The Infinite Library

By Wade Roush

May 2005 - TechnologyReview.com

The Bodleian Library at the University of Oxford in England is the only place you are likely to find an Ethernet port that looks like a book. Built into the ancient bookcases dominating the oldest wing of the 402-year-old library, the brown plastic ports share shelf space with handwritten catalogues of the university's medieval manuscripts and other materials. Some of the volumes are still chained to the shelves, a 17th-century innovation designed to discourage borrowing. But thanks to the Ethernet ports and the university's effort to digitize irreplaceable books like the catalogues—which often contain the only clue to locating an obscure book or manuscript elsewhere in the vast library—users of the Bodleian don't even need to take the books off the shelves. They can simply plug in their laptops, connect to the Internet, and view the pertinent pages online. In fact, anyone with a Web browser can read the catalogues, a privilege once restricted to those fortunate enough to be teaching or studying at Oxford.

The digitization of the world's enormous store of library books—an effort dating to the early 1990s in the United Kingdom, the United States, and elsewhere—has been a slow, expensive, and underfunded process. But last December librarians received a pleasant shock. Search-engine giant Google announced ambitious plans to expand its "Google Print" service by converting the full text of millions of library books into searchable Web pages. At the time of the announcement, Google had already signed up five partners, including the libraries at Oxford, Harvard, Stanford, and the University of Michigan, along with the New York Public Library. More are sure to follow.

Most librarians and archivists are ecstatic about the announcement, saying it will likely be remembered as the moment in history when society finally got serious about making knowledge ubiquitous. Brewster Kahle, founder of a nonprofit digital library known as the Internet Archive, calls Google's move "huge....It legitimizes the whole idea of doing large-volume digitization."

But some of the same people, including Kahle, believe Google's efforts and others like it will force libraries and librarians to reëxamine their core principles—including their commitment to spreading knowledge freely. Letting a for-profit organization like Google mediate access to library books, after all, could either open up long-hidden reserves of human wisdom or constitute the first step toward the privatization of the world's literary heritage. "You'd think that if libraries are serious about providing access to high-quality material, the idea of somebody digitizing that stuff very quickly—well, what's not to like?" says Abby Smith, director of programs for the Council on Library and Information Resources, a Washington, DC, nonprofit that helps libraries manage digital transformation. "But some librarians are very concerned about the terms of access and are very concerned that a commercial entity will have control over materials that libraries have collected."

They're also concerned about the book business itself. Publishers and authors count on strict copyright laws to prevent copying and reuse of their intellectual property until after they've recouped their investments. But libraries, which allow many readers to use the same book, have always enjoyed something of an exemption from copyright law. Now the mass digitization of library books threatens to make their content just as portable—or piracy prone, depending on one's point of view—as digital music. And that directly involves libraries in the clash between big media companies and those who would like all information to be free—or at least as cheap as possible.

Whatever happens, transforming millions more books into bits is sure to change the habits of library patrons. What, then, will become of libraries themselves? Once the knowledge now trapped on the printed page moves onto the Web, where people can retrieve it from their homes, offices, and dorm rooms, libraries could turn into lonely caverns inhabited mainly by preservationists. Checking out a library book could become as anachronistic as using a pay phone, visiting a travel agent to book a flight, or sending a handwritten letter by post.

Surprisingly, however, most backers of library digitization expect exactly the opposite effect. They point out that libraries in the United States are gaining users, despite the advent of the Web, and that libraries are being constructed or renovated at an unprecedented rate (architect Rem Koolhaas's Seattle Central Library, for example, is the new jewel of that city's downtown). And they predict that 21st-century citizens will head to their local libraries in even greater numbers, whether to use their free Internet terminals, consult reference specialists, or find physical copies of copyrighted books. (Under the Google model, only snippets from these books will be viewable on the Web, unless their authors and publishers agree otherwise.) And considering that the flood of new digital material will make the job of classifying, cata loguing, and guiding readers to the right texts even more demanding, librarians could become busier than ever.

