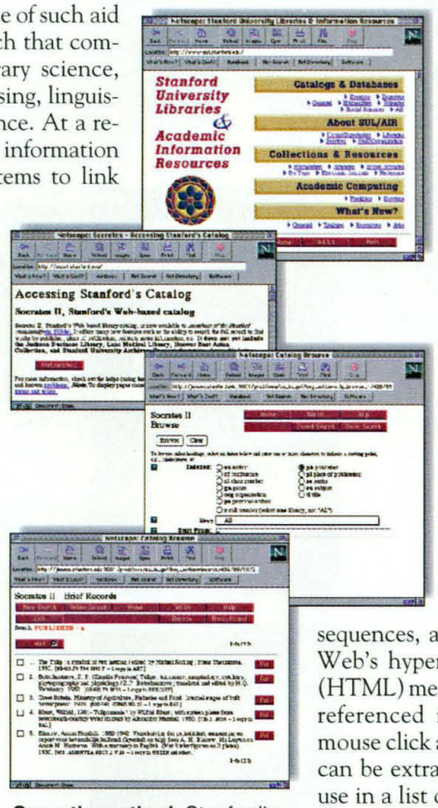
The Search for Mr. Goodfile Generates New Online Tools

HATOYAMA, JAPAN—Swimming alone in an ocean of data may be fine for cybersharks. But most of us would appreciate help as we splash about in the seemingly bottomless pool of information that's available online.

One promising source of such aid is a new wave of research that combines the fields of library science, natural language processing, linguistics, and computer science. At a recent conference here,* information scientists described systems to link the physical resources of traditional libraries to the online digital world. They debated whether traditional indexing and cataloging efforts could be adapted to the World Wide Web. And they presented glimpses of ways to simplify information retrieval.

These efforts are forcing major changes in the world of scholarly information. Stanford University is making the World Wide Web the primary gateway into the library's catalogs and indices of its holdings, notes Michael Keller, director of Stanford University Libraries and Academic Information Resources, because of the user-friendly quality of Web browsers, including their hyperlinks and global accessibility. Keller says the idea is to equip the library's physical resources with the same searching and linking capabilities that are available online.

In addition to the challenge of digitizing and linking catalog information for 7 million books, plus recordings, rare manuscripts, and art, Stanford (which is responsible for the development and maintenance of the *Science* Online Web site) is also developing search and retrieval engines to manipulate and present the information. Starting with dozens of search engines and more than 100 indices in the various Stanford libraries, Keller would like to provide one-stop shopping for users. In addition to returning the requested information, the system would allow the user to review the resources tapped and get information on how deeply they were searched. "These are not trivial goals," Keller says.



Socratic method. Stanford's Socrates II broadens the scope of online searches.

An example of what can be done is Stanford's Web-based catalog interface, Socrates II, which goes beyond simple searches by title or subject to include sorting by publisher, place of publication, and language of the holding. Retrieved records also can be displayed in various sequences, and card records in the Web's hypertext markup language (HTML) mean that additional cross-referenced information is only a mouse click away. Bibliographic data can be extracted and formatted for use in a list of references.

Socrates II, now being tested, has been developed by some of the same people involved in the Digital Libraries Project, a 4-year program begun in 1994 and funded by three federal agencies (*Science*, 7 October 1994, p. 20). Six university-led consortia are tackling the challenges of collecting, storing, and organizing digital information and making it easily accessible. While most have a particular focus, typically by discipline or type of material, Stanford's task is to develop a single interface that would shift the burden of navigating databases with different structures and various search engines from the user to the software. The tools would apply both within the Stanford system and on the Web.

Although a library setting provides abundant challenges, Keller says that the real test is taming the Internet, which lacks the theoretical and taxonomic structures that allow controlled, systematic retrieval of information. Whether such structures could be created for online resources was a matter of some debate. Keizo Oyama, an information scientist at Japan's National Center for Science Information Systems, proposed standard sets of key words that authors would attach to documents. Internet service providers could then have indexes of material on their systems. Better software tools to automatically generate the key words and indexes would help make all this voluntary work easy, he says.

