Overview of the NAS/NRC Conference on Large Data Bases†

MARTHA E. WILLIAMS

Information Retrieval Research Laboratory, University of Illinois, Urbana, Illinois 61801

Received December 9, 1974

In 1971 the Committee on Chemical Information of the National Academy of Sciences, National Research Council, recognized the problem of large data bases and accordingly created a subcommittee to investigate the problem and its impact on chemical information.

The first task of the Large Data Base Subcommittee was to survey organizations that generate and/or process large data bases. Giering[1] has indicated several different ways in which one can view a data base as being large. It can be large in terms of having a large number of entries or records (or bibliographic references). It can be large in the sense of having a large number of fields or data elements, which implies a degree of complexity. It can be thought of as large in the sense of having a large number of searchable or selectable elements, and it also can be large in the most readily understandable sense of having a large number of characters in storage.

The survey was restricted to files containing at least 100 million characters, and, in fact, most of them exceeded several billion characters. It was not restricted to data bases of potential use only to the chemical and/or scientific community, because the subcommittee's interests lay in the more general problems of formatting, reorganizing, storing, maintaining, updating, searching, and otherwise manipulating large data bases. Nonchemical files, then, were included in the hope that cross-fertilization or spin-offs from other disciplines might provide new ideas for handling large files in chemistry.

While principles may be general, in specific applications a variety of factors affect each other in relation to data base handling. The following are some of the factors whose interrelations, influences, and interdependences may bear significantly on the problem of handling large data bases.

backup files
batch search
compression
cost
file structure
fool-proof systems
format
frequency of file activity
(*i.e.*, additions,
purges, corrections)
frequency of request
hardware
inconsistent coding or
indexing for various
portions of the file
input/output
operating systems
organization
personnel
search strategies
size
software
subject matter
time

After surveying the major producers of large data bases to gain insights regarding the problem of largeness as it affects various aspects of information handling, members of the subcommittee concluded that there would be significant benefit to bringing together several large data base experts to share their views, problems, and solutions with others in the field. Accordingly, the Large Data Base Conference was held at the National Academy of Sciences on May 22–23, 1974. The conference committee consisted of M.E. Williams, Chairman; H. Skolnik; R.J. Rowlett; and C.M. Bowman.

Basically the concern of the survey and the concern of the Large Data Base Conference was to determine the nature of the problem(s) of large data bases. In what way is "largeness" a problem? Is it because storage devices are too small to encompass the entire file on any one device? Is it because certain files do not lend themselves to inversion on any reasonable basis? Is it because indexed files are inconsistent—having been indexed in several different ways over a number of years? Is it because a large number of variables are required for entering a complex search question? Is it related to the hierarchical structure of files? Is it because the format or data elements in the file have changed over the years? These are a few of the problem areas. Answers to, or views on, these questions were given by speakers at the conference and their papers appear in this issue of the *Journal of Chemical Information and Computer Sciences.*

LARGE FILES ON-LINE

The on-line bibliographic retrieval systems have received considerable attention in 1974, and, more importantly, they have grown significantly in the number of users accessing the systems and in the number of searches being conducted. There are a number of reasons why the on-line services have achieved the success they have in the past year. One of the most important factors is the increased availability of large data bases for retrospective searching on-line. Several years ago retrospective searching of large bibliographic files was, in effect, not possible because the files did not exist. Most of the major data base producers did not begin generating machine-readable data bases for public use until the late 1960's, and, when they began to do so, they did not convert their back files but began with the current year. Thus, if one wanted to do a five-year retrospective search of *Chemical Abstracts* (CA) in 1970, there would have been only two years of material available in machine-readable form, and the balance would have to have been searched manually. Now, there are six years of CA references available in machine-readable form. A similar situation exists with most of the other major data base producers, and, as with CA, it is now feasible to conduct retrospective searches of significant collections by computer because the files exist in computer-readable form. Another factor that has contributed to the success of on-line bibliographic searching is that the on-line services are now being marketed more extensively; thus more people are being exposed to them, becoming aware of the potential benefits, testing them and eventually purchasing the services. Other factors include: the great decrease in the cost of storage in

† Sponsored by the National Academy of Sciences, National Research Council Committee on Chemical Information at the National Academy of Sciences, Washington, D.C., May 22–23, 1974.

the past several years; the decrease in transmission/communication costs, especially through networks such as TYMNET where the communication cost per hour of on-line searching is typically about $10.00 per hour; the increased reliability of computer operations in the eyes of the general public; the increased awareness of potential users of the capabilities of computers; and the increased awareness of users of data base services.

On-line bibliographic searching is no longer an experimental phenomenon with completely subsidized services; it is now a commercial reality. The large files that are so expensive to search in batch mode when the local demand is low can be searched much more economically when large numbers of users throughout the country and beyond share the same resource and contribute to the operating costs. Based on discussions with a number of operators of on-line search services, one can estimate that in 1974 more than 700,000 on-line retrospective searches were conducted and that the number will probably increase to a million in 1975. These numbers are exclusive of library automation activities such as those at OCLC (Ohio College Library Center).

Among the largest files used for bibliographic or natural language searching are those maintained on-line for interactive searching via the systems at Lockheed, Systems Development Corporation (SDC), Mead Technology Laboratories, and the National Library of Medicine (NLM). These files are large by virtue of the fact that multiple data bases must be up simultaneously for on-line searching.

At the time of the Large Data Base Conference, SDC had 3.5 million records in its disk files representing 3.3 billion bytes of storage and, similarly, Lockheed (Summit[2]) had more than 3 million bibliographic records accessible on-line via its DIALOG system, representing more than 3 billion bytes. SDC's system (ORBIT) is related to the NLM system in that SDC developed the software that is used in NLM's MEDLINE and TOXLINE systems.

