Science & Technology

Overbooked?

Universities are digitizing their book collections to create one enormous Internet library. The unlimited access is changing scholarship and the way we read — and not always for the best.

By
David J. Craig
|
Spring 2008



Keith Negley

It was the original hyperlink. You'd wander into the library stacks with a single call number on a scrap of paper, and then, having located, say, a John Adams biography, you'd spot a collection of his letters to Thomas Jefferson, and a few feet away, Adams's correspondence with his wife, Abigail. You'd plop down cross-legged on the floor, flip through one book after another, and gradually see your research in a new light. Adams, behind his stateliness, was a vulnerable man who could be insecure, playful, and with Abigail, unexpectedly flirtatious: This is rather too coarse a Compliment, but you are so saucy!

Discoveries like this don't take place in the library stacks as often as they used to. Columbia faculty and students now can read some 700,000 books and 80,000 journals simply by logging onto the University's library Web site. Consider, too, the countless rare manuscripts beamed up to Columbia's library site from archives around the world and you might wonder why scholars leave their offices or dorm rooms at all. Forget searching the shelves; it's the keyword search that drives a lot of scholarship in the humanities today.

"I have at my fingertips nearly every word that Abraham Lincoln ever wrote and every letter ever written to him," says Columbia history professor Eric Foner '63CC. "I can accomplish in a few hours at my computer what once would have taken me weeks, months even. To say this has made my research easier or quicker doesn't begin to describe it."

Libraries have digitized so much literature that new types of scholarship are emerging. Historians and literary scholars, for instance, are tracing the evolution of words and ideas across old texts, and in so doing, they are nudging their disciplines toward quantitative analysis.

But the virtual library poses tough new questions for academia: Are students becoming too dependent on the Internet? Are we ignoring books that aren't digitized? And are libraries, by sharing online their unique collections, making themselves irrelevant?

Real page-turners

In the basement of Butler Library, a bibliographic assistant named Baojing Liang stares down at an original copy of William Leete Stone's 1872 book History of New York City, which is cradled carefully in a copy stand. It hasn't been checked out of Butler in six years and its pages have fallen loose from the binding. Liang focuses a digital camera that's suspended over the book, lays a sheet of glass atop one of the open pages, and snaps a picture. Then he lifts up the glass, places it on the facing page, and presses the shutter again.

Creating an e-book is laborious. "This will take a day, maybe more," says Liang. He can finish less fragile volumes in a couple of hours, he says. Then comes the really time-consuming part: Librarians will create Web files that let readers search for words within the text images. The files also must make the book's complicated bibliographic data recognizable to Internet search engines. The whole process will take several days.

Librarians at Columbia have been creating digital facsimiles of rare books and manuscripts for more than 10 years. The first items to go under the camera, back in the mid-1990s, were medieval manuscripts and ancient writings. Next were thousands of items from Columbia Special Collections such as the John Jay Papers, the Greene & Greene Architectural Drawings, and more recently, Barbara C. Adachi's photographs of Japanese Bunraku puppet theatres from the 1920s.

It was only last year that Columbia librarians started digitizing books from the general collection. "Professors might ask for an e-book they want to use in their research or in the classroom," says Janet Gertz, director of preservation and digital conversion for Columbia's library system. "They want a version they can search." The University now employs some 10 people devoted to digitizing library materials, almost as many people as are focused on traditional book preservation and restoration. "It's grown gradually, but has picked up a lot in the last couple of years," Gertz says.

When a dusty old book from Butler is digitized, it first appears on Columbia's library Web site. But History of New York City, say, will have really hit the big time when it's added to an academic database. That's a Web site that collects materials on a single subject, such as Latin hagiography, Renaissance painting, or the history of corporations. These databases unite library holdings scattered across many different institutions in order to give scholars easy access to primary materials in their field.

When depicting old manuscripts, the databases show an image of each page, translations of the text, and lots of scholarly notations.

Many of these databases are created by libraries or library consortiums and are freely accessible, as is the case with a local history database that Columbia is now creating called Digital New York, which will launch later this year and feature History of New York City. Among the some 900 databases available on Columbia's library Web site, however, more than half are commercial sites that get their digital content from libraries in exchange for small royalties. Some of these commercial sites are run by academic publishers and can cost the University up to $200,000 a year in subscription fees and require a Columbia ID to access.

"It's the academic databases that are revolutionizing how we study texts," says Robert Scott '74GSAS, who heads the Electronic Text Service at Butler Library and helps scholars use these tools.

Searched and researched

Eric Foner used to spend his summers sifting through piles of historical documents in small archives scattered across the United States. Today, he's as likely to perch himself in front of a computer, think of a few terms that crop up frequently in his reading, and use simple keyword searches to comb databases for them. "I don't consider myself a high-tech person, and I still visit the archives," he says. "But the amount of historical material that's available now for historians online is unbelievable."

