Learning to Drink from a Fire Hose

Instant access to far-flung databases could soon be a reality, but how will we swallow a trillion bytes per day?

SCIENCE IN THE '90s

Last in a series

"Say I'm interested in a particular galaxy or star," says astrophysicist Larry Smarr, director of the National Center for Supercomputer Applications at the University of Illinois. "Ideally, I'd like to tell the computer, 'Bring me all the data from any telescope and satellite that has ever looked at it. And then combine it into something useful.' " At a keystroke the information would just *be* there, he says, brought from anywhere in the world, filtered, transformed, integrated, imaged in vivid color, and ready for the user to explore—and all without the user knowing or caring where the data originally were.

Call this system the Digital Data Library. Neither Smarr nor anyone else actually has one—yet. But from all reports it's coming fast, conceivably as early as mid-decade. The technological pieces are already falling into place, especially with the spread of nationwide computer networks operating at higher and higher speeds. At least half a dozen laboratories, Smarr's supercomputer center among them, are now doing serious research into how to implement it. And perhaps most important, some kind of digital data library is fast becoming a scientific necessity.

"We're now in a time of large, interdisciplinary projects such as global change, biomedical research, even astronomy," says astrophysicist Barbara Mihalas, who heads up the Illinois supercomputer center's effort to build a prototype digital data library. To get at the underlying causes of, say, liver cancer or diabetes, a researcher may need to integrate information from molecular biology, genetics, cell biology, pathology, patient records—"all the data from all experiments looking at all the processes that are relevant," she says. And yet the reality is that the databases are typically scattered through offices and institutions all over the country, if not all over the world.

As if that were not enough of a problem, she adds, there is the sheer, overwhelming volume of data involved, especially as megaprojects such as the Human Genome Project and the Hubble Space Telescope start pouring forth data by the gigabyte. The most dramatic example is NASA's Earth Observing System (EOS), a series of remote sensing platforms that the agency plans to orbit beginning in the late 1990s to do Global Change research; EOS is scheduled to send down more than one *trillion* bytes of data per day—for 15 years.

Doing research with this kind of data flow is going to be like drinking from the proverbial fire hose, says Mihalas, and massive computer assistance is clearly the only hope. At the same time, she adds, it's important to realize that better computer and network hardware is only one part of the solution. An even tougher challenge lies in getting scientists and administrators to change the way they think about data management.

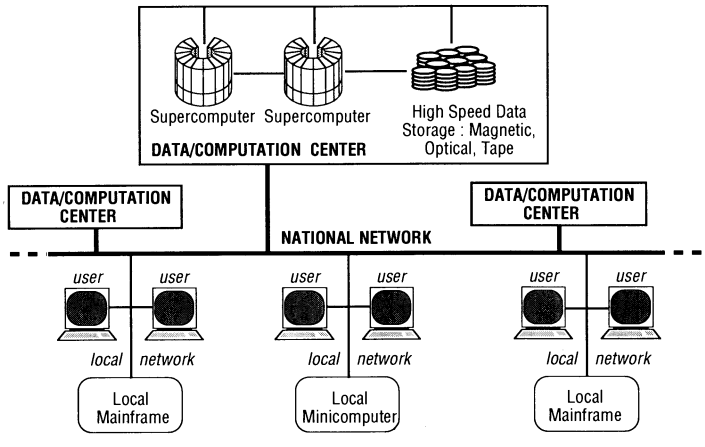
At most institutions, for example, data management has traditionally taken a backseat to more urgent priorities such as buildings, people, and hardware. As a result, databases are not only dispersed geographically, but they are rarely indexed by anything resembling a national card catalog. "It becomes like trying to find 'pineapple farming' in an encyclopedia that doesn't have an index and isn't in alphabetical order," she says.

Under pressure of necessity, however, some of the big, centrally managed programs have begun to pioneer a more systematic approach to data management. Just last year, for example, as part of the legislation authorizing the Human Genome Project, Congress established a Biotechnology Information Center within the National Library of Medicine, which is part of the National Institutes of Health. The new information center, headed by David J. Lipman, was given the task of developing software and database systems for *all* of biology—including the human genome once it is sequenced. In parallel, the Human Genome Project office is now organizing a Joint Informatics Task Force to figure out how to handle massive databases that will be generated by the sequencing effort itself.

NASA, meanwhile, is setting up its Astrophysics Data System, which is intended to give astronomers access, through the NASA science network, to archived data from Space Telescope and its other space astronomy missions—more than a dozen in all by the year 2000. Now scheduled to begin limited operation on 1 July, the system will also cover archives from all European and Japanese space astronomy missions.

As important as such organizational efforts are, however, they represent only one level of the conceptual change that's necessary to create a truly useful data library. An even deeper challenge arises from the fact that scientists—especially the most creative scientists—are very good at asking questions of the data that the database designers never thought of.

"Having this kind of flexibility is a very unusual requirement [to place on a database]," says James Ostell, software chief for the Library of Medicine's new Biotechnology Information Center. Conventional database software has been developed largely for business applications such as airline ticketing, where each record in the database typically contains only a few kinds of alphanumeric data—a name, say, or a social security number—and where the format of the data never changes.

