google plan calls for making the library materials available as part of Google's regular Web service, which currently has an estimated eight billion Web pages in its database and tens of millions of users a day. As with the other information on its service, Google will sell advertising to generate revenue from its library material. (In its existing Google Print program, the company shares advertising revenue with the participating book publishers.)

Each library, meanwhile, will receive its own copy of the digital database created from that institution's holdings, which the library can make available through its own Web site if it chooses.

Harvard officials said they would be happy to use the Internet to share their collections widely. "We have always thought of our libraries at Harvard as being a global resource," said Lawrence H. Summers, president of Harvard.

At least initially, Google's digitizing task will be labor intensive, with people placing the books and documents on sophisticated scanners whose high-resolution cameras capture an image of each page and convert it to a digital file.

Google, whose corporate campus in Mountain View, Calif., is just a few miles from Stanford, plans to transport books to a copying center it has established at its headquarters. There the books will be scanned and then returned to the Stanford libraries. Google plans to set up remote scanning operations at both Michigan and Harvard.

The company refused to comment on the technology that it was using to digitize books, except to say that it was nondestructive. But according to a person who has been briefed on the project, Google's technology is more labor-intensive than systems that are already commercially available.

Two small start-up companies, 4DigitalBooks of St. Aubin, Switzerland, and Kirtas

Technologies of Victor, N.Y., are selling systems that automatically turn pages to capture images.

[Copyright 2004 The New York Times Company](#) | [Home](#) | [Privacy Policy](#) | [Search](#) | [Corrections](#) | [RSS](#) | [Help](#) | [Back to Top](#)