However, relying on authors or service providers poses a serious problem, says Karen Sparck-Jones, a linguist at the University of Cambridge. Authors are not necessarily the best judges of how to index or catalog their own works, she says, because of their limited knowledge of how the documents might be accessed and used. In addition, she notes, few people who put information on the Internet are interested in cataloging or indexing the material. A better approach in this era of full-text retrieval, she suggests, is improved search engines and techniques.

Better retrieval methods is one of the promises held out by developments in natural language processing. NLP, which grew out of attempts in the 1950s to automate library cataloging and indexing, uses computational methods to try to discern meaning from a text. It may count the number of times words appear, or look at the relationships among phrases in a sentence. The goal is for computers to recognize meaning in human language and to transmit messages for humans to read or hear.

Researchers had hoped that NLP would improve information retrieval through an increased understanding of queries and better filtering of documents. For a question about legal suits, for example, NLP would realize that the user was interested in litigation and lawsuits and reject documents related to clothing. But progress has been slow. Sparck-Jones says NLP techniques have not proven significantly better than search techniques based on multiple simple terms. "The generally good information-retrieval strategy is just to use more single terms in the query," she says.

NLP enthusiasts hope that the increased availability of uncataloged, full-text documents will raise demand for solutions, and the workshop offered examples of such advances in NLP techniques as term recognition and parsing of sentences. But commercial search engines now make only rudimentary use of NLP techniques. Sparck-Jones believes that NLP's greatest contribution may come in summarizing documents once they are retrieved.

While the payoff from natural language processing may be far off, work continues on refining the dominant method of searching the Web through an iterative cycle involving request-review-modification. Yoshiki

* Workshop on New Challenges in Information Retrieval and Dissemination, 7–8 April, Advanced Research Laboratory, Hitachi Ltd., Hatoyama, Saitama, Japan.

Niwa, of Hitachi Ltd.'s Advanced Research Laboratory here, showed an interactive search scheme, called Dual-Navi, that presents search results both concretely and abstractly in side-by-side windows.

Starting with a search string, Dual-Navi presents the typical list of retrieved titles on the left side of the screen while a graph of key words extracted from those retrieved documents appears on the right side. The key words are displayed according to their frequency of occurrence, and associated words are joined by solid lines.

Although the popular AltaVista search engine uses a similar graph, Dual-Navi provides interactive links between the two views. To narrow the search, users select additional key words in the graph view and click a button. The documents containing those words will come to the top of the title list. Conversely, click on a title, and the key words found in that document are highlighted in the graph. Additional titles with similar characteristics can then be gathered. These processes can be repeated, with the graph and list views constantly changing to reflect the latest stage of the search.

Even more user-friendly, however, would be a system tailored to a person's needs. Stanley Peters, a mathematical linguist at Stanford's Center for the Study of Language and Information, presented one approach based on concepts, or groups of synonymous words, extracted from a person's e-mail. The idea, says Peters, is to exploit the "idiosyncratic associations" among words to come up with customized searches.

In one test, researchers generated associations based on 3 months of an individual's e-mail and, for comparison, a database of 42,000 Associated Press (AP) news wire articles. They then searched a target database using four key words—race, identity, Asian, and dating. The results were strikingly different. The documents retrieved using the associations generated from the AP articles were primarily about black-white race relations, while those retrieved using the associations gleaned from the individual's e-mail were much more closely related to issues involving Asian race relations, the individual's primary research interest.

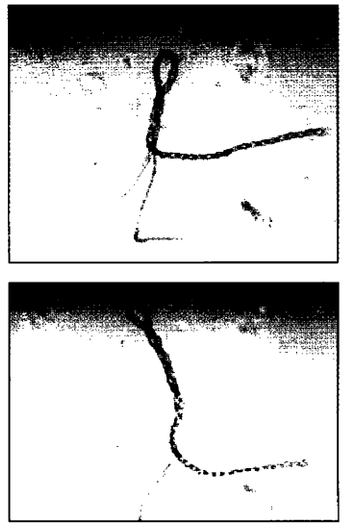
Peters believes this approach could be extended. Civil engineers and stamp collectors, for example, could use sets of associations generated from databases of civil engineering journals or philatelic magazines to narrow the range of retrieved documents when searching something like the Web. But even this feature has its limitations. "There is not likely to be one approach that suits all particular needs," Peters says. So, while trolling through the ocean of information may get easier, it is still going to take work to stay afloat.