"I chafe at the presumption that once you digitize, there is nothing left to do," says Donald Waters, a former director of the Digital Library Federation who now oversees the Andrew W. Mellon Foundation's extensive philanthropic investments in projects to en-

hance scholarly communication. “There is an enormous amount to do, and digitizing is just scratching the surface.”

Digitization itself, of course, is no small challenge. Scanning the pages of brittle old books at high speed without damaging them is a problem that’s still being addressed, as is the question of how to store and preserve their content once it’s in digital form. The Google initiative has also amplified a long-standing debate among librarians, authors, publishers, and technologists over how to guarantee the fullest possible access to digitized books, including those still under copyright (which, in the United States, means everything published after January 1, 1923). The stakes are high, both for Google and for the library community—and the technologies and business agreements being framed now could determine how people use libraries for decades to come.

“Industry has resources to invest that we don’t have anymore and never will have,” points out Gary Strong, university librarian at the University of California, Los Angeles, which has its own aggressive digitization programs. “And they’ve come to libraries because we have massive repositories of information. So we’re natural partners in this venture, and we all bring different skills to the table. But we’re redefining the table itself. Now that we’re defining new channels of access, how do we make sure all this information is usable?”

Breaching the Walls

Even for authorized users, access to the Bodleian Library’s seven million volumes is anything but instant. If you are an Oxford undergraduate in need of a book, you first send an electronic request to a worker in the library’s underground stacks. (Before 2000 or so, you would have handed a written request slip to a librarian, who would have relayed it to the stacks via a 1940s-era network of pneumatic tubes.) The worker locates the book in a warren of movable shelves (a space-saving innovation conceived in 1898 by former British prime minister William Gladstone) and places it in a plastic bin. An ingenious system of conveyor belts and elevators, also built in the 1940s, carries the bin back to any of seven reading rooms, where it is unpacked, and the book is handed over to you.

The process can take anywhere from 30 minutes to several hours. But once you finally have the book, don’t even think about taking it back to your dorm room for further study. The Bodleian is a noncirculating legal deposit library, meaning that it is en titled to a free copy of every book published in the United Kingdom and the Republic of Ireland, and it guards those copies jealously. The library takes in tens of thousands of books every year, but the legend is that no book has ever left its walls.

But a digital book needn’t be loaned out to be shared. And Oxford’s various libraries have already created digital images of many of their greatest treasures, from ninth-century illuminated Latin manuscripts to 19th-century children’s alphabet books. Most of these images can be examined at high resolution on the Web. The only catch is that scholars have to know what they’re looking for in advance, since very few of the digital

pages are searchable. Optical character recognition (OCR) technology cannot yet interpret handwritten script, so exposing the content of these books to today's search engines requires typing their texts into separate files linked to the original images. A three-person team at Oxford, in collaboration with librarians at the University of Michigan and 70 other universities, is doing just that for a large collection of early English books, but the entire effort produces searchable text for only 200 books per month. At that rate, making a million books searchable would take more than 400 years.

That's where Google's resources will make a difference. Susan Wojcicki, a product manager at Google's Mountain View, CA, campus and leader of the Google Print project, puts it bluntly: "At Google we're good at doing things at scale."

Google has already copied and indexed some eight billion Web pages, which lends credibility to its claim that it can digitize a big chunk of the 60 million volumes (counting duplicates) held by Harvard, Oxford, Stanford, the University of Michigan, and the New York Public Library in a matter of years. It will be a complex task, but one that is in some ways familiar for the company. "It's not just feeding the books into some kind of digitization machine, but then actually taking the digital files, moving those files around, storing them, compressing them, OCR-ing them, indexing them, and serving them up," points out Wojcicki. "At that point it becomes similar to all of Google's other businesses, where we're managing large amounts of data." But the entire project, Wojcicki admits, hinges on those digitization machines: a fleet of proprietary robotic cameras, still under development, that will turn the digitization of printed books into a true assembly-line process and, in theory, lower the cost to about $10 per book, compared to a minimum of $30 per book today.

