Electronic Databases

Martha E. Williams

Much of the world's scientific and technical information is contained in the more than 2800 databases that are now available online, and that number is increasing weekly if not daily. The problem for many scientists and engineers is learning what databases there are and what systems allow access to them. The systems, often called information utilities or database vendors, provide online databases together with software for search and retrieval, data manipulation and modeling. The online search services are called information utilities because they permit you to use information even though the database is not in your possession. They permit the wide sharing of tremendous resources: databases, large computers, and sophisticated software through communication networks.

Summary. Electronic databases corresponding to most of the world's currently published literature and many other types of information are publicly available through online systems. Scientific databases that give references for publications are numerous and widely used; scientific numeric databases that are open to the public are far fewer and less used. Online retrieval systems are becoming easier to use as a result of the introduction of artificial intelligence techniques and user-friendly front ends and gateways. Issues related to electronic databases include public-private sector competition, transborder data flow, copyright, downloading, and the changing roles in database generation and processing.

Scientists, engineers, businessmen, lawyers, and many other users of databases learn quickly that online searching is not only useful but quite inexpensive. The average expenditure per hour for online database use is just about $100 including all costs (prints, connect time, online displays, and so on), but the average search requires only 10 to 15 minutes or only between $16 and $25. During the 10 to 15 minutes the user may search through several million records equivalent to the Library of Congress catalog or 20 years of *Chemical Abstracts*, or 10 years of the *New York Times* and dozens of other newspapers. Those who have done manual literature searching know that one would be hard pressed to pull 20 years of *Chemical Abstracts* index volumes off the shelf in 10 minutes, let alone look up just one compound and list the numbers for the references that refer to the compound and read the abstracts. For extensive exhaustive searching as is often done in the pharmaceutical industry and by patent attorneys, online searching can compress literally days or weeks of manual search time to a matter of hours or minutes, and this translates to a reduction in costs (for labor primarily) from thousands or even tens of thousands of dollars to hundreds or less.

Databases now serve nearly every major field—the traditional scientific disciplines, law, politics, social sciences, arts, and humanities—and specialized databases for such areas as oil spills, child abuse, automobile recalls, robotics, shipping, and even shopping. The areas best served and most used are business, law, medicine, and chemistry. Databases used in science, engineering, and technology are all referred to under the umbrella term "scientific." It would be difficult to separate these overlapping areas, as Harrison observed: "The sum of scientific, engineering, and technological knowledge is a continuously expanding resource of unprecedented richness and value. Without a term for this body of knowledge, it is frequently referred to as scientific knowledge" (*1*). She observed that the legacy of science, engineering, and technology is an array of concepts, an array of methodologies, and a body of knowledge consisting of a database. Much of the knowledge she referred to is recorded in publications and computer files and ultimately appears in electronic form in publicly available databases. It is those publicly available electronic databases that are of concern here.

There are many ways to classify databases, and a distinction among word-oriented, number-oriented (numeric), and picture-oriented (pictorial) types is made on the basis of the different machine representations for data. The computer software packages for handling these three types differ considerably. Word-oriented databases are comprised primarily of strings of characters, and consequently software that is adept at handling such strings is used; the computer processes strings of alphabetic and alphanumeric characters. The file structures and indexing in the computer for these may vary for the different word-oriented databases, but the string handling operations are basically the same.

In contrast, numeric databases require retrieval of numeric data, capabilities for manipulating them (for example, various statistical routines, regression analyses, and time-series) and routines for presenting the data in a format that is familiar to the users. There is less fetching and more processing involved. Statistical routines and other manipulation programs operate in the same way whether the data relate to sociological, economic, production, or chemical phenomena; numbers, symbols, or both, representing numbers are manipulated.

Pictorial databases contain primarily pictorial representations of, for example, chemical structures, physical particles, fingerprints, anatomical parts, or geographic and geologic maps. The types of computer processing involved are less likely to depend on string matching and computational capabilities than on feature extraction, pattern-matching, and coordinate matching. Since there are no publicly available online pictorial databases (other than chemical structure files wherein matching is done through connection tables rather than pattern matching), they are not discussed further here.

Databases, essentially organized collections of computer-readable information in a defined subject area, are sometimes classified in that way. The defined area may be a single subject or discipline (chemistry), multidisciplinary (chemical-biological activities), problem-oriented (environmental pollution), mission-oriented (space), or oriented toward certain types of transactions (trading stocks and bonds). Other classifications distinguish bibliographic (literature references) from nonbibliographic (all others), databases (bibliographic) from databanks (numeric), numeric from non-numeric, or refer-

The author is professor of information science, Coordinated Science Laboratory, University of Illinois, Urbana 61801.

ence (containing pointers to data sources) from source (containing the information or data).

Bibliographic Databases

Electronic databases in science, engineering, and technology appeared in the 1960's; they were word-oriented and contained bibliographic references to published literature. In the mid-1960's there were only a few dozen. A database directory published in 1976 listed 301 (*2*). By the end of 1984, there were more than 2800 publicly available electronic databases of all types (*3, 4*). Key databases, among the hundreds of scientific databases, cover disciplines such as chemistry, physics, biology, engineering, and medicine as well as interdisciplinary and problem-oriented areas.