–Dennis Normile

MICROBIOLOGY

Physics, Biology Meet in Self-Assembling Bacterial Fibers

Twenty years ago, when Neil Mendelson first described a mutant strain of bacteria that twisted itself into ropy helical fibers, his fellow microbiologists considered it just a curiosity, one of many in the world of microbes. As Mendelson, a professor at the University of Arizona, narrowed his research to focus on the quirky twists and turns of these bacteria, the scores awarded to his grant applications took a nose dive and so did the number of papers he published in peer-reviewed microbiology journals.



Self-starters. Bacterial filaments *(top)* twist spontaneously into helical fibers *(above)*.

Lately, however, Mendelson's odd microbes have been making a name for themselves in some unexpected settings, far from microbiology. They have won devotees among mathematicians, engineers, and physicists who, collaborating with Mendelson, have used the microbial fibers to help solve long-standing problems in elasticity theory, model solar flares, and make a new siliceous material that could be used in medical implants. Mendelson was "a pioneer and out of the mainstream," says Ralph Slepecky, a professor emeritus of microbiology at Syracuse University in New York. "But his work is proving useful."

What has sparked all this interdisciplinary effort is a mutant form of a common, rod-shaped bacterium called *Bacillus subtilis*, about 4 micrometers long and 0.7 micrometer in diameter. Back in 1975, Mendelson discovered a strain lacking the enzymes that normally cleave daughter cells after cell division, so that the daughters grow stuck to their parent cells like beads. Individual filaments of linked bacteria spontaneously twist and double back on themselves many times to form a thick, rope-like helical coil of up to 100 filaments (*Science*, 3 January 1992, p. 32). Mendelson dubbed these coils "macrofibers," and while their unique penchant for self-assembly may have left some microbiologists cold, it piqued the interest of physical scientists.

For example, when Michael Tabor, head of applied mathematics at the University of Arizona, first saw a film of macrofibers self-assembling in 1992, he realized that he had found a living, dynamic model for flexible elastic filaments. Mathematicians have been modeling the way these filaments twist for more than a century, but most models were static—describing only starting and ending structures, rather than the complex stages in between, says Tabor. His group, including post-doc Alain Goriely, spent several years observing *B. subtilis*—live and on video—and developed new dynamic equations to describe the twisting and coiling of macrofibers. Their analysis describes a seemingly unpredictable aspect of macrofiber coiling: how two-dimensional twisting causes the fibers to kink or rise up from a flat surface, adding a third dimension to their shape. For many natural systems, the spontaneous twisting of the bacterium is a better model than, say, a rubber band, which twists only in response to an outside force, says Tabor.

Mathematicians call this move into a third dimension "writhe." Tabor's model shows that writhe stems from subtle mathematical instabilities: When the researchers altered certain variables in their equations, the solutions required a shift into three dimensions. These mathematical instabilities are likely to be a general property of elastic filaments, Tabor says, although he doesn't know to what physical properties they might correspond. Because elasticity theory is used to model everything from the supercoiling of DNA to the twisting of magnetic field lines in a star, the new model will likely have plenty of applications, adds Mendelson's collaborator John Thwaites, a mechanical engineer at Cambridge University.

In fact, Tabor's model of elastic filaments has already been applied to the behavior of so-called magnetic flux tubes in the sun. These structures, made up of bundles of magnetic field lines, cause sunspots and can trigger the enormous magnetic detonations on the surface of the sun known as solar flares. The tubes emerge from the sun's interior as narrow strands of magnetic field, which float to the surface and appear as sunspots. The long tails trailing back into the interior can be mod-