Turning an Info-Glut Into a Library

An exploding amount of information has made the Internet into the digital equivalent of a used book store. A new NSF program aims to bring some order to it

In Rome, the Vatican is digitizing its library, capturing rare and one-of-a-kind materials in electronic form. Soon any scholar with a personal computer and a modem—not just the favored few with Vatican library privileges—will be able to pore over such papal treasures as the four oldest known manuscripts of Virgil's poems or a magnificently illustrated copy of *The Divine Comedy*. Meanwhile, in Ithaca, New York, Cornell University is saving thousands of crumbling books by recording them on optical disks. Now they will survive as facsimiles, available to be viewed on computer screens or resurrected as hard copies.

Around the globe, information that was once analog—text, audio recordings, videos, maps, photographs, diagrams, scientific data—is turning digital, and much of it is flooding out into the vast public space created by electronic networks such as the Internet. "Almost overnight, we've gone from a high-mass, paper-based world to an almost massless, paperless world," says computer scientist Dan Atkins at the University of Michigan. But this digital cornucopia has a serious downside: The information has been growing much faster than the ability to keep up with it.

The Internet has come to resemble an enormous used book store with volumes stacked on shelves and tables and overflowing onto the floor, and a continuous stream of new books being added helter-skelter to the piles. Recently some relief has appeared, such as the Mosaic software program that makes it easy to browse through much of the information available on the Internet by creating "links" between related items at separate sites (*Science*, 12 August, p. 895). But even with Mosaic, finding what you're looking for is hit-or-miss. The solution, say Atkins and other experts, is to transform this used book store into one or several "digital libraries," complete with the electronic equivalents of neatly ordered shelves, a catalog, and a helpful staff.

Last week, the most ambitious effort so far to develop these virtual libraries got under way, with the announcement of major grants to six university-led consortia. Over 4 years, the National Science Foundation (NSF), together with the Advanced Research Projects Agency and the National Aeronautics and Space Administration, will spend $24.4 million to develop systems for collecting, storing, and organizing digital information and making it easily available to anyone who wishes to use it, from high school students to research scientists. Some of the consortia focus mainly on the problem of finding information scattered throughout cyberspace; others will concentrate on developing automated means to catalog new kinds of digital information, such as images and videos. And by including libraries, museums, publishers, schools, and computer and communications companies in the consortia, NSF hopes to lay the foundation of an infrastructure that will serve all parts of society.

"We view building partnerships between researchers, applications developers, and users as essential to achieving success," explains Y. T. Chien, director of NSF's division of information, robotics, and intelligent systems. At the end of the 4 years, each of the six teams should have created a working digital library that serves a large public while also acting as a testbed for further research and development. And after that, NSF hopes to scale up the libraries, adding more information, more advanced information-handling tools, and more users.

NSF is looking well into the future with this ambitious project, but it is following the lead of other digital library projects, aimed at a more immediate payoff. One of the more innovative was announced in August by IBM and the Institute for Scientific Information (ISI). The companies plan to create a prototype digital library based on 1350 scientific journals in the life sciences, complete with a catalog that allows searching by journal, author, title, subject matter, key words, and so on. That much is relatively standard on existing databases, but IBM is adding a twist: Instead of having every customer deal directly with one central library—which would demand tremendous communications and computing capacity—ISI will set up "branch libraries" for clients, such as universities or corporate research labs.

Each branch will be run on a server computer attached to a local area network. The server will hold a subcollection of the available works—customized to the needs of that customer—plus a catalog of everything that can be found in the main collection, which would be kept at one central location. Anyone on the network can search the entire catalog and request copies of any article.