Chemical Abstracts Service (CAS) led the way with Chemical Titles (CT) database in 1961 with limited coverage of 750 journals providing references to 68,400 articles (*5*). CAS introduced CA Condensates in July 1968 and in its first 6 months included 112,137 references. CA Condensates was replaced by CA Search, which corresponds to the references in the printed *Chemical Abstracts* and covers the chemical literature from 1967 to the present. The CAS Registry and Nomenclature database, which provides registration numbers, CA nomenclature, and synonyms for 6,910,000 chemical substances, was started in 1965. The average number of substances registered per year has been 345,000, but an average of 600,000 per year will be added during the next several years as substances cited in *Chemical Abstracts* from 1920 to 1965 are registered. CAS Registry numbers will provide links within and between multiple chemical databases (both CAS-produced and others), and the potential exists for linking virtually all chemical databases in a chemical network (*6*). The CAS Registry and Nomenclature system was developed to overcome the naming problem within CAS's own databases but already serves a wider purpose. There are now 10,565,000 names of chemicals in the system for an average of 1.53 names per compound, but most substances appear only once and have only one name. Compounds that appear more frequently in the literature or that are produced by multiple manufacturers exceed the average. The extreme case is polyethylene with 1411 names, followed by polychloroethylene with 927 names.

Most of the major scientific bibliographic databases started in the late 1960's and early 1970's. The American Chemical Society's CAS was soon followed by BioSciences Information Service with the BIOSIS Previews database; it corresponds to the printed *Biological Abstracts* and *BioResearch Index* (now called *Biological Abstracts/RRM*). BIOSIS Previews became available commercially in 1969 and by the end of 1984 had grown to include 4,264,000 references. The major databases in engineering are Engineering Information's COMPENDEX database that corresponds to the *Engineering Index* and the Institution of Electrical Engineer's INSPEC-B and -C databases that correspond to *Electrical and Electronics Abstracts* and *Computer and Control Abstracts*. COMPENDEX, which began in 1969, had 1,685,669 references by the end of 1984. The INSPEC databases started up in 1970 (with information going back to 1969). At the end of 1984, INSPEC-B had 776,569 references and INSPEC-C had 504,582. Major databases in physics are the American Institute of Physics' SPIN database and INSPEC-A that corresponds to *Physics Abstracts*. SPIN, dating from 1970, had accumulated 375,201 through 1984, and INSPEC-A, 1,509,176 items.

The two major databases in medicine are Medline of the National Library of Medicine (NLM) and Excerpta Medica of Excerpta Medica, B. V. Medline corresponds to the printed *Index Medicus*. The Excerpta Medica database corresponds to the print product of the same name.

The largest multidisciplinary scientific database is the Institute for Scientific Information's SCISearch; it covers virtually all areas of science and technology and corresponds to the *Science Citation Index*. SCISearch includes not only references to articles from over 4000 journals but also cited references that appear in the bibliographies of these articles. Apicultural Abstracts and Chemoreception Abstracts are specialized life science databases. Envirotapes and Energytapes are problem-oriented databases that include material from many disciplines.

Full-Text Databases

A relatively new trend in electronic databases is the creation of full-text databases for search and retrieval on the systems that traditionally have provided bibliographic databases. A full-text database contains the entire text of documents such as wire service stories, legal cases, statutes, encyclopedia articles, journals, or textbooks (*7*). The Mead Data Central (MDC) LEXIS database is not only one of the largest databases in the world but was one of the first full-text databases, having begun in the 1960's. LEXIS contains the full text of legal cases as well as statutes and many other types of legal documents. Competing with LEXIS in the legal database area is WESTLAW of West Publishing Co. Initially, the two took different approaches, with LEXIS going the full-text route and WESTLAW following the more conventional surrogate (digest) route. WESTLAW provided indexing and surrogates, together with case names, citations, headnotes, and key numbered topics. In 1978, West added full-text information to the WESTLAW database to remain competitive.

Law was the first economically successful application area for full-text databases because lawyers often require the full texts of cases together with applicable statutes, rulings, and so on. The legal databases keep growing as the body of knowledge expands with new cases and with new and changing statutes and regulations. Unlike many fields where recent findings replace prior ones, in law the earliest case of a given type may set the precedent and become the linch pin of a subsequent case. The same is true with respect to patent databases—new patents do not supplant older ones; the object of most patent searches is to determine whether there is prior art. The database is constantly expanding and the recorded data, regardless of age, are essential for practitioners. These are two of the reasons for the tremendous success of legal databases.

Newspapers (for example, the *New York Times* and the *Wall Street Journal*), wire services (for example, United Press International and Associated Press), news magazines (for example, *U.S. News and World Report* and *Newsweek*), and newsletters (for example, *World Environment Report* and *Financial Management Advisor*, which are among the 250 newsletters offered by the NewsNet search service) are examples of full-text databases providing access to news. For a number of years, the New York Times Information Service (NYTIS) was in the forefront of news databases offering its InfoBank database online in 1973. In 1981, NYTIS first put the full text of *New York Times* articles online, but MDC beat them with NEXIS, a full-text database covering newspapers, news magazines, and the wire services. In 1983 MDC took over the NYTIS databases through a long-term license agreement, making MDC the leader in news databases as well as legal databases.

Among the full-text databases available for researchers are journals, reference works, and encyclopedias. The electronic version of *Encyclopaedia Britannica* is available through MDC, the *Academic American Encyclopedia* from Bibliographic Retrieval Services, Inc. (BRS), Dow Jones, and CompuServ, and, of interest to chemists, is the *Kirk-Othmer Encyclopedia of Chemical Technology*. In 1961 CAS produced the first electronic database offered for searching by the public, and by 1983 the full texts of 18 ACS journals were available for searching on BRS. BRS and BRS/Saunders provide more full-text scientific journals than any other online vendor. BRS/Saunders' Colleague system provides easy access to databases in the Comprehensive Core Medical Library of databases, including major medical journals such as the *New England Journal of Medicine* and *Lancet* and textbooks (*8*).

