Indexing the Internet

The World Wide Web makes it simple to retrieve information from cyberspace. Now researchers are devising systems for tracking down the information in the first place

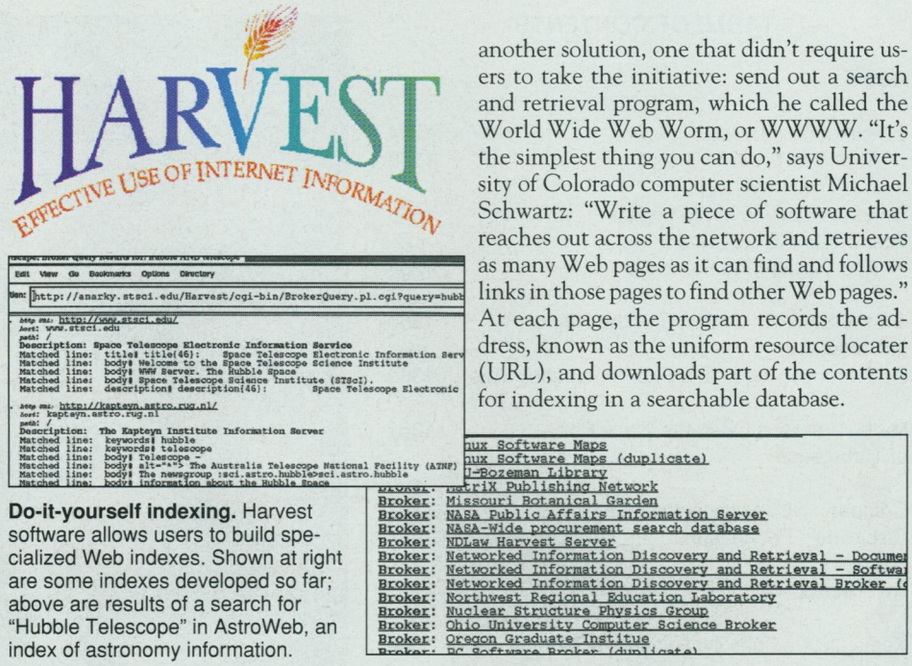
Two years ago, the riches of cyberspace started to pale on David Filo and Jerry Yang, electrical engineering students at Stanford University. To be sure, says Filo, they were finding "plenty of cool sites and stuff" on the World Wide Web, the web of linked Internet sites that invites users to wander from one to the next with the click of a mouse. But whenever they went looking for a specific topic, they soon got lost. "So we started this thing," Filo says, "that allowed us to quickly categorize sites as we came across them, and we put it out on the [World Wide Web]."

The result was called, for lack of a better name, "David and Jerry's Guide to the Web." As the service grew, however, Filo and Wang opted for an acronym. Filo says they considered "yet another something," and so they looked up words starting with "ya" and found Yahoo, which seemed to fit. "So we ended up calling it 'Yet Another Hierarchical Officious Oracle.' " Yahoo today indexes 60,000 Web sites that Filo and Yang consider noteworthy, sorted into 10,000 categories. That's just a fraction of the Web's tens of millions of documents and resources. Even so, Yahoo is considered one of the more complete guides to the Web.

And that sets the scale of the next challenge confronting the architects of the Internet: indexing its entire contents, so that a user seeking a specific piece of information—from box scores in yesterday's *Beijing Times* to the first treatise on rural electrical lighting—can quickly hunt it down. As Robert Wilensky, head of a digital library project at the University of California, Berkeley, puts it, the goal is to "invert the Internet," opening the way for users to "browse by content, navigate through concept space, rather than having to find things based on where someone put them up on their home page."

By that standard Yahoo and its closest competitor, an indexing system called Lycos, developed at Carnegie Mellon University, are just the beginning. Already on the horizon are systems for gathering tens of thousands of documents a day and indexing them with the aid of programs that can classify their contents precisely. Other approaches would rely on specific user communities to compile their own indexes, which could then be linked to form a global guide to the Web.

In trying to devise such indexes, Internet architects are coping with one consequence of their own success. The point-and-click ease of access brought by Web browsers like Mosaic has sparked a boom in usage and in the volume of available material—and made tracking down anything specific far more difficult. "The amazing thing about the Web," says Bruce Schatz, a research scientist at the National Center for Supercomputing Applications, where Mosaic was developed, "is why it's so popular if it doesn't do anything useful. All it does is let you surf. It certainly doesn't let you solve problems, and it doesn't search out information for you."



Do-it-yourself indexing. Harvest software allows users to build specialized Web indexes. Shown at right are some indexes developed so far; above are results of a search for "Hubble Telescope" in AstroWeb, an index of astronomy information.