Foner is writing a book today that he says would have been prohibitively time-consuming to research properly five years ago. To understand how Abraham Lincoln's views on slavery were influenced by public debate over the course of his life, Foner is examining online nearly all of Lincoln's correspondence, transcripts of congressional debates he participated in, and huge numbers of 19th-century newspapers, magazines, and antislavery pamphlets housed in archives from Alabama to Ohio. "By searching for terms like 'gradual emancipation' or 'colonization,' which referred to the practice of sending blacks out of the country, I can spot historical patterns very quickly, and I know I'm not missing anything," says Foner, who is widely considered to be the leading contemporary historian of

American Civil War Reconstruction. “This saved me years.”

Some historians are conducting entirely new types of research with databases. Dan Cohen, a George Mason University history professor, for instance, recently analyzed on his computer tens of thousands of personal testimonials that Americans wrote about their experience on September 11, 2001, in order to map the regions where the largest numbers of people reported praying that day. Other historians, like Foner, are using databases primarily to access a larger number of sources and to work more efficiently. “I would say that databases are changing how most people in the field work,” says Robert Townsend, an assistant director of the American Historical Association who writes about research trends. “Theorists haven’t tackled yet what this means for the nature of historical research.”

In English departments, trawling databases for language patterns tends to be controversial, in part because it’s at odds with the tradition of studying great works in the context of the canon. “Literary scholarship has always been about narrowing the number of works we study, making more out of less,” states Clifford Siskin, a former Columbia English professor who now teaches at New York University. “What’s exciting about database research, and potentially threatening to English departments, is that we can suddenly step back and examine all of 18th-century literature as one big, searchable text. New types of knowledge are going to come out of this, although it’s too early to say what it will look like, exactly.”

Especially suitable for database research is British and American literature from the 18th and early 19th centuries. Former microfilm publishing companies like Gale, ProQuest, and Chadwyck-Healey in recent years have digitized the vast majority of English-language creative works published from 1700, when typography becomes neat enough for today’s computers to recognize, until about 1840, when the publishing industry exploded.

Alice Boone ’09GSAS, a PhD candidate in English literature, scours databases that cover the 18th century to pinpoint when British book editors started publishing ironic footnotes poking fun at elaborate forms of error correction that developed during the Johnsonian period. She explains that by searching various types of literature for terms like “errata” and “footnote,” she’s stumbling across “relevant material that no one in my field would have been familiar with, like a cultural study of China put out by a small British publisher. . . . This introduces a new element of serendipity to my

work.”

Digitized texts are influencing the work of humanities scholars who don’t rely heavily on database research, too. Consuelo Dutschke, curator of medieval and Renaissance manuscripts for Columbia’s libraries, was preparing a lecture recently about the cursive typescript Bâtarde when she recalled that it was considered too informal for medieval church or legal texts written in Latin. To confirm the point, she examined a database that contains images of some 90,000 manuscripts containing Latin text. “I found two or three in Bâtarde,” she says. “So whereas before I would have said that we think Bâtarde was rarely used for Latin, now I’ll say that we know it. This is how the humanities are becoming a bit more empirical. We have access to more data now, so we’re demanding harder evidence.”

Lost works?

Database analysis has technical shortcomings, though. Old books tend to contain lots of typographical errors, for starters. This can throw off keyword searches, as can the slightest imperfections in digital images that compose e-books, Scott says.

Typos are the least of some scholars’ concerns. Jenny Davidson, a Columbia associate professor of English, worries that students will waste time on Internet fishing expeditions, drowning in data when they should be boning up on their Hemingway. “There’s widespread concern about there being overenthusiasm for this kind of thing,” she says. “The applications are quite narrow, I think, more for a historical type of literary scholarship. We don’t want to see lots of graduate students doing projects where they don’t distinguish between important books and totally obscure texts.”

Isn’t it possible that a keyword search could unearth, say, a long-forgotten memoir by the confidant of an important author? Samuel Roberts, a Columbia assistant professor of history and sociomedical sciences, says researchers might be better off exploring physical archives or even the stacks. “Sometimes you need to sit down and read 50 pages of a book before you really appreciate what you’re looking at,” he says. “It’s easy to not peruse on the Internet. You type in a keyword and zoom in on that one passage that a search engine spits back.”

Foner, too, sees a dilemma: “Students are starting to think that research is all about the Internet. If it isn’t online, it might as well not exist.”

In fact, libraries have digitized merely a sliver of their holdings. Columbia has converted less than one-tenth of 1 percent of its special-collections material — all of those letters, first drafts of novels, and other primary-source materials that are considered essential to professional scholarship. Collections that get digitized tend to be those in hot demand among Columbia faculty, who often work with librarians to solicit grants for digital conversion projects.