The legal full-text databases have been eminently successful in terms of wide use and high revenues, but the record, to date, with encyclopedias and full-text journal databases, such as the Harvard Business Review database, has not been remarkable. Full-text databases may well be more successful on the optical disk medium, which can accommodate extremely high volumes of information (billions of characters) that can be recorded on a single disk and distributed to the user at relatively low cost. Optical disks may replace some magnetic disk databases and may complement others by providing a means of displaying pictorial information which is often needed with medical texts, for example.

Numeric Databases

A numeric database is an electronic database that contains primarily numeric data. A numeric datum consists of a numeric value and one or more attributes of the value. Numeric data are measured, observed, or calculated quantities, such as the boiling point of a specific compound, wind velocity at a specified location and time, or the average price of a class of products over time. Naturally, some non-numeric data are included in numeric databases; the labels and legends associated with tables and graphs, for instance, and the nominal numeric designators (code numbers, document numbers, and product numbers) are not numeric data. Many numeric databases are actually database systems because they do not exist apart from a system for locating, retrieving, processing, and analyzing the data that they contain (*9–12*). Numeric databases, like word-oriented databases, include a wide variety of types and are frequently distinguished by subject or discipline, multidisciplinary areas, problems, missions, and the like.

There are scientific and nonscientific numeric databases. Within the first category, a distinction is made between the databases that are strictly science-oriented and those that are more business-oriented. Science-oriented databases provide data resulting from, and used in, the conduct of scientific research, development, testing, and evaluation in the physical, engineering, and life sciences (theoretical, observational, and calculated data regarding physical, chemical, and biological materials, objects, and phenomena). Business-oriented databases that are scientific deal with the business aspects of science, engineering, and technology (production data, sales export and import figures, transportation data, plant capacities, plant locations, and so on). These are practical distinctions based on observation of the databases that are now on the market. A rigorous taxonomy of types of information and data would likely produce a continuum with considerable overlap among and between types of data, but many of the data represented in such a taxonomy would probably not be found in current publicly available databases.

If one considers a continuum of types of numeric data ranging from the abstract physical to the social (physics, chemistry, engineering, biology, medicine, social science, politics, and economics-finance-business), certain general observations may be made about the data. Proceeding from the physical to the social, data range from time-independent to time-dependent, from stable to highly changeable (because of social factors), from those associated with a few variables to many variables, from predictive to nonpredictive, from less susceptible to subjective interpretation to more susceptible to subjective interpretation, and from highly reproducible to less reproducible; they also range from observations and measurements of physical phenomena, properties and events, through observations and measurements of biological organisms, to individual humans, to groups of humans or social organisms such as markets, companies, industries, cities, or countries. It is the human-social character of data at the social end of the continuum, where business data fit in, that contributes to the high use of business data online. The data are needed for business decisions, and the data are changing continuously. Therefore the users must have access to systems that provide the data in real time (for example, commodities markets, monetary exchange rates, and credit checks).

Some bibliographic databases, such as PHYSCOMP, point users to articles containing data compilations, and there are hundreds of bibliographic databases that direct the user to published articles and reports containing research results; but there are relatively few publicly available online compilations of hard scientific data. Many databases (such as the DRI Chemical Data Bank), deal with the business aspects of science. Similar types of data are provided for electronic products, agricultural products, and others. Such databases are essential to the business aspects of science, technology, and engineering but do not contain the scientific data required for scientific research.

DIALOG Information Services, Bibliographic Retrieval Services, System Development Corp. (SDC), the National Library of Medicine, and the other vendors that grew up in the bibliographic world do provide some scientific numeric databases but for the most part they are textual-numeric and not primarily numeric. Those vendors that originated in the numeric data time-sharing environment provide access to business databases primarily, and even though some (Data Resources, Inc., and I. P. Sharp, for instance), have added scientific databases, these are business-oriented scientific databases. The largest service that provided primarily scientific numeric databases was the Chemical Information System (CIS), which is no longer operated by the government. Even though CIS was the largest such service in 1983, the total number of connect hours for that year was under 7000.

Although the public use of scientific numeric databases is low, the use of scientific word-oriented databases is not. Scientific word-oriented databases make up slightly more than half of the word-oriented databases offered online and more than a third (36 percent) of the U.S. usage of databases provided through those systems is of the scientific databases (*13*). Use of these scientific databases is outranked only by legal databases. Legal databases make up only 2 percent of the word-oriented databases, but the usage of legal databases represents 38 percent of the total U.S. usage of this class of databases. The two largest vendors in this group are Mead Data Central and DIALOG, both of which experienced approximately half a million hours of use in 1983 in the United States alone. The use of scientific databases might be even greater if scientists and engineers could pass on the cost of searching to customers as lawyers can. Scientists and engineers do pass on the cost ultimately in the form of research and development costs, but such compensation may be many years removed from the consumer's purchase of the resultant products or services.

The number of scientific numeric databases that are publicly available through online and time-sharing systems, excluding the business-oriented ones, is also quite small. If one includes the 63 textual-numeric databases such as the TOXICOLOGY Data Bank of the National Library of Medicine, there are 83, but if one excludes these (36 of which are chemical) there are only 20 scientific, nonbusiness oriented numeric databases (*3*). Certainly there are thousands of in-house, company-restricted, and government-restricted scientific numeric databases in electronic form, but this article is concerned only with those databases that are available to virtually anyone in the free world who is willing to pay for access to them.

