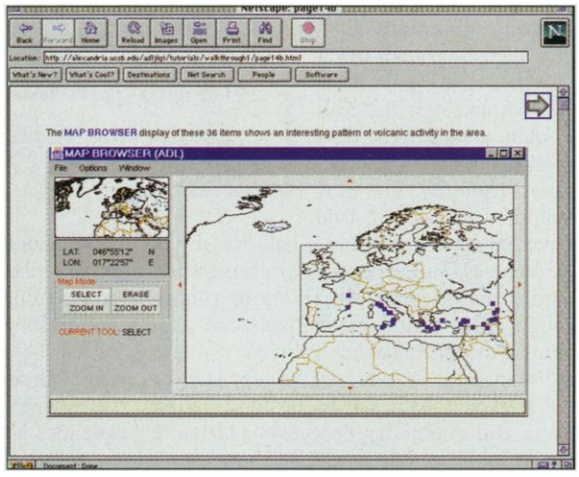
The "universal library," an amalgamation of all recorded human knowledge, searchable from your personal computer, sounds like a fantasy. But the elements are now under development

Assembling the World's Biggest Library on Your Desktop

Ask Raj Reddy or Michael Shamos what the library of the future might look like, and they tend to get carried away. Imagine, they say, sitting at your computer and having access via a lightning-fast Internet connection to the entire corpus of recorded human knowledge and creation. Want to see the inside of Saint Peter's? Some keystrokes, a few mouse clicks, and you are off on your own walking tour. A painting high on a wall captures your interest—click on it and you can find out who its creator was and where you might see other examples of that artist's work, work of similar style, or perhaps a history of Italian baroque art. Or say you have observed what you believe is an unusual interaction between a species and its environment. A quick online query provides examples of similar interactions along with maps showing where those other habitats exist, videos of those species, sound clips of their vocalizations, and digests of research germane to your particular pairing—all in Spanish, your native language.

As Reddy and Shamos, who are computer scientists at Carnegie Mellon University in Pittsburgh, see it, you will access this universal library from your desktop via a simple interface akin to today's search engines such as HotBot or Yahoo! Behind it, however, will lie multiple search tools that convert your request for information into formats and languages that can query thousands of individual digital libraries housing text, two-dimensional and three-dimensional images, music, maps, and other types of data. "We're talking about a means of bringing together materials from many libraries that have little in common, in a way that makes the process transparent to the user," says Reddy. "Are we close to having such a library today? Not even remotely so. Can we get there? Yes, I believe that over the next 20 to 50 years, the truly universal library will exist and, importantly, that it will be worth the effort and money that we will spend to make it a reality."

Reddy, who first coined the term "universal library," is working with colleagues at Carnegie Mellon on methods of indexing and searching video clips, as well as creating searchable transcripts and abstracts of video images. He is also acting as chief negotiator, getting collection holders as large as entire governments to support this endeavor. And thanks in part to Reddy's evangelism, enthusiasm for the universal library has spread beyond Carnegie Mellon to computer science laboratories worldwide. In the United States, they're building on projects funded over the past 4 years by the federal Digital Libraries Initiative (DLI-1), which developed schemes to collect, store, and organize information in digital forms and make it available over communication networks. Now, with DLI-1 winding down, the challenge is to meld those individual libraries into a seamless whole—a feat that will require breakthroughs in information processing and retrieval, in addition to teraflops and gigabucks. U.S. funding agencies are set to kick in $50 million over 5 years for DLI-2, but industry will have to spend hundreds of millions more to make the universal library a reality.



Virtual map room. A screen display from the Alexandria Digital Library shows sites of volcanic activity near the Mediterranean.

First, however, there are some tough nuts to crack. At the top of almost everyone's list is developing new, fast methods for finding information in what will be a widely distributed library. "It's amazing that we find anything using today's browsers and search engines," says Susan Dumais, a member of the natural language processing group at Microsoft Research in Redmond, Washington. Storing and analyzing images, music, and other nontext information is a second challenge. Developing abstracting programs that can rapidly summarize search results or display results in some manageable form, so that users will not be swamped with too much information, is a third.

"We have enormous problems to solve before we can efficiently search all the available text documents, let alone nontext information such as mathematical formulas or music or art, in a comprehensive and useful manner," says Paul Kantor, director of Rutgers University's Distributive Laboratory for Digital Libraries. "But I don't doubt that those problems are solvable." Adds Gloriana St. Clair, head librarian at Carnegie Mellon, "This effort now has enough momentum to make the universal digital library a reality."