NSF-FUNDED DIGITAL LIBRARY PROJECTS

Lead Institution (Amount)	Partners	Content
Carnegie Mellon University ($4.8 million)	Microsoft, DEC, Bell Atlantic, QED Communications, Open University, Fairfax VA County Schools	Digital video with math and science focus
University of California, Berkeley ($4 million)	Xerox, Resources Agency of California, California State Library, Sonoma County Library, San Diego Association of Governments, The Plumas Corp., Shasta County Office of Education, Hewlett Packard	Environmental information
University of Michigan ($4 million)	IBM, Elsevier Science, Apple Computer, Bellcore, UMI International, McGraw-Hill, *Encyclopaedia Britannica*, Kodak	Multimedia with earth and space science focus
University of California, Santa Barbara ($4 million)	State University of New York–Buffalo, University of Maine, industrial partners	Geographical information, including images and maps
Stanford University ($3.6 million)	Association for Computing Machinery, Bellcore, Dialog, EIT, Hewlett Packard, ITC, Interval Research, O'Reilly and Associates, WAIS Inc., NASA Ames, Xerox PARC	Technologies for a single, integrated virtual library
University of Illinois ($4 million)	National Center for Supercomputing Applications, University of Arizona, IEEE, APS, John Wiley & Sons, *U.S. News and World Report*	Engineering and science journals and magazines

SOURCE: TIM COLE, BILL MISCHO, BRUCE SCHATZ/UNIVERSITY OF ILLINOIS

If the article is held in the branch collection, it will appear on-screen immediately. If not, the user can ask the server to request a copy from a computer attached to the main collection.

The IBM-ISI group faces some tricky challenges. One is the problem—common to all digital information sources—of protecting intellectual-property rights. As IBM's Robert J. T. Morris notes, a customer who requests a copy of a journal article can be charged for it, but there is little to prevent that customer from distributing the article electronically to dozens of colleagues who pay nothing. The library's developers will also have to tackle difficult technical jobs, such as devising a means of quickly searching through tens of thousands of abstracts.

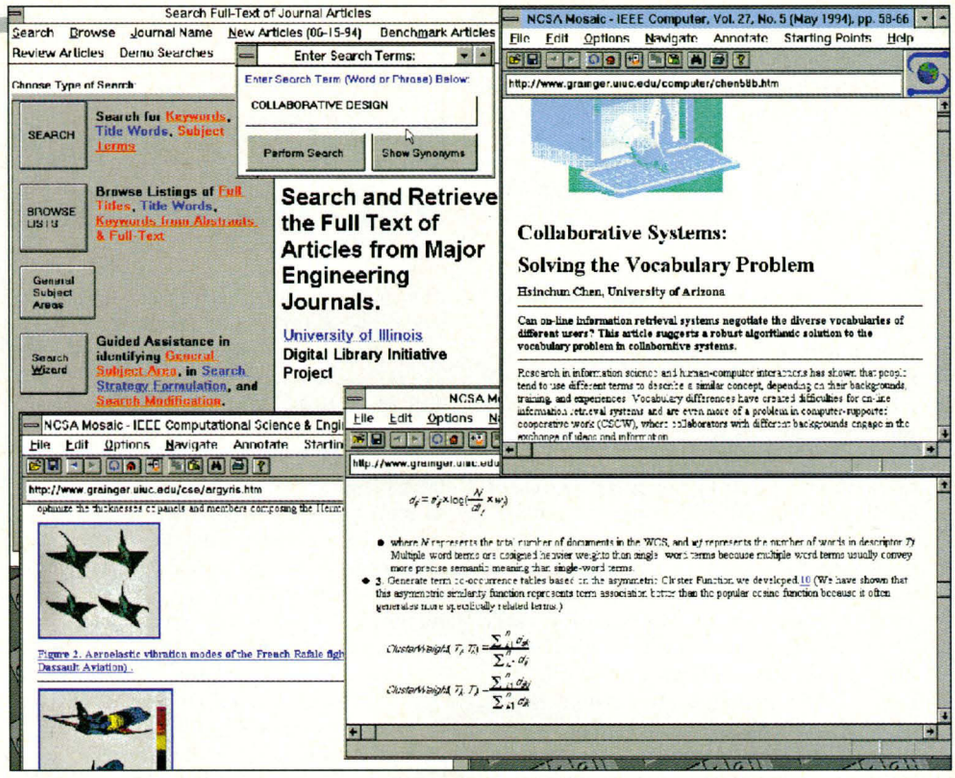
The center cannot hold

But in one important respect those challenges are easier than the one facing the six NSF-funded consortia. The contents of the IBM-ISI library will all be in one location—the main collection—making it simple to keep an up-to-date catalog for distribution to the various branches. That's not possible for any digital library that takes advantage of the riches of the Internet, where pertinent information may be scattered across dozens or even thousands of locations.