The Chemical Information System was originally developed by the National Institutes of Health and operated by the Environmental Protection Agency (EPA) starting in 1973. It provided access to 20 chemical databases including physical and chemical properties (x-ray and thermodynamic), spectroscopic data (mass spectra, infrared spectra, and carbon-13 nuclear magnetic resonance), biological data (nucleic acid sequences), toxicological, regulatory, and environmental data as well as an electronic mail service and linkages through the CAS Registry Numbers (for chemical substances) to bibliographic databases available on DIALOG, SDC, and NLM (*14*). In response to recommendations made to EPA and because of alleged mismanagement (*15*), CIS was dropped by EPA at the end of November 1984. To ensure continued availability of its databases to the 600 CIS subscribers, CIS databases were licensed by the government to two private sector organizations, Fein-Marquart Associates and Information Consultants, Inc.

At the hub of CIS is the Structure and Nomenclature Search System (SANSS) containing more than 800,000 names for more than 250,000 substances. SANSS functions both as a substructure search system and as a locator utility to gain access to other CIS databases. Within the EPA CIS family of databases, SANSS, which was used most, totaled fewer than 2000 hours in 1983, followed by the Mass Spectral Search System and the Oil and Hazardous Materials Technical Assistance Data System, which together were used only about 120 hours per month or 1500 hours in a year. The next most used database had only 30 hours of use per month, and most CIS databases had fewer than 10 hours of use per month which, at rates of $55 or $85 per hour, could not support their shares of the maintenance cost of the system and could not touch the database production costs.

Technical Database Services is currently developing a line of hard data including physical, thermodynamic, transport, and other properties and specifications for chemicals, materials, and mixtures. It has implemented for online search and retrieval Log P (produced by the Medicinal Chemistry Project of Pomona College) (*16*) and will soon introduce thermodynamics and transport data from the Texas A&M thermodynamics database. Most numeric databases provide information in business and economics rather than science and technology. Publicly available scientific databases are still relatively few in number, and their use is low in comparison with that of either bibliographic databases or business-oriented numeric databases. Usage of business numeric databases, including commodities quotation systems, was certainly in the tens of millions of connect hours (because of the kinds of price structures used for numeric database services, they do not measure performance in terms of connect hours but by revenue) whereas use of the (hard) scientific numeric databases was less than 30,000 connect hours.

If new scientific numeric databases are to become successful (become sufficiently used to remain viable), they must provide not only high-quality data packaged in a manner that is amenable to the intended user, but they should include one or more attributes of the databases that have proven successful. Chances of success may be increased if databases include data that (i) are continuously changing and therefore need to be updated frequently (as seen in the commodities market); (ii) are needed by many people daily as tools of their trade (as seen in legal databases); (iii) are of permanent value and are not replaced by new data so that the entire database is always of value to users regardless of its continued growth (as is the case in legal and patent databases); and (iv) are too voluminous for individual organizations to maintain but that are recognized as essential to the functions of specified industries. It is regrettable that in a country where considerable expenditures of time and dollars are made for scientific research we have been unable to produce and maintain, on an economically stable basis, scientific numeric databases that could be shared by many researchers.

Access to Databases

The principal mode of access to electronic databases is through online search service organizations. These vendors, so called because they sell access to databases produced by others, also provide software used for search and retrieval (including a host of commands, features, and functions) and the added value associated with the way they load the file (break the records into fields with subfields and then sort and index them to permit search and display either separately or in predefined sets).

There are 362 vendors (*3*) of online and time-sharing search services worldwide but only a few account for most of the search activity. Vendors provide the windows through which users see the data in databases. Vendor organizations can be classed according to the majority of the kinds of databases they process—that is, mostly word-oriented databases or mostly numeric databases. As their efforts to expand markets continue, most are branching into other areas. Vendors of word-oriented databases are acquiring numeric databases and vice versa, and both types are adding services such as electronic mail and messaging systems. Although these other services are not database services, when used habitually they make the users more dependent on the vendor. The closer a vendor comes to being a one-stop information and computer service organization, the more likely it is to keep its customers.

Among the vendors of online numeric databases, there are those that deal primarily with business databases—including the business aspects of science—and those that deal with (hard) scientific databases. Among numeric database services are Data Resources, Inc. I. P. Sharp, General Electric Information Services Company, Chase Econometrics and Chemical Information System. All but the latter are principally business-oriented. Among the vendors of word-oriented databases are Mead Data Central, DIALOG Information Services, Bibliographic Retrieval Services, System Development Corporation, the National Library of Medicine, CompuServ, The Source, Dow Jones and Company, Inc., and NewsNet, Inc. The first five are examples of vendors whose markets are largely in institutions and their databases are geared for serious research. The second four are vendors whose customers are largely individuals and their databases are consumer-oriented or geared for popular appeal. Consumer systems are not discussed further.

Data Resources, the largest of the systems that provide business-oriented numeric databases, began in 1969 with about 3000 time series (data values associated with a defined phenomenon observed at equally spaced times); at the end of 1984, there were approximately 20 million time series. DIALOG, the largest online search service, provides mostly word-oriented bibliographic databases. It began providing commercial search services in 1972 with two government produced databases, ERIC and NTIS. In 1972 those two databases contained approximately 100,000 references and the service was operated on an IBM 360/30 (less power than many personal computers have today). At the end of 1984 DIALOG had 200 databases with 100 million items (bibliographic references and other types of records) occupying 250 gigabytes of storage on two large-scale dual processors (Hitachi NAS 9000 series machines).