What isn’t online? Roberts, an expert on tuberculosis trends among African Americans, says the public health records that he studies reside in small private and government depositories across the country that can’t afford to digitize their materials. He often brings students to archives at Columbia and around New York City to teach them how to conduct hands-on archival research. “Database analysis might be a viable research strategy if you’re studying the Civil War because all of that stuff is online for you,” Roberts says. “For less popular fields, that’s not necessarily the case. I remind students that what gets digitized is partially market driven, and that it’s slanted toward the United States and the English language, just like everything else on the Internet.”

Columbia students do get better exposure to the University’s physical archives now than ever before. The Rare Book and Manuscript Library has promoted the educational use of its collections vigorously under current director Michael Ryan. Last fall it hired a research librarian, Gerald Cloud, who’s devoted solely to helping faculty at Columbia and at nearby institutions incorporate the University’s rare-print materials in their teaching.

“We’ve found that when we put collections on the Web it actually brings more people into the archives to poke around,” says Ryan. “It piques interest in what we’re doing, which is what we want. We’re open for one, open for all.”

Equal-access artifacts

Universities haven’t always made their most valuable library materials accessible to the world. Departments of special collections in the U.S., which grew out of the antiquarian book trade in the prosperous years following World War II, were at first

“intensely territorial and object-focused places run by connoisseurs,” says Ryan. “They weren’t eager to share the wealth.”

That started to change in the 1970s and 1980s, when university administrations liberalized, and put pressure on, rare-books librarians to serve more scholars. “These departments needed to be brought in line with the library profession’s deep public service ethic,” Ryan says. “By the early 1990s, that had largely occurred.”

Some small groups of scholars retained their grip on library holdings, however. Papyrologists, who study ancient writings on papyrus manuscripts, are a case in point. Access to papyri has traditionally been “based on friendships, personal networks,” acknowledges Roger Bagnall, a prominent papyrologist and Columbia’s Jay Professor of Greek and Latin and Professor of History emeritus. The custom is to keep outside scholars away from papyri at your own library so you can publish on them first. Bagnall has never liked that, so in the mid-1990s he launched the Advanced Papyrological Information System (APIS), a free database operated by Columbia’s libraries in partnership with Duke and Tufts. APIS features images, searchable translations, and bibliographic data for thousands of Columbia’s own Latin, Greek, and Roman papyri, as well as for papyri from 10 other institutions.

Scholars at colleges around the world, by surfing APIS, can now see where papyri are housed and can easily request visits to view the real artifacts. “That makes it harder to restrict access,” says Bagnall. “And the broader access is leading to new kinds of work. We have lots of classicists and historians of early Christianity looking at this stuff now.”

That’s good news for the scholarly community at large, but it means that a prestigious university’s library holdings are losing some draw as a means of recruiting faculty and students. Librarians are quick to point out that old manuscripts don’t reveal all of their secrets on a computer screen, and that’s true: You can’t smell the acidic odor that identifies a Renaissance-era palimpsest, a piece of parchment reused after its original ink was washed off with urine. And you might not see online that a medieval manuscript has hairlike fibers protruding from both of its sides, which indicates it was made in Germany, where artisans didn’t scrape one side smooth with a knife. But there are plenty of things you can tell about an old book or manuscript online. And reading the text, seeing the basic condition of the paper, or skimming the bibliographic data in many instances tells scholars all they

need to conduct their research.

“There isn’t the same imperative now for scholars to be at a university whose archives align with their research interests,” says Christian Dupont, a librarian at the University of Virginia who chairs the American Library Association’s rare books and manuscripts division. “To a university, the benefit of investing in a great library now is becoming more about the prestige.”

It’s not just special collections that are opening up, either. Columbia and dozens of other U.S. research institutions have signed deals with Google and Microsoft in the past couple of years, allowing them to digitize library books for their competing e-book sites (see sidebar on page 30).

Does public availability of all of this content devalue the holdings of private research libraries? James Neal ’73GSAS, Columbia’s top librarian and vice president for information services, says sharing information is an ethical decision for a university before it is a business decision. “The notion that our research collections would be an exclusive resource for those with the wherewithal, the knowledge, and the resources to travel to New York is a very elitist view,” he says sharply. “The value of our material is in its use. In any case, the world still produces lots of print-based information, we continue to buy it, and only a small percentage of it is going online anytime soon.”

Even if Ivy League libraries could magically digitize all of their holdings, access to the information wouldn’t be entirely democratized because the materials are organized most usefully in academic databases, most of which are very expensive. Popular search sites like Google, scholars say, generate results that are too vast and too comprehensive to be useful in conducting serious research. “You get 90,000 hits for everything,” says Foner. “I can’t use that.” Only top institutions like Columbia, furthermore, license some of the most expensive databases, like ProQuest’s collection of historical African American newspapers, or Gale’s Web site The Making of Modern Law, which contains legal treatises from the past 200 years.

