The Brave New World of Book Research

For research over the Internet, one or the last great
frontiers still largely unexplored is that of books.
Say, for instance, that you are looking For background
information to help with a business presentation. For
some time now you have been able to search the Web For
government data; information put out by companies,
nonprofits, and hobbyists; and newspaper and magazine
articles. You can use free web-search tools such as
Google and Yahoo! or professional services such as
Dialog (www.dialog.com) and LexisNexis (www.lexis-
nexis.com).

But, with some exceptions, you cannot search through
the huge repository of knowledge found in books. You
need to go to a library for that. This situation will
change.

The two largest and most prominent Internet search
companies have recently locked horns in what will
likely be an epic battle to bring consumers the same
productivity and convenience benefits with books that
are derived from using the Web to search through other
media.

Google, with Google Print (print.google.com), and
Yahoo! with the Open Content Alliance
(www.opencontentalliance.org), are taking previous
book-archiving efforts to a new level.

The first widely accessible effort to place the full
text of a large number of books online was Project
Gutenberg (www.gutenberg.org), where you can currently
download, read, or search through more than 16,000
books.

These are primarily out-of-copyright volumes published
before 1923, from Shakespeare's work to the Sherlock
Holmes stories of Arthur Conan Doyle. You can also find
interesting odds and ends, such as the Human Genome
Project and pictures of prehistoric cave paintings from
the south of France--all free of charge.

Another useful book site that has been around for a while is the On-Line Books Page (onlinebooks.library.upenn.edu), where you can find a listing of more than 20,000 books that are freely available through the Web. Unlike Project Gutenberg, at this site you cannot search for words in the books, but you can search by title and author for books whose full text is available on the Web. The site features a section where you can link to books that were once banned, such as Ulysses and Fanny Hill.

Yet another oldie but goodie is the Internet Public Library (www.ipl.org), which provides a window not only to full-text books on the Web but also to almanacs, dictionaries, quotation collections, directories of associations, and more.

But Google and Yahoo! are pushing ahead into new territory. Google is working with Harvard University, Stanford University, Oxford University, the University of Michigan, and the New York Public Library to digitize all or portions of their collections and make them searchable by keyword. This project represents potentially hundreds of thousands--even millions--of books.

You will be able to read the full text of out-of-copyright books, small portions of library books still under copyright, and new books--for free. Searching through new books this way is similar to Amazon.com's feature "Search Inside This Book."

With Google Print, new books will be accompanied by links to book reviews, publishers, and online booksellers that have the particular books in stock. Library books will be accompanied by links where you can type in your zip code to find a library near you that has the book.

Announced in October 2004, Google Print is still in beta, or preview, mode, but it is open to the public and worth a look. One limitation, although a small one, is that the books digitized by Google will be searchable only through Google. In response, Yahoo! has recently teamed with the University of California, the University of Toronto, the National Archive in England, Microsoft, and others to create the Open Content Alliance.

As with Google Print, Yahoo! plans to digitize hundreds
of thousands of books, but it will make them searchable
through Yahoo! and any other search engine, including
Google. Announced last year, its efforts are not as far
along as Google Print's.

The new is often controversial, and these efforts are
no exception. The Authors Guild has sued Google for
copyright infringement, contending that Google Print
will hurt authors. There are competing interests here,
all legitimate. On the one hand, authors need
compensation for their work to undergo the considerable
efforts in producing them. On the other hand,
researchers and the world in general benefit from
having easily accessible knowledge. The fair use
doctrine of the copyright law is designed as a
compromise between these interests, allowing for wide
use of a small portion of copyrighted works.

Though I am an author, I believe that the Authors Guild
is overreacting. But complex issues are involved, which
you can read about at lwn.net/Articles/153452.

~~~~~~~~

By Reid Goldsborough