Neither Google nor its partner libraries have announced exactly how the process will work. But John Wilkin, associate university librarian at the University of Michigan, says it will go something like this: "We put a whole shelfful of books onto a cart, keeping the order intact. We check them out by waving them under a bar code reader. Overnight, software takes all the bar codes, extracts machine-readable records from the university's electronic catalogue, and sends the records to Google, so they can match them with the books. Then we move the cart into Google's operations room."

This room will contain multiple workstations so that several books can be digitized in parallel. Google is designing the machines to minimize the impact on books, according to Wilkin. "They scan the books in order and return the cart to us," he continues. "We check them back in and mark the records to show they've been scanned. Finally, the digital files are shipped in a raw format to a Google data center and processed to produce something you could use."

The Book Web

Exactly how readers will be able to use the material, however, is still a bit foggy. Google will give each participating library a copy of the books it has digitized while keeping another for itself. Initially, Google will use its copy to augment its existing Google Print pro-

gram, which mixes relevant snippets from recently published books into the usual results returned by its Web search tool. A user who clicks on a Google Print result is presented with an image of the book page containing his or her keyword, along with links to the sites of retailers selling the print version of the book and keyword-related ads sold to the highest bidders through Google's AdSense program.

Does it bother librarians that Moby-Dick might be served up alongside an ad for the latest Moby CD? "To say we haven't worried about it would be wrong," says Wilkin. "But Google has a 'good citizen' profile. The way they use AdSense doesn't trouble me. And if suddenly access were controlled, and there was a cost to view the materials, we could still offer them for free ourselves, or at least the out-of-copyright materials."

In fact, Google may put the entire texts of these public- domain materials online itself. In the future, Google could even use those materials to create a kind of literary equivalent of the Web, says Wojcicki. "Imagine taking the whole Harvard library and saying, 'Tell me about every book that has this specific person in it.' That in itself would be very powerful for scholars. But then you could start to see linkages between books"—that is, which books cite other books, and in what contexts, in the same way that websites refer to other sites through hyperlinks. "Just imagine the power that that would bring!"

(Wojcicki's example shows how history can, indeed, come full circle. Google founders Larry Page and Sergey Brin developed BackRub, the predecessor to the Google search engine, while working on an early library digitization project at Stanford that was funded in part by the National Science Foundation's Digital Libraries Initiative. And PageRank, Google's core search algorithm, which orders sites in search results based on the number of other sites that link to them, is simply a computer scientist's version of citation analysis, long used to rate the influence of ar ticles in scholarly print journals.)

The Michigan library, says Wilkin, may do whatever it likes with the digital scans of its own holdings—as long as it doesn't share them with companies that could use them to compete with Google. Such limitations may prove uncomfortable, but most librarians say they can live with them, considering that their holdings wouldn't be digitized at all without Google's help.

Closed Doors?

But others are more cautious about the leap Google's partner libraries are taking. Brewster Kahle, who is often described as an inspiring visionary and sometimes as an impractical idealist, founded the nonprofit Internet Archive in 1996 under the motto "universal access to human knowledge." Since then, the archive has preserved more than a petabyte's worth of Web pages (a peta byte is a million gigabytes), along with 60,000 digital texts, 21,000 live concert recordings, and 24,000 video files, from feature films to news broadcasts. It's all free for the taking at www.archive.org, and as you might guess, Kahle argues that all digital library materials should be as freely and openly accessible as physical library materials are now.

That's not such a radical idea; free and open access is exactly what public libraries, as storehouses of printed books and periodicals, have traditionally provided. But the very fact that digital files are so much easier to share than physical books (which scares publishers just as MP3 file sharing scares record companies) could lead to limits on redistribution that prevent libraries from giving patrons as much access to their digital collections as they would like. "Google has brought us to a tipping point that could define how access to the world's literature may proceed," Kahle says.