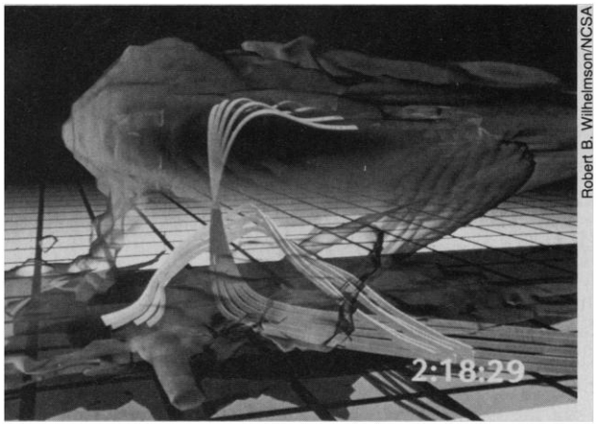


Data, data everywhere. *If the network were fast enough, any desktop workstation would act as though it held all the data in the world.*

"But in science," he says, "you actually have a hierarchy of information." The genome, for example, can be viewed as a sequence of DNA bases, as a genetic map, as a collection of chromosomes—all at different levels of organization. Moreover, new views of the same data arise constantly as new understanding develops. So how do you design a database that can be looked at in multiple overlapping ways, Ostell asks. And how can you revise it to reflect new ideas without instantly making it unreadable by every piece of software that anybody has ever written to use the data?

The full answers to those questions are still on the cutting edge of database research. But there does seem to be a consensus on the general approach: make the data "self-describing." That is, store certain key information about the data in the data file itself.

This might sound trivially obvious, says Mihalas. But look at how most laboratories handle data now. "Someone will make up a program to put data on a disk, in some order that seems reasonable to him," she says, "and then he writes down in a notebook how it's done." Or he may not even do that, she says: the only documentation may reside in the computer code itself. But either way, the information is *not* part of the data file.



Robert B. Wilhelmson/NCSA

Seeing the world anew. *Data alone means nothing until you can comprehend it. This supercomputer-simulated thunderstorm is a state-of-the-art example of computer visualization.*

In the self-describing approach, by contrast, that information would be encoded in the file right along with the data, according to some standard, widely known format. In addition, the file could include higher level information, such as whether a particular string of binary digits represents a DNA sequence, a gene map, chromosome crossing data, three-dimensional protein structure, or what have you. It might well incorporate commentary about the calibration of the instruments that took the data and the reliability and correctness of the results. And it could even include preprogrammed computer code that already knows how to read the data.

In practice, unfortunately, it is still not at all straightforward to make such a self-describing database work well. But prototypes are under active development both in commercial software companies and in a number of laboratory groups such as Milhalas'. And the payoff, she says, would be to transform a data file into something much more than a passive string of bits. The file would function instead like an expert consultant, ready to tell users exactly what they want to know with a minimum of fuss. Users could analyze the data any way they wanted with any software they wanted, and neither they nor their software would need to know how the data file was actually structured because the knowledge would be contained in the file itself. "If you don't just happen to love sorting out data formats," she says, "that makes a big difference."

However, even this doesn't go quite far enough. What about the researcher sitting at his or her desktop workstation who doesn't know precisely what questions to ask—and who certainly doesn't want to spend hours or days of precious research time trying to find out precisely which data sets might be relevant? What that user needs is a good librarian—preferably an intelligent, on-line, electronic librarian.

Perhaps the most ambitious proposal along these lines comes from Vinton G. Cerf and Robert E. Kahn, co-founders of the Corporation for National Research Initiatives in Reston, Virginia. Kahn and Cerf want to fill the national research network with a swarm of "knowbots"—autonomous computer programs reminiscent of benign computer viruses.

Knowbots, they explain, could be created, copied, or destroyed as needed. They could take up permanent residence in a given computer, or they could migrate through the network from one computer to another. They would communicate with each other by means of electronic messages. And they would have sufficient artificial intelligence to carry out the bidding of any given user—or, for that matter, to run the network and data system as a whole.

Suppose, for example, that a researcher in Idaho were to ask "What do we know about the genetics of disease X?" A user knowbot residing in his computer would interpret the request, determine how to go about answering it, and then create one or more messenger knowbots that would travel out to the appropriate databases. Once there, the messengers would request the necessary data from database knowbots, then convey the information back to the user's machine. Finally, still more knowbots would integrate the data and arrange to display it in a meaningful way to the user himself—who never has to know about any of this.

So how long before this nationwide hive of knowbots starts buzzing? "Steps are being taken even as we speak," says Cerf. His company already has funding from the National Science Foundation and the Library of Medicine to develop some of the basic knowbot software. Meanwhile, the National Science Foundation's NSFNet already provides nationwide networking service at 1.5 million bits per second, and will start going to 45 million bits per second by the end of the year. And several bills are now circulating in Congress to establish a national network operating at several billion bits per second. The upshot, says Cerf, is that some components of a large-scale digital library could be in operation as early as 1993.

Whatever the time scale, however, the advent of a nationwide digital data library could have profound implications for the conduct of science, he says. "As digital publishing comes on, we're going to see an enormous increase in the flow of information," he says. And yet as intelligent knowbots are developed, he says, "we hope to make it far more likely that researchers can get relevant information in a timely way."

Each individual user, for example, could have a personalized "filter" knowbot acting as his or her representative. "It would wait out in the system watching the data go by," says Cerf. "And when it saw something you'd like, it would snap it up."

In short, the knowbot-equipped digital data library would help researchers drink from the fire hose without being drowned. Indeed, as Kahn and Cerf wrote in their original 1988 proposal on the subject, "tireless knowbots making their endless journeys though information space," could ultimately augment human brainpower not just in science, but "in the collection, use, and creation of information in virtually every aspect of our lives." ■ **M. Mitchell Waldrop**