SOURCE: MICHAEL SCHWARTZ/UNIVERSITY OF COLORADO

The spider stratagem

One of the first attempts to solve the problem relied on anyone who posted a new home page to list the service in a central index—the "Mother of All Bulletin Boards," as its developers at the University of Colorado, led by Oliver MacBryan, called it. Users of the bulletin board could then search it for whatever subjects interested them. For the Mother of All Bulletin Boards to work, however, everyone posting documents on the Web had to know about it—and use it correctly. And even then the result would be a list rather than a subject index. "It was a good notion in principle," says Paul Ginsparg, a Los Alamos National Laboratory physicist who created an electronic preprint archive. "But it didn't really solve the problem."

MacBryan was also among the first to try another solution, one that didn't require users to take the initiative: send out a search and retrieval program, which he called the World Wide Web Worm, or WWWW. "It's the simplest thing you can do," says University of Colorado computer scientist Michael Schwartz: "Write a piece of software that reaches out across the network and retrieves as many Web pages as it can find and follows links in those pages to find other Web pages." At each page, the program records the address, known as the uniform resource locater (URL), and downloads part of the contents for indexing in a searchable database.

Since then, these search-and-retrieval programs, now called Web crawlers or spiders, have proliferated (to the distress of some people running Web sites—see box on next page). When Filo and Yang found Yahoo's popularity growing, for instance, they quickly added a Web crawler to their service. And at Carnegie Mellon, Lycos was born as a simple spider program created by John Leavitt, to which his colleague Michael Mauldin added an indexing program in the spring of 1994. "When it would fetch a document," says Leavitt, "it would create an outline or table of contents of the document by stripping out all the headers, and then it took the first 20% or 20 lines, whichever was smaller, as an excerpt or abstract. It also took a group of 100 words, which were statistically the most salient for the document, as key words" for indexing it. Because Mauldin ran the program on his workstation at night, he and Leavitt named it after the Lycosa family of wolf spiders, which are night hunters.

Lycos, which Leavitt and Mauldin have since transferred to a private company and licensed to Microsoft, now consists of a flock of spiders that have already indexed over a million documents and are adding new ones at the rate of 20,000 each day. In addition, it has gathered partial information about another 4 million Web documents referenced by the ones it indexes directly. Leavitt says

Lycos now gets several million hits a week from users, who can search the database for key words and then, with a mouse click, go directly to the relevant documents.

Lycos's spiders may crawl too slowly, however, to keep up with a system unleashed this month by Berkeley's Eric Brewer. Brewer's system, which runs on four workstations networked together to form a parallel supercomputer, controls crawlers that can bring in and index over 100,000 documents a day. At the same time, it can serve several million search requests a day.

But the caches of documents amassed and indexed by Web crawlers aren't a full answer to the challenge of making the Internet searchable, says Schatz. The problem, he says, is that such automatic indexing "is so haphazard." For example, says Berkeley's Wilensky, "the word 'film' is ambiguous." If a human indexer comes across the word "film" in a document, he or she knows whether it refers to a movie, photographic film, or dirt residue, and will catalog the document accordingly. A computer indexing program has no such intuition, so it simply tags "film" as a key word and leaves it at that; as a result, searching for "film" in the current automated indexes will bring up documents on all possible meanings of the word.

Wilensky's group is trying to solve this problem through a technique known as lexical disambiguation. The algorithm builds a reference database from a statistical analysis of the contexts in which a word is found in a wide range of documents. It can then compare a new keyword and its context with those in the database to choose the most likely meaning. "You learn, for instance, that the movie meaning of 'film' tends to occur a lot in contexts where words like 'actor' appear," says Wilensky. " 'Soiled covering' tends to appear in contexts where you hear about 'dirt' a lot. When you have those associations, you can make pretty good guesses about the right sense of the word." That classification procedure could eventually allow crawlers to build huge subject indexes that would be searchable with a precision and efficiency beyond anything available today.

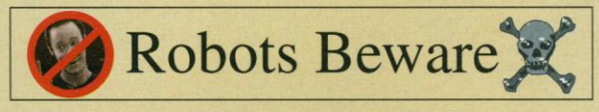
The human factor

This sensitivity to context demands large amounts of computing time, and even then, the approach isn't likely to be capable of subtle cataloging judgments anytime soon. As a result, some computer scientists are looking for ways to harness human expertise to build high-quality Web indexes. Schatz, for instance, predicts the creation of community repositories maintained by particular disciplines—groups ranging from martial-arts enthusiasts to computer scientists. "Once you have that kind of community," explains Colorado's Schwartz, "you can have somebody who knows the subject, who is

The Web-Crawler Wars

Efforts to organize the millions of documents on the World Wide Web have set the Web aswarm with programs known as crawlers, spiders, or sometimes robots (see main text), which visit remote sites and automatically download their contents for indexing. Robots are a clever solution to a tedious task. But Paul Ginsparg, a physicist at the Los Alamos National Laboratory, regards them as "mindbogglingly stupid."