The National Library of Medicine began operation as an online search service organization in October 1971 with one database, Medline with 147,000 references. At the end of 1984 Medline alone had grown to 4,504,089 references and the full set of 21 databases offered included 8,682,798 items. Mead Data Central introduced commercial online service in 1973 with its own LEXIS database and the National Automated Accounting Research System (NAARS) database, operating on an IBM 370/155. The 1973 database included 208,000 documents or 2.5 billion characters. By the end of 1984, MDC was processing databases within ten major services containing 18,290,000 documents or 80 billion characters of data, operating on four Amdahl 5860's with a total capacity of 52 million instructions per second. In 1973 a staff of 12 handled the databases and search services; the staff had grown to 1300 by the close of 1984.

Aids to Online Retrieval

Most of the world's currently published literature, referenced or included in more than 2500 databases, is readily accessible not only to the trained information searchers in corporate information centers and special, academic, and public libraries throughout the world but also to scientists, engineers, and researchers of all types. Although public online search services have been available since the early 1970's, virtually all of the initial users were trained searchers who learned to deal with the specific commands, protocols, features, responses, and messages of each system that they used. In the early years fewer than a dozen online systems were open to the public. Now with 362 online services (*3*) the problem has increased. While few searchers maintain proficiency in the use of more than five to ten systems; even that is much more than end-users can manage. Not only is there a problem in being "conversant" with multiple systems, but there is the difficulty of knowing what databases exist and what systems allow access to them. These problems have long been recognized and addressed by researchers in information science (*17*).

Research programs going back to the early 1970's paved the way for the development of "transparent systems," "user-friendly front ends," "intermediary systems," and "gateway systems" to online services. These new systems all have the same aim—to make the complexities of online searching transparent to users. Virtually all of the transparent systems that are commercial products have one feature in common; they all provide automatic dialing and log-on to the search service.

The terminology seen in the trade literature is imprecise and evolving, as is the case with much of the terminology associated with computer technology. The following definitions may be helpful. User-friendly front ends are software interfaces to online systems which are easy for the user to operate and run a microcomputer when the microcomputer is being used for online access. They are easy to use and simplify the access to the search service, hence "user-friendly." They are front ends in the sense that such a system is used in front of, or before, the target search service. User-friendly front ends permit one to enter search terms off-line and to connect to the search service by merely striking a single key (specific keys are programmed for specific systems—one for DIALOG, another for Bibliographic Retrieval Services and the like). The software on the user's microcomputer performs the search automatically, downloads (transfers from the host mainframe to the user's microcomputer) the search results, and logs off. The user does not need to know many specifics about the search service command language, syntax, protocols, and so on; thus the novice user can perform a search more easily. Also, since the keying of search terms is done off-line on the user's microcomputer, connect time is reduced as are the associated charges. User-friendly front ends can aid the novice searcher and the casual or occasional searcher who does not wish to memorize the detailed specifications required for online searching. They may also be useful to the experienced user who is familiar with the terminology and contents of specific databases. On the other hand, user-friendly front ends remove from the user the interactive aspect of the search service and the facility for iteratively developing and improving search strategies. The cost of the interface software and the added cost of the searcher's time expended while developing the search off-line are considerations, too. Since a novice user does not know how to exploit the full potential of the search software, the user-friendly interface is a good way to overcome the initial resistance that many have to conducting information searches on a computer. The advantages are more for novices than for experienced users.

Intermediary systems are microcomputer software packages that are designed to function as substitutes for intermediary searchers. They are intended to help the user, as an intermediary searcher does, in negotiating a search question on an online search system. They are generally menu-driven (giving a series of prescribed options) or prompted (providing suggestions and instructions) systems that guide a user through the search process of selecting terms, phrases, and term fragments (truncated terms) and relating them through Boolean and proximity logic operators. Intermediary systems are usually database specific—that is, they help the user conduct a search in a specific database or family of databases. They are often developed by the database producer to reach a larger market. Examples are Sci-Mate for ISI databases, Disclosure, Inc.'s Microdisclosure for searching the Disclosure database on DIALOG, and Search Helper, designed for use with the Information Access Corporation family of databases (Magazine Index, National Newspaper Index, Legal Resource Index, Trade and Industry Index, Newsearch, Management Contents, and the *Computer Database*.)

A gateway system directs a user to one or more of multiple online systems and ultimately to one or more databases within the system. It is more than a simple automatic log-on. Gateway software should contain intelligence tha' automatically routes a question to ai appropriate system without the use needing to know telephone numbers protocols, and passwords. Research o gateways began in the mid-1970's (17) and operational systems started in the early 1980's. Among the more sophisticated current gateway systems are the Intelligent Gateway System of Lawrence Livermore National Laboratory, iNet of Bell Canada, and the International Institute for Applied Systems Analysis IIASA TPA/70 Gateway in Austria.

Most commercially available front ends and intermediary systems are provided as floppy disk software packages, but there is no technical reason why they cannot be located at other points within the network. For example, an intermediary system might reside on a host search service. In fact, intermediary software could be developed for every database, or, a generic system could be developed for the online search service itself. Intermediary systems and other such facilitators could reside in a separate facility and be called up as needed to provide entrée into required databases and database search systems.