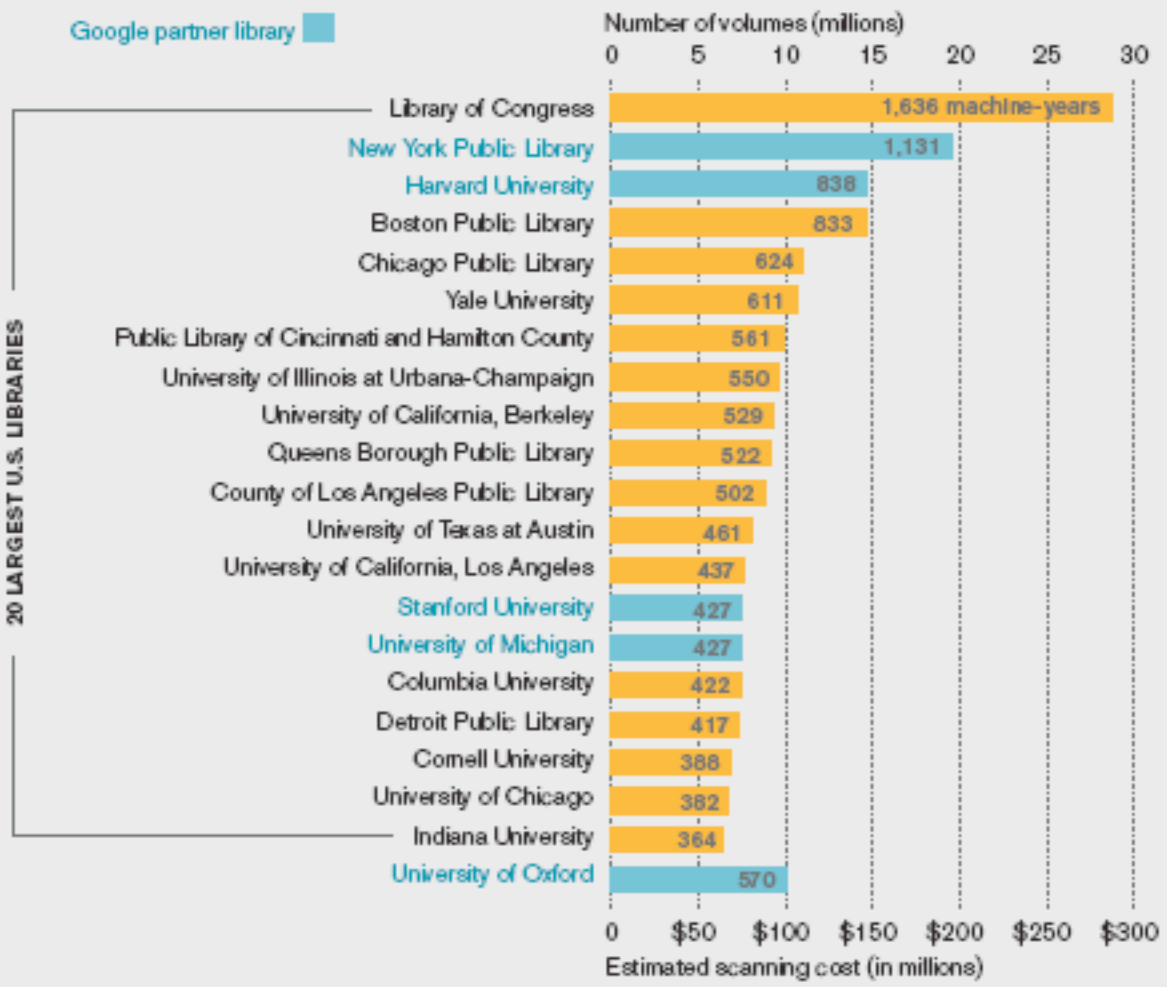
In Kahle's view, every previous digitization effort has followed one of three paths; with a bit of oratorical flourish, he calls them Door One, Door Two, and Door Three. (Kahle acknowledges up front that his picture is simplified, and that these aren't necessarily the only paths open to libraries today.)