Ginsparg has reasons for his vitriol. He runs an electronic archive where some 30,000 physicists and other researchers post electronic preprints and read the latest work of their colleagues. It's the Web home to over 100,000 documents, which is why Ginsparg has found himself waging an ongoing altercation with Web crawlers that methodically download his entire archives, clogging his server and inconveniencing more sentient users.



No androids, please. Preprint archive warns off Web indexers.

SOURCE: PAUL GINSPARG/LANL

Ginsparg woke up to this unexpected consequence of Web indexing efforts in February 1994, when he discovered that a computer from England had been requesting thousands of documents from his archives, one after another. "I didn't know what was going on," he says, "and I had no idea how to get in touch with the person. I tried fingering the machine to see who was logged on. Nobody was logged on. I tried sending e-mail to the machine. No answer." A week later, by serendipity, he learned that the machine was the home of an aspiring Web indexer, and he was being hit by a Web crawler.

This Web crawler seemed to have been designed under the misconception that all Web sites are small, says Ginsparg, and thus it could retrieve and index each site's entire contents without tying up its server for long. But Ginsparg's archive includes virtually every paper written in physics in the past 4 years, perhaps a few billion bytes of data. The Web crawler's effort to download it all to a machine in England wasn't just a minor annoyance, he says. "It's not only slowing down my system, but it's essentially hurting everybody trying to use the trans-Atlantic links."

To get help and alert others, Ginsparg posted a message about the incident to an Internet news group, only to find that he wasn't the only one worried about Web crawlers. In particular, he heard from Martijn Koster, a computer scientist with the British information systems company Nexor, who said that he had been tracking robots since they first appeared and had drawn up common-sense guidelines for robot builders. As Ginsparg describes them: "Announce that you're running this thing; when generating the requests, say who you are and how to contact you; don't run it unattended; have a limit on the number of requests to any single site; don't keep sequentially requesting from the same site; in no case make multiple requests without time outs, etc."

People making serious attempts to index the Web seemed to have gotten the message. John Leavitt, one of the creators of the commercial Web-indexing robot Lycos, says his group went out of its way to make its crawlers "Internet-friendly." Eric Brewer, a computer scientist at the University of California, Berkeley, who has recently created the world's fastest crawler, says, "We actually do more than the protocols ask for, because our search rates are so high. We go to great lengths to randomize and distribute targets of the crawlers, so even though we can bring in 100,000 documents per day, we'll only hit a single site a small number of times. And so far, no one seems to have even noticed us."

Ginsparg, meanwhile, has written his own counterattack robot. "It just cuts them off," he explains, "and automatically looks up the administrative and technical contacts for their domain and sends a canned irate e-mail message to them as well as to the offending machine."

For humans viewers, he also added a "Robots Beware" notice to the front page of his archive. Click on it and you'll find a page headed by a picture of the android, Data, from *Star Trek*, within the circle-and-slash symbol for "prohibited." Next to Data is a skull and crossbones. The text explains that the archives are huge and should not be downloaded. And at the bottom, says Ginsparg, "It says 'Click here to initiate automated seek-and-destroy against your site.' You wouldn't believe how many people click on that."

–G.T.

responsible for making decisions on content." And in turn, a properly linked array of community indexes could serve as a loose-knit index covering all subjects.

For information scientists, the challenge is developing software that allows each community to choose and retrieve documents for a do-it-yourself index. Schwartz's Harvest Program at the University of Colorado is considered the most ambitious attempt so far. The Harvest software, he says, makes it "pretty easy" for anyone "to list a set of uniform resource locaters and documents they want to have included in the index." The software then retrieves the data, extracts the content for indexing purposes, builds the index, and handles queries.

To make the program even more efficient, Schwartz and his colleagues have designed Harvest to be split into two distinct parts, a gatherer and a broker. The gatherer resides at a remote site, retrieving documents to be indexed, while the broker stays on the home machine and builds the index. "Imagine a scenario," says Schwartz, "where you're trying to index a lot of information at NASA, for instance. You can start distributing the process by putting gatherers on all the machines where the data are. Each gatherer can then extract data much more efficiently, because it doesn't have to go across the network to do it. ... Instead of a gatherer sitting on one machine reaching out, you can have gatherers on 100 different machines, each one boiling down the information locally and then sending it across [to a single broker]."