Still, it’s true, Neal says, that the Internet is “in many ways breaking down the barriers that have existed historically and have distinguished among different types of libraries. Now professors at Oberlin College, University of Wyoming, or Peking University can access much of the same archival material that our scholars have at Columbia. What’s still distinct, though, is the level of service you provide. How do

your librarians work with individual scholars to integrate information into teaching and research? That’s where depth exists at a place like Columbia.”

New directions

The academic library of the future, as Neal envisions it, is an outward-looking place where technologists support nearly every aspect of scholarship. Neal convened a task force this spring to create a Columbia University–wide digital repository that would store science lab data. Meanwhile, his team of Web experts maintains Gutenberg-e, an Internet site that features new history books rejected by publishers merely for lacking big sales potential. And last fall, Neal established the Copyright Advisory Office, which helps scholars understand intellectual-property laws related to online and print publishing.

According to Dupont, these are the kinds of gaps that libraries are trying to fill at universities across the country in order to stay relevant. He says that while library Web sites provide the best tools for online research, there is widespread anxiety among librarians that corporate sites like Google Scholar, which provides access to a few hundred journals, will expand their content and refine search functions to better compete for researchers’ attention. “There are some ways in which libraries will never be able to keep up with Google, such as in system speed,” Dupont says. “When I go to the University of Virginia library catalog and it takes five seconds to download an entry, that’s too long. There’s a real threat there.”

Of course, academic libraries have an advantage over Web companies: They know scholars’ needs intimately. To meet those needs, Neal is planning three new centers for digital research: One opens in the Lehman Social Sciences Library this fall, another in Butler Library for the humanities in the fall of 2009, and another in the new Interdisciplinary Science Building on the northwest corner of the Morningside Heights campus in the fall of 2010. The Digital Humanities Center will train scholars in the rudiments of database research as well as in high-end applications such as cluster analysis, which lets researchers identify words that appear in close proximity to one another in a given body of literature. “By searching for these groupings, you’re then looking for ideas rather than for particular words,” the librarian Robert Scott says. “You can see where a term enters a particular discourse, or when one

term replaces another."

But is it possible that librarians, in their eagerness to serve scholars' technological needs, will sprint too far ahead of them, spending money on research tools that might prove to be faddish? The typical humanities professor doesn't really want to be glued to a computer screen, does he? "I think that's wrong," Neal says. "Every discipline is moving in this direction, including those in the humanities. Just imagine all of the new things you can do. You can go into the Herbert Lehman papers and search for 'Israel,' or combine a search for 'Israel' with another scholar's name. These are such powerful tools for discovering new relationships across texts.

"It's not enough to give people access to lots of information now," he continues. "We need to help people use information differently, to do new things with it."

Google's stack attack

Google is coming, this much we know.

The California-based company this fall will start digitizing large numbers of Columbia's library books to add to Google Book Search, a Web site that Google loftily claims will one day feature a digital facsimile of every book ever published.

Which books will Google digitize? How many? How will the imaging be done, and where? What are the financial arrangements of the deal? Columbia librarians are mum because Google, which has made similar deals with dozens of other major research libraries throughout the U.S., insists that its business tactics remain secret. What the librarians can say is that the University will get copies of the electronic files Google creates, for use in teaching and research. Columbia also will create links from book entries in its online library catalog, CLIO, to corresponding e-books on Google Book Search.

There's reason to suspect that the project is big. Google can reportedly scan at least one million books annually, and Columbia, whose library system is the fifth largest in the United States, has lots of distinctive holdings to offer the company — titles that complement those Google has digitized already at institutions like Harvard, Stanford, the University of Michigan, and New York Public Library. "They're pretty voracious,"

says Janet Gertz, who is Columbia's top preservation librarian and is managing the University's work with Google. "We're going to give them as many books as we can."

But Columbia has laid down some parameters: Google may pick only books from the University's general collections, not rare books or artifacts, which require a level of care that mass digitization projects can't provide, Gertz says. And Google may digitize only books that are known to be in the public domain, which means those published before 1923. (Google Book Search has generated controversy by posting books from other libraries that are still copyright protected.)

Currently, Columbia libraries are allowing Microsoft to digitize books for its Live Search Books site as well. This collaboration is similar to the Google project, librarians say, but smaller in scope. Links to these e-books already have begun appearing in CLIO.

"Libraries want to share with the public as much of our material as possible," says Gertz, whose staff have digitized just a few hundred books on their own. "Google and Microsoft come at this at a scale that the academic world could never afford."

Read more from ***David J. Craig***

Guide to school abbreviations

All categories >

Read more from ***David J. Craig***