The entire area of transparent systems for aiding online database searching and the development of user-friendly interfaces, front ends, and the development of gateways is ripe for the application of artificial intelligence (AI) techniques and the development of specialized expert systems.

Artificial intelligence is concerned with the "design of intelligent computer systems: that is, systems which exhibit the characteristics commonly associated with human intelligence—understanding natural language, problem-solving, learning, logical reasoning, etc." (18). The proponents of AT predicted its use in information retrieval for understanding natural language documents in the 1960's. Considerable research was underway, but progress was limited to applications for very small databases in very limited domains. It has been only in the past decade, with the increased speed and decreased cost of computers and the facilities for handling large files (containing knowledge bases) that wider scale testing of AI techniques has been possible and AI software for very specific applications has become operational in real world situations. Artificial intelligence is finally entering the world of online retrieval and electronic databases.

Artificial intelligence techniques are used in expert systems. An expert system "combines the storage capacity of a computer for specialized knowledge with its ability to mimic the thought processes of a human expert" (18). The mid-1980's are witnessing the development of expert systems to aid online retrieval by making them more transparent. Intelligent gateways and intelligent intermediary software will mimic the thought processes and actions of expert searchers. The achievements are in limited domains and the greatest strides have occurred in medicine.

Issues and Trends

Public-private sector. Databases are produced in both the public (government organizations) and the private sectors (not-for-profit and commercial organizations). All producers of publicly available databases charge for their use and vendors charge for the search services. Charges are levied for various reasons: to cover the costs of database generation, to develop new products, and to generate profit (in the commercial sector); to cover the cost of data production and new product development (in the not-for-profit sector); to comply with Office of Management and Budget imposed mandates to recover costs (in the government sector); or, as in the case of the National Library of Medicine, to limit the amount of use to stay within the bounds of computer limitations. The fees charged by government agencies are generally lower than those charged by industry. The rationale for lower prices is that because the databases are produced with the use of tax funds, there is a social responsibility to make them available to a wide community of users, and price, they say, should not constitute a significant barrier to use.

The use of low prices by government database producers is viewed by the private sector as unfair competition. The private sector is concerned that low pricing by the government tends to create an impression, among users, that higher prices represent exploitation, whereas in fact, since government databases are partially subsidized by tax funds, the pricing may be artificially low. Also, any successful database produced by the government is viewed by the private sector as a lost opportunity for them. Needless to say, if the government databases had not proved successful, no one would have cared. However, the successful government databases did not start out that way. The government carried out and sponsored much of the pioneering work in the development of electronic databases and online systems. The problem becomes more acute when the same subject area is covered both by a government database and a private sector one. This occurred in the instance of the National Library of Medicine whose Medline database, which is one of the three most used word-oriented databases (13), was considered by Elsevier to constitute unfair competition for its Excerpta Medica database (19). In general, government databases are priced low and tend to deal with defense, essential services (education, security, emergency and disaster information, and so on), and biomedicine, whereas commercial databases are more business-oriented and are more likely to produce profits in the near term.

Transborder data flow. Transborder data flow is a problem because there is disagreement between those who favor the free flow of information and those that are concerned about safeguarding trade, business, and financial as well as scientific, technical, and governmental secrets. In the world of commerce, maintaining proprietary information within a company is of utmost importance for achieving and maintaining a competitive position in the marketplace. A country's advances in science and technology and the use of the resulting information lead to the development of products and processes and ultimately provide benefits in terms of national defense and prosperity for the country. On the other hand, information is a societal good and can be used for the benefit of mankind regardless of national borders. The usual safeguards employed are copyright, data rights, and trade restrictions, on the one hand, and reciprocal agreements or bilateral agreements on the other. One of the deeper problems that goes beyond formal agreements for giving or exchanging information is the fact that data in electronic form can often be transmitted without the knowledge or permission of the data owner, and the transgression is difficult if not impossible to detect.

Transborder data flow is an issue that is tied in with the public-private sector problem and with copyright. Those who wish to keep databases in the public sector for the public good are generally those who also favor the free flow of information. They view information as more of a societal good than as a commodity. Those who argue that information is an economic resource and a commodity generally favor keeping databases in the private sector with the full protection of copyright. Emotions are strong on both sides, but it is not an either-or situation. Information is generated, processed, and used by all parts of the public and private sectors and there cannot and should not be exclusive claims to particular domains.

Copyright and downloading. Microcomputers (including inexpensive personal computers) allow access to databases by both trained searchers and end users. This is a mixed blessing as far as database producers are concerned. The expansion of the number of users is positive as is the fact that databases are beginning to be used more widely by end users. On the other hand, since microcomputers have both processing power and memory, it is possible to program them (or acquire programs that are commercially available) to download information (transmit from a mainframe computer, for instance, the online vendor's computer, to a local memory device such as a microcomputer or intelligent terminal) and store it on the microprocessor for reuse. This presents a problem because the databases are generally under copyright (*20*), and the database-producing organizations are dependent on use fees or royalties from database usage to pay the cost of developing and maintaining the database. Producers do not want their databases to be copied and reprocessed locally without their knowledge and fair payment. Many permit downloading for one-time use, that is, in order to reformat or rearrange search output, but this does not include continued storage and re-searching.