The project was started in late 1993, and Schwartz estimates that Harvest users scattered throughout the Web have put together almost 1000 independent indexes. To link them, Schwartz and his colleagues plan to create software that will allow each distinct index to have pointers leading to other Harvest indexes. And the Harvest researchers are also working on indexes that will support not just documents but actual scientific data, linked to programs that can interpret the data to produce meaningful information. "Right now the Internet is mostly used for data that humans look at," says Schwartz, "Eventually you want to build systems in which programs actually go along, collect data together, and do computations on it." Such indexes will do more than sort through what's on the Internet; they will make sense of it.

But in spite of the best efforts of Schwartz and other indexers, says Robert Kahn, one of the original architects of the Internet, who is now with the Corporation for National Research Initiatives in Reston, Virginia, there won't be any single best way to index the Internet. The solutions, he suggests, will be "technology dependent and sociology dependent," balancing such issues as whether the desired information is distributed throughout the Internet or is more concentrated and whether users care most about reliability, completeness, or speed.

"It's hard to generalize," he says. "It's like discussing transportation systems. You can ask me what I think will be the best transportation system in the future, and I'll say it depends on what you want to do and where you want to go."

–Gary Taubes

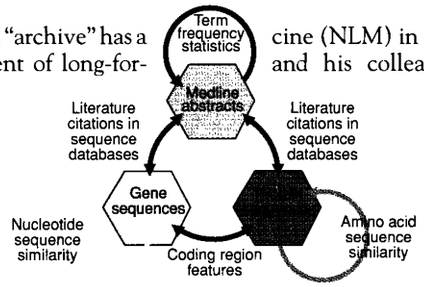
BIOINFORMATICS

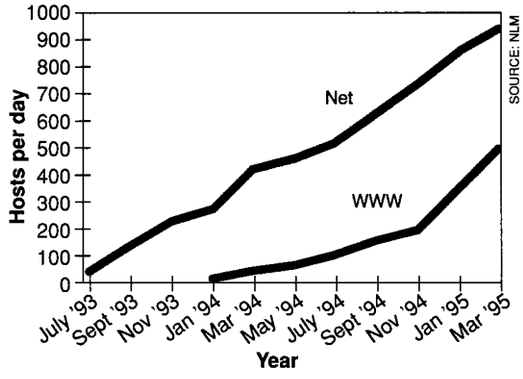
On-Line Archives Let Biologists Interrogate the Genome

To many of us, the word "archive" has a dry, dusty sound, redolent of long-forgotten library shelves. Not even computers necessarily change that perception; they just make electronic archives easier to search. To molecular biologists, however, on-line archives are anything but dry and dusty these days.

As researchers confront a fast-growing mountain of gene sequence data—it now stands at well over 200 million base pairs from humans and more than 8000 other species—electronic databases are fast becoming the lifeblood of the field. It's no coincidence that the major biomedical archives have found themselves on the cutting edge of database management. Witness the dual saga of GenBank and the Genome Sequence Data Base (GSDB), the former a nucleotide database that has been a fixture of molecular biology for more than a decade, and the latter its upstart rival.

At GenBank, "there's been a real turnaround in the past few years," says Mark Boguski, senior investigator at GenBank's current home, the National Library of Medicine (NLM) in Bethesda, Maryland. He and his colleagues have transformed GenBank from a simple archive into an intricately cross-linked array of databases where investigators can now search for similarities among gene and protein sequences, trace their evolution, and jump from sequence data to the relevant literature. By providing all these resources in a centralized archive, Boguski boasts, "our staff is playing a catalytic, proactive role in moving research ahead."





Making connections. The NLM's linked databases *(top)* draw increasing use, much of it over the Web.

Much the same feeling of excitement prevails in Santa Fe, New Mexico, where a new not-for-profit corporation called the National Center for Genome Resources is being funded by the Department of Energy (DOE) to develop the GSDB: an experimental, highly ambitious archive that originated as a spin-off of GenBank but takes a very different approach. Its goal, says the company's chief scientist, Christopher Fields, is to function not as a centralized archive but as a decentralized database, built up and maintained by the efforts of many people in many laboratories—in much the same way that science itself is. "Our task is to explore methods for building federations of distributed and autonomous databases," he says, the aim being to create an arena where scientists can pursue more complex and unexpected questions than they can at GenBank.

GenBank and its spin-off are hardly alone in the biological database world. Two others also try to maintain full archives of nucleotide sequences: the European Molecular Biology Laboratory's nucleotide sequence database (*Science*, 4 August, p. 630) and the DNA Data Bank of Japan. (The four large databases exchange newly submitted sequences every day so that each is complete.) And then there are many smaller scale, special-purpose databases, such as the archive of gene fragments at The Institute for Genomic Research (TIGR), a private research foundation in Gaithersburg, Maryland (*Science*, 16 December 1994, p. 1800). But GenBank and GSDB typify two leading approaches to coping with molecular biology's data flood.