Building blocks

Some of the building blocks for the universal library are already being laid down: Hundreds, probably thousands, of digital libraries are quietly amassing huge collections of nearly every type of data imaginable and making them accessible via the Web. Some, such as Project Gutenberg (sailor.gutenberg.org), resemble traditional libraries, updated for the digital world. As of May, Gutenberg had 1306 out-of-copyright classic texts available at a mouse click for downloading in ASCII format. A step up in complexity are searchable libraries such as JSTOR (www.jstor.org), which now contains current and back issues of 83 academic journals. Then there are what could be called virtual digital libraries, such as the Electric Library (www.elibrary.com), a commercial gateway through which subscribers can search hundreds of other private and public-access libraries including newspaper and magazine archives and transcripts of government hearings.

Still more complex are libraries of materi-

als that are not as easily searched as straight text files. The Alexandria Digital Library at the University of California, Santa Barbara (www.alexandria.ucsb.edu), which is getting ready to open its doors to the public, is a collection of maps searchable by location. Still to come are libraries containing digital representations of three-dimensional objects. Takeo Kanade at Carnegie Mellon, for example, has constructed a geodesic imaging dome fitted with cameras at each of the dome's 51 corners that produces a three-dimensional, manipulable image. At Stanford University, Marc Levoy uses a boom-mounted laser that travels around an object to create images with millimeter resolution, fine enough to see an object's imperfections. The Vatican will use this system to image its enormous collection of art in Rome. And the National Ethnological Museum in Osaka, Japan, uses five video recorders and a laser dimension-measuring system to digitize every object in the museum's collection for eventual online display.

But this proliferation of libraries is far from the ideal that Reddy and others espouse of a universally accessible information source. "To begin with, we don't even know of all the digital libraries that exist and what's in them, so just as a start we need a worldwide registry of all the libraries and their content. This isn't that hard to do, we just have to set up the right mechanisms," says Shamos, operations director of the Carnegie Mellon project. GenBank (www.ncbi.nlm.nih.gov), a library of human genome sequences administered by the National Institutes of Health, might be one model. Journals require investigators to register sequence data with GenBank before publishing papers on the work. Similarly, digital libraries might have to register their presence and content before being granted a uniform resource locator (URL) Web address.

Unearthing buried text

A far bigger challenge is devising ways to extract information from all these libraries. "We're getting very good at building data tombs, huge repositories in which information becomes buried forever. What we need now are ways of getting that information out of these repositories," says Usama Fayyad, director of data-mining studies at Microsoft Research.

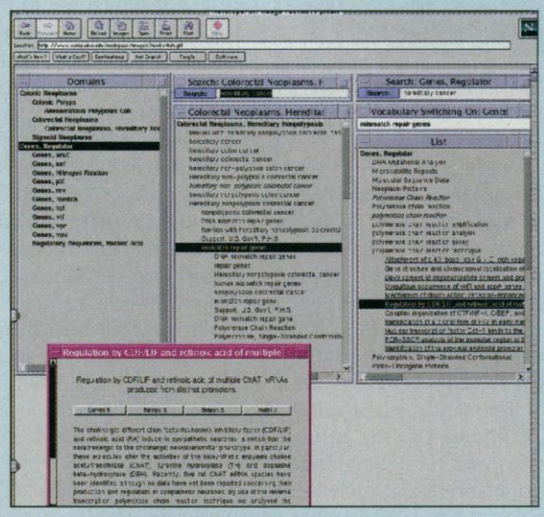
Taming MEDLINE With Concept Spaces

At 10 million entries and growing, the National Library of Medicine's MEDLINE is one of the world's largest public databases and a well-used resource for the biomedical research community. But this online collection of papers and abstracts is also unwieldy. "If you work hard at it and you have a lot of time, you can usually, but not always, find the information that you are looking for," says Richard Berlin, a surgeon and medical director for Health Alliance, a regional health maintenance organization in Champaign, Illinois. "If you need an answer right away, forget it."