The obvious solution might seem to be to keep a central catalog and have each unit—the Vatican, Cornell, and every other information source encompassed by the library—report its contents there. But it's not that easy. "The problem with centralization is that it doesn't scale," explains Robert Wilensky, a computer scientist at the University of California, Berkeley, and principal investigator for one of the six groups that won NSF grants. Keeping track of information from several locations or even several hundred might be workable, but the resources needed to maintain a central record grow much faster than the number of locations.

As a result, Wilensky's group at Berkeley and several of the other groups will be exploring an alternative strategy: inventing electronic helpers that can visit the locations and report on what's there. With no middleman—no central catalog or clearinghouse—between the user and the information source, the result will be radically different from a traditional library, Wilensky says. Perhaps a better term for what he and the others are developing, he suggests, would be "digital nonlibraries."

One example of this kind of decentralization is the three-pronged strategy mapped out by a consortium centered at the University of Michigan. Atkins, the principal investigator, says his team will develop three types of software "agents," which will collaborate. The first will work with individual users, asking questions to determine their needs—not just the information they seek, but how much detail they want and how much they're willing to pay for it. A second group will take users' requests and range over the Internet to find the locations that can fill them. And the third group will be attached to individual information providers. Each of these agents can search its particular location for requested information and will know how much the information costs and whether it is restricted to certain users.



Check it out. The Illinois Digital Library testbed searches and retrieves items from its collection of full-text journal articles on science and engineering.

That approach brings the challenge of writing software that can—in a reasonable amount of time and with acceptable accuracy—search hundreds or perhaps thousands of locations, each with thousands of individual items or more. Making the job more difficult is the great variety of items within each collection. Many will be text or numerical data, but an increasing number will be photographs, drawings and other images, maps, and audio and video recordings. Each of these types of nontext documents will demand its own specialized searching methods.

The NSF-funded group at the University of California, Santa Barbara, for example, is at work on search techniques for digital images—in this case maps, satellite photos, and other geographical information. Terence Smith, a co-leader of the group, explains that people who use such a library will generally want to search by three criteria: location ("Find all maps and photos containing 34°N, 120°W"); content ("Find a map of Amelia Earhart's last flight"); and time—when the photo was taken or the map was drawn. The time and spatial coordinates are easy to catalog, but content is not. Conceivably, specially trained librarians could examine each map and image as it was digitized and enter an exhaustive list of the contents, but that would take too long and cost too much for the millions of images Smith and his colleague Michael Goodchild are eventually aiming for. Instead they will try to develop pattern-recognition software that can automatically characterize each image as it is added to the collection.

Multimedia libraries

While the Santa Barbara team is learning to put maps and images into a digital library, researchers at Carnegie Mellon University will be taking on an equally difficult task as part of another NSF project: assembling a digital library of video clips—a total of 1000 hours, to start. Video, with its moving images, demands innovative search techniques. To determine the content of each video, the Carnegie Mellon team will run the soundtrack through Sphinx 2, a speech-recognition system developed at the school, in tandem with a natural-language processor, a system that analyzes grammar and meaning. The library will also keep tabs on the images in the collection. Given an image from one video clip—of the Eiffel Tower, say—the search program will be able to seek other clips with the same structure, even if it's shown from a different perspective. This capability will rely on software that constructs a three-dimensional representation of an object and calculates what it would look like from different angles.

Assuming that such search techniques can be perfected, the ultimate goal will be to weave together many different kinds of libraries containing both text and nontext items into giant, multimedia libraries. Although the search software for different kinds of items may vary, says Terry Winograd, principal investigator for a consortium based at Stanford University, the user shouldn't have to worry about that. A student should be able to put in a single request for, say, information on the assassination of John F. Kennedy and receive a video of the shooting, a map of the final route JFK took through Dallas, text from newspapers, magazines, and books about the event, photos of Lee Harvey Oswald, and so on—all from dozens of sources around the Internet.

Besides the sheer power of having access to so much information, digital libraries offer another, potentially greater capability, says Bruce Schatz of the National Center for Supercomputing Applications and the University of Illinois, who is principal investigator for the Illinois project. In the past, libraries have provided for only a one-way flow of information—from the authors and other creators of information to the users. At most, a reader might write in the margins of a library book and thus offer something to future readers, but otherwise the information was fixed. Digital libraries offer a way for the users to play an active role.