Database producers provide their databases to the online search services or vendors under a license agreement that provides for compensation back to the producer. Pricing strategies are developed by the vendor that will permit compensation to the vendor and to the producer. Additional revenues are generated for each use of a database. If databases are downloaded for repeated local searching, revenues will decrease. While downloading provides a cost savings for users, it may ultimately be self-defeating. Loss of revenues to database producers (especially the nonprofit producers whose margins are generally low) and vendors may force them to raise prices and may ultimately cause the demise of databases. Another worrisome prospect is the possibility that commercial organizations will download substantial subsets of databases and then reformat the data or augment it with additional data to produce a new database product, or both. This practice not only would decrease revenues of a database producer but would make the producer of the original database compete with the new producer who would be using the original producer's own data. The downloading phenomenon is growing rapidly because it is easy to do and because there are so many commercial software packages (front ends, and so on) that provide for it. Database producers are attempting to develop pricing schemes that will permit downloading for reuse and resale but with adequate compensation to the producers.

Related to the problem of copyrighted databases and producer proprietary interests is the contention between authors and publishers regarding the proprietorship of intellectual property. Authors are becoming increasingly concerned as their works become part of electronic databases, and it becomes easy for others to copy them. They are also concerned about compensation as the electronic publishers make additional uses of the authors' works. These and related issues are being addressed by the U.S. Congress Office of Technology Assessment Advisory Panel for Intellectual Property Rights in an Age of Electronics and Information.

Optical disk. Databases have traditionally been distributed as magnetic tapes, and small specialty subsets have been distributed in the form of floppy disks. Now optical disk or videodisk technology provides a new means of database distribution and a compact way of storing archives. Optical disks offer several important features. They permit extremely dense storage (500 million to more than a billion characters per disk side depending on the manufacturer), the ability to store not only text but graphic or pictorial information, and a low cost per copy, after the master is created. Optical disks are a new way for database producers to distribute their products.

Some optical disks are read-only, that is, once created, data cannot be altered. More recently, read-write (write once but one cannot overwrite) disks have become available. Disks containing entire databases, that would otherwise require dozens of magnetic tapes, can be reproduced for only a few dollars each, but this does not mean that database producers will lease, license, or sell their databases for a price that is close to the cost of the physical disk plus copying cost. As is the case with other types of publications, the physical production costs are a tiny fraction of the total database creation costs, which must be recovered. The potential exists for disks to replace online usage of databases. However, if optical disks were to make significant inroads in the online market, then the producers would lose the revenue from online use. The advantage optical disks would provide to producers is that the producer would know who his customers are, whereas with online database use the online vendors maintain that the users are their customers and do not readily communicate full usage information back to the producers. Producers must weigh the benefit of user contact against the potential loss of income from online use of their databases when deciding whether to distribute in the disk medium.

Changing roles in database generation and processing. In the early days of database development and online services, there were clear lines of demarcation among publishers of hard copy primary journals, database producers (who were also the publishers of secondary abstracting and indexing journals), online search services, and telecommunication companies. These distinctions are breaking down as the roles become intermixed. Some of the changes have resulted from mergers and acquisitions in the electronic database industry. Many new alliances among all types of database producers, online and time-sharing services, primary publishers, secondary publishers, and even telecommunications companies have been formed. While the merger-acquisition phenomenon indicates that the industry is maturing, there are some potential dangers, such as the possibility of monopolies, the possibility of sacrificing database quality to profits, and, for the United States, the possibility that the control of major and essential information resources could be transferred out of the country as many U.S. information organizations have been bought by foreign organizations.

Not all changing roles are attributable to mergers and acquisitions. Some online services have begun developing databases in order to increase their online offerings in specific fields. Some database producers have begun to offer their own and other databases online, thus becoming vendors. They see this as a way to regain contact with their users. Such moves toward vertical integration may well help the producer-vendor, but they are viewed by some as tending toward monopolies. Depending on one's perspective, horizontal integration or extending the supermarket (one stop) approach to database offerings can also be viewed as problematic.

Most of the major publishers of books and journals view themselves as electronic publishers. They are creating full-text databases in conjunction with their computerized composition, and they are eagerly looking to put the databases online in order to capture additional revenues without jeopardizing their print sales. This creation of full-text databases and their use online has not yet proven profitable, but more of them are going online and it may take a critical mass in given subject areas to make them really attractive and useful.

Conclusion

Online databases for use by the general public have been available for little more than a decade, but in that short time the volume of items available has grown more than a thousandfold. Although limited in subject coverage initially, there are now databases that appeal to users in virtually all disciplines. The number of services has increased as has the value received per dollar (even though the price per hour for given databases may have doubled in 10 years, the number of records in the databases has far more than doubled). Online systems have become much more sophisticated and at the same time intermediary systems have been developed to make them easier to use. The fledgling activities of the early 1970's are a part of a successful industry with entrepreneurs appearing everywhere. Research and development is continuing and the research and development, together with the enthusiasm and excitement of the entrepreneurs, may well lead to the day when using online databases is as routine an activity as using telephones.

