SCIENCE

7 JULY 1989
VOLUME 245
NUMBER 4913

Retrieval of Scientific and Technical Data

At one time it was possible for an individual to become knowledgeable about new developments impinging on his or her field by simply scanning a few publications. But with the continuing proliferation of new journals and of interdisciplinary work, this is no longer feasible. Electronic equipment is creating an enormous flood of data. At the same time, computers and electronics are providing means of storing and retrieving new and old information so that in many instances an individual can tap information even more effectively than did predecessors a generation ago. Procedures for making information more readily available to the user are evolving.

Starting in the early 1970s, a few bibliographic databases became accessible on-line. Since that time, the number of publicly accessible databases has increased to about 4200 and their character has changed. Full text is available in some instances. Numerical databases have come to constitute a substantial fraction of the total. A majority of the bases are devoted to legal or business matters, but about 30% are related to medicine, science, or engineering.

The National Library of Medicine draws on the world's medical literature and that of related fields to create 20 important databases, including Medline. The Chemical Abstract Service of the American Chemical Society has joined with associates in Germany and Japan in a network (STN) that provides 85 databases. Their resources include 8 million chemical abstracts dating back to 1965 and information concerning the properties of 9.5 million substances. The consortium is expanding its coverage of biology, physics, mathematics, engineering, and materials science.

Although a few organizations are the dominant creators or vendors of databases, as many as 1700 different entities maintain at least one base. This individual initiative, coupled with the procedures of vendors, has led to a wide variety of systems, services, command languages, protocols, and terminologies—in short, a kind of electronic chaos that limits the value of the vast store of information for the average user. The situation demands more and helpful front-end software to ease access. It also requires agreed-on standards.

The costs of on-line searches average about $120 per hour and range from about $26 to several hundred dollars per hour. Such costs do not seem to discourage business or the legal profession. However, universities find it necessary to limit the use of on-line vendors. Instead, they are tending to buy or lease diskettes or compact disc read-only memory databases (CD-ROMs) that can then be tapped by the campus computer network. Some campuses are being equipped with high-speed glass fiber networks, and professors and students have their own personal computers.

The CD-ROM is already having an increasing role in data storage and retrieval. A disc only 12 centimeters in diameter can contain the equivalent of 200,000 pages of book text. For databases that do not change rapidly or at all, the disc is an ideal repository. Currently the commercially available CD-ROMs are priced at about $6000 each. This covers the cost of creating the master from which the discs are produced as well as a profit. Once the master is paid for, the cost of producing additional CD-ROMs is only a few dollars per copy. The cost of a CD-ROM reader is only a few hundred dollars. When demand for the discs becomes substantial, their price will fall. Already competition has become evident. There are now nine producers of CD-ROMs containing the information available on Medline. Price ranges from about $1000 to $1500. Special beneficiaries of cheaper CD-ROMs would be physicians, scientists, and engineers in Third World countries. They would need some simple equipment, but the cost of the information provided would be a tiny fraction of what it has been. Already CD-ROMs containing vast amounts of geophysical data are being supplied for a few dollars.

In the United States, computer networks are playing an increasing role in the research and education communities. More than 600 networks with over 100,000 computers and workstations are currently linked by INTERNET. NSFNET, one of the key components of INTERNET, links 250 institutions and major laboratories. Its traffic has been increasing rapidly. During April 1989 it carried 800 million packets. A larger, faster, federally sponsored network is being advocated. This would facilitate enhanced exchange of complex numerical data and transmission of figures and images.—PHILIP H. ABELSON