The simplest way will be to offer annotating capabilities. Like the scribbler in a library book, users will be able to write notes to themselves and offer comments and extra information to any later user who wants to see what predecessors have added. Both the Illinois and Berkeley teams plan to offer this capability in their libraries. And Schatz and his colleagues are planning to equip their test library—a collection of thousands of articles on science and engineering—with an even more powerful interactive tool derived from Mosaic. A user who discovers an interesting relation between items will be able to set up links between them so that later users who called up one article would be referred to the others.

Ultimately, Schatz says, users should be able to add to the library themselves. A scientist might, for instance, take an equation from a journal article, use it to analyze data from another place on the Internet, use a visualization program from a third source to view the results, and then leave a record of the work for others to see and work with.

With this new suite of capabilities, the Internet would become something else altogether—a space of connected information instead of a collection of machines. Schatz calls it the "Interspace." Others may offer other names, but certainly no one will call it a used book store.

–Robert Pool

PLAGUE EPIDEMIC

Bottleneck Keeps Existing Vaccine Off the Market

As epidemics of deadly bubonic and pneumonic plague sweep India, no plague vaccine exists that can stop the epidemic in its tracks. There is a vaccine, formulated half a century ago, but it is far from ideal for halting an epidemic: It takes weeks to evoke immunity, and it does not provide permanent protection. No new vaccines have been developed in recent years, because there's little commercial incentive for companies to invest in an improved vaccine for a disease that rarely strikes industrial countries in significant numbers. As a result, promising leads on new vaccines haven't been followed up.

Even the current, flawed vaccine would be better than nothing, however—especially to protect vulnerable groups such as health or relief workers. But right now, it isn't commercially available in India—or in the United States. In India, 3-year-old stocks of the vaccine are only now being distributed by a government institute (see box on p. 23). And stocks in the United States, stored in the refrigerators of a small company in North Carolina, are not for sale. The reason: The U.S. Food and Drug Administration has taken nearly a year to confer its approval on the serum, even though the vaccine is identical to one that had been produced by another U.S. company for decades—and despite U.S. Army trials that have shown the vaccine to be safe and effective.

AP

Thereby hangs a tail. Sanitation worker in Bombay collects rats; a bounty has been put on rats in that city to avert the spread of plague.

"It's heartbreaking to say there's nothing we can do," says William White Jr., president of Greer Laboratories, the small family-owned manufacturer located in Lenoir, N.C., which bought the rights to produce the plague vaccine 2 years ago. "We have a major plague in India, vaccine in the warehouse, people calling for it, and we can't get it to them." Most of those calling his company, White says, are relief groups and U.S. companies whose personnel work in the area.

The bottleneck holding up the plague vaccine in the United States results from FDA regulations. Those regulations require that if rights to a vaccine are sold to a new producer, the new company's preparation must be tested as if it were a new product—to ensure that the new company can consistently make a product identical to the one already approved.

Those regulations became a barrier to availability of the plague vaccine in March 1992, when Greer acquired the rights to the vaccine from Cutter Laboratories Inc. of Berkeley, California, a subsidiary of Miles Inc., which is itself an offshoot of the multinational A. G. Bayer Co. of Germany. The vaccine would bring Greer less than $2 million in revenues, but that is a significant sum for a company whose revenues are $10 million a year.

With no likely competition, Greer would have a virtual monopoly on the vaccine and a lock on the vaccine's sole significant U.S. customer—the Army, which must protect its troops in the event they are sent to a region of the world where plague is endemic. Patrick Murphy, professor of medicine at the Johns Hopkins School of Medicine, says "the Armed Forces are the only people who have a use for the vaccine, because the disease is so rare there's no reason to immunize the general population."

The Army's needs are not merely theoretical: The Cutter vaccine was an essential part of medical protection for U.S. troops in Vietnam. According to Walter Brandt, a civilian research employee of the Army, U.S. troops, all of whom were immunized, suffered only one case of plague per 1 million person-years of exposure. In contrast, the South Vietnamese army, which was not vaccinated, suffered 333 cases of plague per 1 million person-years of exposure.

Because of the armed forces' need for a plague vaccine, the testing of the Greer vaccine has been a collaborative effort between Greer and the Army. To satisfy the FDA's requirement that the Greer vaccine be identical to its predecessor, White says, the company used the same 50-year-old strain of the