References and Notes

1. A. J. Harrison, *Science* **223**, 543 (1984).
2. M. E. Williams and S. Rouse, Eds., *Computer-Readable Bibliographic Data Bases: A Directory and Data Sourcebook* (American Society for Information Science, Washington, D.C., 1976).
3. *Directory of Online Databases* (Cuadra Associates, Santa Monica, Calif., Fall 1984), vol. 6.
4. M. E. Williams, L. Lannom, C. G. Robins, Eds., *Computer-Readable Databases: A Directory and Data Sourcebook* [American Library Association, Chicago and Elsevier, Amsterdam) in press (in two volumes)].
5. Statistics regarding database sizes and vendor file sizes represent information as of January 1985.
6. M. E. Williams and K. MacLaury, in *Computers in Chemical Education and Research*, E. V. Ludena, N. H. Sabelli, A. C. Wahl, Eds. (Plenum, New York, 1977), pp. 3–23.
7. C. Tenopir, *Annu. Rev. Inf. Sci. Tech.* **19**, 215 (1984).
8. BRS/Saunders is the name of the joint venture between BRS and W. B. Saunders.
9. J. A. Luedke, Jr., G. J. Kovacs, J. B. Fried, *Annu. Rev. Inf. Sci. Tech.* **12**, 119 (1977).
10. D. R. Lide, Jr., in *Development and Use of Numerical and Factual Data Bases* (AGARD Lecture Series 130, Nevilly sur Seine, France, 1983), pp. 8-1 to 8-6.
11. J. R. Rumble, Jr., in *Proceedings of the Fifth National Online Meeting*, M. E. Williams and T. H. Hogan, Comps. (Learned Information, Medford, N.J., 1984), pp. 325–330.
12. S. V. Meschel, *Online Rev.* **8**, 77 (1984).
13. *Information Market Indicators: Information Center/Library Market* (No. 5, Information Market Indicators, Monticello, Ill., 1984).
14. G. W. A. Milne, C. L. Fisk, S. R. Heller, R. Potenzone, Jr., *Science* **215**, 371 (1982).
15. J. L. Fox, *ibid.* **225**, 483 (1984).
16. The Log P database is a compilation of partition coefficients (P) (equilibrium ratio of the concentration of a substance in an organic solvent to the concentration of that substance in water) and has 30,000 measurements for more than 5,000 organic compounds. Log P values correlate with physical-chemical phenomena such as solubility, absorption, and transport. They also correlate with biological and medical properties of molecules and can be used to predict these properties. Because the octanal-water system, which is the standard for the Log P, emulates the chemistry of a cell membrane (lipid-aqueous) it is a good indicator of uptake and absorption and for predicting biological activity. Consequently, Log P is useful for environmental analysis and in the design of drugs and pesticides.
17. R. S. Marcus and F. J. Reintjes, *Computer Interfaces for User Access to Heterogeneous Information Retrieval Systems* (Report ESL-R-739, Massachusetts Institute of Technology, Cambridge, April 1977); V. Hampel, S. K. McGrogan, L. E. Gallo, and J. E. Swanson, "The 'LLNL' meta-machine: a flexible extensible and practical technique for interactive data management, modeling and distributed networking," paper presented at the Fourth Berkeley Conference on Distributed Data Management and Computer Networks, August 1979. M. E. Williams and S. E. Preece, *Proc. Am. Soc. Info. Sci.* **14** (1977); A. E. Negus, Development of the EURONET-DIANE command command language," in *Proceedings of the Third International Online Information Meeting* (Learned Information, Inc., Medford, N.J., 1979), p. 95–98.
18. A. J. Meadows, M. Gordon, A. Singleton. *Dictionary of New Information Technology* (Kogan Paye, London, 1982), pp. 14 and 67.
19. M. M. Cummings, *Science* **205**, 265 (1979); R. S. Willard, *ibid.* **217**, 586 (1982); J. Leiter, *ibid.*, p. 982.
20. Databases are fully covered by copyright (section 101 of the Copyright Act of 1976). They are considered to be literary works, collective works, and compilations and factors such as form, media, language, and coding do not alter the copyrightability.

Intelligent Tutoring Systems

John R. Anderson, C. Franklin Boyle, Brian J. Reiser

Summary. Cognitive psychology, artificial intelligence, and computer technology have advanced to the point where it is feasible to build computer systems that are as effective as intelligent human tutors. Computer tutors based on a set of pedagogical principles derived from the ACT* theory of cognition have been developed for teaching students to do proofs in geometry and to write computer programs in the language LISP.

Computer systems for intelligent tutoring are being developed to provide the student with the same instructional advantage that a sophisticated human tutor can provide (1, 2). A good private tutor understands the student and responds to the student's special needs. From its beginnings, the computer has been viewed as capable of providing such instruction, thereby having the potential to improve the quality of education. Of particular importance is the improvement of training in the mathematics and science topics that are requisite for entrance to the scientific community and to the high-technology world.

There are now over 10,000 pieces of educational software available. Almost all of this software can be classified as computer-assisted instruction (CAI) in contrast to intelligent computer-assisted instruction (ICAI) or programs that simulate understanding of the domain they teach and that can respond specifically to the student's problem-solving strategies. A large fraction of CAI software is of low quality and accounts for much of the teacher disenchantment with the computer (3, 4).

There have been attempts to bring artificial intelligence techniques to bear in development of ICAI (2, 5), but this has been viewed as impractical and has been largely relegated to the research laboratory. One of the reasons was the high cost of ICAI. It was common to require a million-dollar machine to interact with one student, and often the response time of the machine was slow. A second reason was the large amount of time associated with creating educational software. It is thought to take 200 hours to create 1 hour's worth of conventional CAI, and the time associated with ICAI is thought to be an order of magnitude greater. Finally, there was no established paradigm for enabling students to acquire knowledge. Early ICAI efforts often were ill-focused attempts to interact intelligently with the student without any clear understanding of the impact of those interactions on learning.

These obstacles to past efforts at ICAI

The authors are on the staff of the Advanced Computer Tutoring Project, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.