Door One, says Kahle, is epitomized by Corbis, an image- licensing firm owned by Microsoft founder Bill Gates. Since the early 1990s, Corbis has acquired rights to digital reproductions of works from the National Gallery of London, the State Hermitage Museum in St. Petersburg, Russia, the Philadelphia Museum of Art, and more than 15 other museums. In some cases, it's now impossible to use these images without paying Corbis. "This organization got its start by digitizing what was in the public domain and essentially putting it under private control," says Kahle. "The same thing could happen with digital literature. In fact, it's the default case."

Behind Door Two, parallel public and private databases coexist peacefully. Here Kahle cites the Human Genome Project, which culminated in two versions of the DNA sequence of the human genome—a free version produced by government-funded scientists and a private version produced by Rockville, MD–based Celera Genomics and used by pharmaceutical companies to identify new drug candidates. The model has worked well in genomics, and Google seems to be setting out on a similar path, as it keeps one copy of each library's collection for itself and gives away the other. Kahle worries, however, that the restrictions Google imposes on libraries will prevent them from working with other companies or organizations to disseminate digital texts. Libraries might be barred, for example, from contributing material to projects such as the Internet Archive's Bookmobile, a van with satellite Internet access that can download and print any of 20,000 public-domain books.

Books to Bits

Google's digitization effort will be expensive and time consuming. The graph below shows how long a single book-scanning machine would take to scan the collections of the largest libraries in the United States, plus the University of Oxford. Of course, the more machines Google uses, the sooner they'll finish.



NOTES: SCANNING MACHINE TIME ASSUMES AN AVERAGE SPEED OF 30 MINUTES PER BOOK. SCANNING COSTS BASED ON AN ESTIMATE OF $10 PER BOOK. SOURCES: AMERICAN LIBRARY ASSOCIATION, LIBRARY OF CONGRESS, NEW YORK PUBLIC LIBRARY

Door Three, Kahle's favorite, hinges on new partnerships in which private companies offer commercial access to digital books while public entities, such as libraries, are allowed to provide free access for research and scholarship. Here his main example is the Internet Archive's collaboration with Alexa, a company founded by Kahle himself in 1996 and sold to Amazon in 1999. Alexa ranks websites according to the traffic they attract, and its servers, like Google's, constantly crawl the Internet, making copies of each page they find. But after six months, Alexa donates those copies to the Internet Archive, which preserves them for noncommercial use. "Jeff [Bezos, Amazon's CEO] was okay with the idea that there are some things you can exploit for commercial purposes for a certain amount of time, and then you play the open game," says Kahle. "Libraries and

publishing have always existed in the physical world without damaging each other; in fact they support each other. What we would like to see is this tradition not die with this digital transformation."

So which alternative comes closest to Google's plans? Google is no Corbis, says Wojcicki, but is nonetheless limited in what it can share. "Door One was never our intention, nor is it even practical," she says. "And we can't do Door Three, because we're not the rights holders for much of this material. So Door Two is proba bly where we're headed. We're trying to be as open as possible, but we need to hold to our agreements with different parties."

Precisely to avoid questions about copyright, Oxford librarians have decided that only 19th- and early 20th-century books will be handed over to Google for digitization. "Some of the other libraries, including Harvard, have agreed to have some in- copyright material digitized," says Ronald Milne, acting director of the Bodleian Library. "They are quite brave in taking it on. But we didn't particularly want to go there, because it's such a hassle, and we didn't want to get on the wrong side of the book laws."

At the same time, though, the American Library Association is one of the loudest advocates of proposed legislation to reinforce the "fair use" provisions of federal copyright law, which entitle the public to republish portions of copyrighted works for purposes of commentary or criticism. And two of Google's partner universities—Harvard and Stanford—are also supporters of the Chilling Effects Clearinghouse, a website that monitors allegations of copyright infringement brought against webmasters, bloggers, and other online publishers under the controversial Digital Millennium Copyright Act (DMCA) of 1998. Mass digitization may eventually force a redefinition of fair use, some librari ans believe. The more public-domain literature that appears on the Web through Google Print, the greater the likelihood that citi zens will demand an equitable but low-cost way to view the much larger mass of copyrighted books. "I think this will be another piece of good pressure, another factor in the whole debate over the DMCA," says Wilkin.