MEDLINE's problem is that in spite of its size and the many disciplinary boundaries that its entries cross, it is constructed along much the same lines as most other databases. Human indexers assign an average of 14 key words, known as Medical Subject Headings, or MeSH terms, to each bibliographic entry, of which an average of four so-called central MeSH terms provide the most specific categorization. Users query the database and retrieve all abstracts whose MeSH terms match those in the query. Where MEDLINE fails, though, is in searching for information that crosses even minor disciplinary boundaries—developmental genetics research on two different species, for example. "If you are a biologist working on *Drosophila* genes and you are trying to find out what's been done in the *Caenorhabditis elegans* literature, MEDLINE is not going to give you many good answers because the terminology that the two disciplines use, and the MeSH terms, are [different]," says information scientist Bruce Schatz of the University of Illinois, Urbana. Now, with the help of one of the largest computations ever, Schatz has come up with a scheme for bridging these vocabulary differences.

Two years ago, as part of the Digital Libraries Initiative, Schatz and colleague Hsinchun Chen of the University of Arizona, Tucson, used semantic indexing techniques to classify more effectively a collection of 1 million abstracts in the engineering literature (*Science*, 7 June 1996, p. 1419). These techniques involved using key words to partition the large database into smaller collections that correspond to community repositories for specialized disciplines. Each smaller collection presumably shares a common set of jargon. Computations using statistical methods create sets of related entries, called concept spaces, listing all the terms that occur with one another within the collections. Terms that appear together in one concept space—"highway" and "pavement," for example—are likely to be linked to the same underlying concept. The result is a thesaurus of alternative keywords for searching the entire database. Building this semantic index was a computationally intensive task, one that required 10 days of processing time on the world's fastest supercomputers at the National Center for Supercomputing Applications (NCSA) on the Illinois campus.

Now, Schatz has repeated this feat on the MEDLINE database, creating the first semantic index for a complete scientific discipline. The 10 million MEDLINE entries were partitioned into 10,000 community repositories, by using the central MeSH terms classifying each entry. Each repository thus represents a specialized scientific subdiscipline, such as colon cancer or *C. elegans* genetics. The current largest NCSA supercomputer, the 128-node Silicon Graphics Origin 2000, then produced the semantic indexes, taking roughly 8 days for testing and 2 days to process the entire database. The resulting concept spaces are then searchable as part of the Illinois Digital Library test-bed.

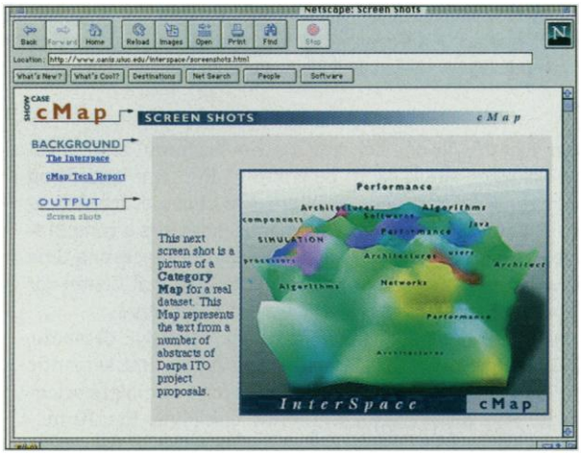


Searching by concept. A prototype "semantic index" of MEDLINE suggests alternative search terms.

Early demonstrations (www.canis.uiuc.edu) have gone over well with physicians who have had access. Querying is still done using word matching, but the searchers now rely on the semantic index rather than on MeSH terms. "It's wonderful," says Jonathan Silverstein, a surgeon and member of the informatics faculty at the University of Illinois in Chicago. "I'm now getting far more useful information out of MEDLINE, and I'm getting it in a time frame that I was actually able to get information I needed while the patient was in my office."

For Schatz, this success is merely a precursor to what he hopes will be a semantic index of the entire Net. "I see this as a model for the day 10 years from now when we have a billion small community libraries distributed around the world and we will need a fast, searchable index of that body of knowledge," says Schatz. **–J.A.**

Today's commercial search engines do a mediocre job at best at finding text-based items that satisfy a particular query, and they are not designed to find image- or sound-based information. The problem, say researchers, is that the information on the Web is so heterogeneous in structure as to be almost unsearchable using any one approach. "Key-word matching, the heart of your basic text search engine, is fine if you are searching a small database," says Bruce Schatz, director of the digital library project at the University of Illinois, Urbana. "But that only works if all your documents are in one language, if everyone consistently uses the same words in the same way, and if the person conducting the search already knows all the terms germane to a particular topic."



Topography of information. Search results are displayed on a landscape, where broad concepts sit on hilltops.