The Mixing Chamber

If you're over 30, today's libraries are probably nothing like the ones you remember from childhood. Enter any major library today and you'll find an armory of computers and a platoon of specialists, from the reference librarians who are expert at accessing online resources, to the acquisitions officers who decide which books, CDs, DVDs, and subscriptions to purchase, to the computer geeks who keep the building's network running.

Digitization and the growing power of the Internet are making all of these people's jobs more complex. Acquisitions experts, for example, can no longer just rely on the traditional quality filter imposed by the publishing industry; they must evaluate a much larger mass of material, from newly digitized print books to the millions of Web pages, blogs, and news sites that are born digital. "On the Internet, publishing is a promiscuous activity," observes Abby Smith of the Council on Library Information and Resources. "Libraries are confused and challenged about how to collect and select from that material."

Then there are the problems of cataloguing and preserving digital holdings. Without the proper "metadata" attached— author, publisher, date, and all the other information that once appeared in libraries' physical card catalogues—a digital book is as good as lost. Yet creating this metadata can be laborious, and no international standard has emerged to govern which kinds of data should be recorded. And considering the limited life span of each new data format or electronic storage medium (have you used a floppy disk lately?), keeping digital materials alive for future generations will, ironically, be much more costly and complicated than simply leaving a paper book on a library shelf.

But even if every book is reduced to a few megabytes of 1s and 0s residing on some placeless Web server, libraries themselves will probably endure. "There is no one in the field of librarianship who thinks the library is disappearing as a physical space," says Smith. Seattle's exuberant new Central Library, for example, is built around a four-story spiral ramp that enables an unprecedented immediacy of access to its physical book collection. But at the same time, the library provides 400 public-use computers (compared to 75 in the library that previously occupied the site), buildingwide Wi-Fi access, and a high-tech "mixing chamber" where an interdisciplinary reference team uses an array of print and electronic resources to answer patrons' questions. More than 1.5 million people visited the new library in 2004—almost three times the entire population of Seattle.

"The real question for libraries is, what's the 'value proposition' they offer in a digital future?" says Smith. "I think it will be what it has always been: their ability to scan a large universe of knowledge out there, choose a subset of that, and gather it for description and cataloguing so people can find reliable and authentic information easily." The only difference: librarians will have a much bigger universe to navigate.

Stephen Griffin, the former director of the National Science Foundation's Digital Libraries Initiative (a Clinton-era project that funds a variety of university computer-science studies on managing electronic collections), takes a slightly different view. Ask him how he thinks libraries will function in 2020 or 2050—once Google or its successors have finished digitizing the world's printed knowledge—and he answers from the reader's point of view. "The question is, how will people feel when they walk into libraries," he says. "I hope they feel the same—that this is a very welcoming place that is going to help them to find information that they need. As we bring more technology in, the notion of libraries as places for books may change a bit. But I hope people will always find them a comfortable place for thinking."

Good time to get your copy of The Infinite Library and Other Stories. Shortlisted for both the United Kingdomâ€™s 2018 International Rubery Book Award for the best books by independent writers, self-published authors, and books published by independent presses, and by the Asian Books Blog for the 2017 Book of the Lunar Year Award. "The Library of Babel" (Spanish: La biblioteca de Babel) is a short story by Argentine author and librarian Jorge Luis Borges (1899â€“1986), conceiving of a universe in the form of a vast library containing all possible 410-page books of a certain format and character set. The story was originally published in Spanish in Borges' 1941 collection of stories El JardÃn de senderos que se bifurcan (The Garden of Forking Paths). That entire book was, in turn, included within his much-reprinted Ficciones (1944)