Steve Cousins, a search expert at Xerox's Palo Alto Research Center in California, took on the challenge while he was a graduate student at Stanford. His system, the Digital Library Integrated Task Environment, provides a simple user interface hiding a translation system that can reformulate search requests into many different forms capable of interacting with different data structures at a contextual level rather than simply matching words. Now he is developing what he calls wrappers, programs that would sit over a digital library and translate its particular data structure into a form understandable by the particular search engine querying that library.

Another approach to gleaning useful results from many different kinds of libraries is to analyze context as well as search for key words. At Microsoft Research, Susan Dumais's language processing group is using statistical models to analyze how often words occur together in documents as a means of drawing inferences about what a document says, rather than just the words it contains. Schatz's team has taken a similar approach to create searchable concepts within the 10 million items in the National Library of Medicine's MEDLINE database of biomedical literature (see sidebar on p. 1785).

Making search engines more discriminating won't solve another problem that plagues Web searches today: the overwhelming number of potentially useful Web pages they often return, with little indication of the content of the pages or how they may be related. To ease this problem, Marti Hearst of the University of California, Berkeley, has developed Cha-Cha, a program that determines the home page for each item retrieved, records the shortest path to get from that home page to the retrieved page, and then groups together information that shares pathways. The result is an organizational map of the relationship between different pieces of information. Although this still requires manually browsing one page in each group of pages, it allows the user to narrow a search quickly to a smaller subset of information. "This seems simple, but it turned out to be a surprisingly effective way to organize what was otherwise perceived as a disconnected jumble of pages," she explains. Cha-Cha is now being tested on the 300,000 Web pages in Berkeley's system (cha-cha.berkeley.edu).

In another effort to make sense of search results, a software development team at the University of Illinois, led by Schatz, Bill Pottenger, and Kevin Powell, has created cMap, which displays search results on a colored three-dimensional surface. Broad concepts sit on the tops of "hills" that the user can easily spot; diving down the face of the hills leads to more focused topics. Neighboring hills contain related concepts to which the user can easily jump. "It's a way of visually organizing data that reveals more about the relationships between different sets of information," says Schatz.

Others are working on systems that rely on an iterative, give-and-take approach with the user. Stanford's SenseMaker, designed by Michelle Q. Wang Baldonado, now at Xerox, and her Ph.D. adviser Terry Winograd, allows the user to explore a subset of an initial search result, observes the user's reactions to assess his or her preferences, and refines the search criteria to suit them.

Getting the picture

Searching images, including maps, artwork, photographs, and video, is a more challenging task. "For all the problems associated with searching text, it is still an inherently searchable medium," says Shamos. "But with images, we don't have that built-in verbal content handle to grab onto." The Getty Information Institute in Los Angeles is leading an effort to add that verbal content by developing standard reference terms for cataloguing digital images.

Others are working on methods to search images based on characteristics of an image itself. Shi-Fu Chang, an expert in content-based image processing at Columbia University, and his colleagues have developed VisualSEEk (www.ctr.columbia.edu/VisualSEEk), a collection of three search engines that query online art museum collections based on subject matter, physical properties, or similarity to a random sample of object types. For example, the SaFe search engine (disney.ctr.columbia.edu/safe/) looks for objects that match color, texture, and composition elements specified by the user.

Ultimately, the universal search engine will probably comprise several task-specific information seekers—one for images, another for video, a third for music, perhaps several for text—that will find all relevant knowledge to satisfy a particular query. But such a search engine will need multiple user interfaces. "There are many different cognitive search styles among users," says Carnegie Mellon's St. Clair. One user, for example, might feel frustrated having to weed out extraneous information from a search, while others may want information peripheral to their main question.

Besides these research problems, a number of practical issues are looming, most of them as yet with no solution at hand. Foremost is language—the Web is very English-centric today, but a universal library implies access for everyone, including those who speak Chinese, Swahili, and Urdu. This will require some type of translation system that ideally would provide literal translation on the fly and, at a minimum, translate concept and content information.

Copyright and reimbursement issues could also restrict the availability of collections through a universal library. And then there is the problem of bandwidth—moving all this data around the world in a reasonable amount of time. The Alexandria Map Library, for example, ends up shipping its digitized maps to users on tapes via Federal Express because sending a gigabyte file via the Internet is not practical today.

But Stephen M. Griffin, who heads the National Science Foundation's digital library efforts, is not worried. "This infrastructure issue is going to take care of itself because the commercial demand is there. The scientific and social issues that we have yet to solve are the real bottlenecks in creating a universal library." **–JOSEPH ALPER**

Joseph Alper is a writer in Louisville, CO.