NLM's TOXLINE file contains bibliographic references to the toxic properties of chemicals, drugs, and pollutants and is growing at a rate of 100,000 records per year. The TOXLINE system is now operated at NLM using the ELHILL software (a variation of the SDC ORBIT programs) which provides free-text information retrieval. Prior to March 1974, TOXLINE was operated under the NASA RECON software which was developed by Lockheed. A comparison of the file structures, commands, and response time for the two software systems was made by Hummel.[3] The comparisons are made in light of the requirements of TOXLINE users.

PROBLEMS OF LARGE FILES

One of the problems associated with large files as well as small files is the problem of maintaining data in a secure manner. Orceyre[4] discussed six things that can happen to data. Data can be disclosed, modified, or destroyed either accidentally or intentionally. Another problem is the adequacy of character sets to represent the data symbols fundamental to individual data bases. Rule 5 pointed out that in the fields of chemistry and biology approximately 1000 different symbols are needed. However, each discipline's 1000 symbols do not exactly overlap.

The field of mathematics requires about 800 symbols and other disciplines require equally large sets. Though chemistry requires 1000 characters, 94.5% of the total occurrences of characters can be represented by 36 basic symbols (disregarding variations of the same symbols such as upper and lower case, superscripts, and subscripts, etc.), 64 basic symbols account for 99.59% of the occurrences, and 88 account for 99.95%. Related are the problems of standardization for character sets; compatibility, at the bit representation level, of the same character in different character sets; collating sequences; and the compatibility between input device conventions, internal storage representations, and output conventions.

While the representation of individual characters is a problem, the unambiguous representation of individual chemicals is a more difficult problem. Chemicals can be named in a variety of ways, all of which are correct, and their structures can be represented in various notation schemes for machine handling (Vasta[6]).

Data in data bases are represented in terms of characters; characters of data are input to files for machine manipulation and searching or output from files as a result of processing. Assuming a keying rate of 50 words per minute, the man-effort required to input 100 million characters would be two man years. Snyder and Skolnik[7] discussed this and other significant problems associated with input and output of data. We are all familiar with the great speed with which computers can manipulate data, but the bottlenecks with respect to keyed input and production of hard copy output remain.

DATA BASE MANAGEMENT

One problem associated with the generation and handling of large data bases is the operational management problem. Ideally, one should be able to handle the input and processing functions for several products, or outputs, in a manner that permits sharing of the processing activities and some records or portions of records (data elements, etc.). This requires, however, the establishment and employment of standard data element definitions and file structures.

Among the data base producers that have streamlined their operations for generating multiple products from an integrated system of files are Chemical Abstracts Services (CAS) and the Institute for Scientific Information. CAS (Wigington[8]) has designed an integrated processing system for producing and maintaining a number of hard copy products and data base products. The overall activity, both chemical information and basic data processing, involves approximately 200 master files and many transient files maintained on tape and direct access media. The largest of the files contains 1 billion bytes, and several others exceed 100 million bytes. Several years ago, at the Institute for Scientific Information (ISI), 15 separate products were produced from three separate product lines. These have now been amalgamated into one master file and a single processing stream, thereby effecting significant economies and increasing efficiency (Weinstock[9]).

A data base management system at the University of Illinois's IRIS (Illinois Resource Information System), which stores data representing attributes of land parcels, includes a clever monitoring scheme that checks user and system use parameters after each search and, based on use features, automatically triggers a batch job for changing the hash coding schemes and restructuring the files for improved efficiency (Alsberg[10]). This dynamic restructuring of files contributes significantly to reducing the cost of future searches.

COST

Large data bases are expensive to acquire whether through lease, license, or purchase. Because of their size they are expensive to process, i.e., reformat, invert, etc., in preparation for searching. They are expensive to store, maintain, and update, and they are expensive to search. And, as Cuadra[11] has said, the hardest part of operating a large data base searching service is learning how to make ends meet when you are operating in a nonsubsidized commercial environment.

A EUSIDIC working group[12] has looked at the economic factors associated with processing large files and developed some interesting ballpark estimates (converted here from British pounds to U.S. dollars). The factors looked at are: (a) the value of the stored information, which may include cost for generating the information, maintenance and storage, and the subjective value to the user; (b) manual-intellectual file processing, which is estimated at $625-megabyte of input; (c) file conversion, which is estimated at $25–50/megabyte for inversion plus $125–250/megabyte for compression, clustering, or tree structuring (if used); (d) storage, which is estimated at $1,250/megabyte/year for on-line storage (100 times the cost of magnetic tape storage); (e) file search, which is estimated at $2.50–10.00/megabyte for a sequential search for one question or $0.50/megabyte if the file is inverted;* (f) communication cost, which is approximately one-third the cost of a query; and (g) overall investment, which is directly proportional to file size.

RESEARCH

Efficiency and economy in relation to production are not the only advantages that accrue to the generation of large integrated files. Such files provide capability for generating additional subset products and spin-offs. They also permit analyses and experimentation with statistical validity of a very high order. This is possible because of the sheer size of the file. In other words, the large size of the file makes it possible to do studies that could not have been done with the same degree of confidence if the files were small. Some of the types of studies that benefit from size are those related to science policy, journal evaluation, research evaluation, vocabulary analyses, clustering, and automatic indexing (Weinstock[9]). When working with small data bases one can make only intuitive judgements for developing algorithms. With large data sets, however, one can develop reliable algorithms, assuming that data of sufficient frequency and with high statistical validity are available. That is, today, the characteristics of the data contents do not become evident until the file reaches a certain size, and statistical validity does not obtain until the file reaches an even larger size.

SOLUTIONS

Among the approaches that have been taken for solving the problem of retrosearching of large files are file subsetting, clustering, generation of search key surrogates (Lefkowitz[13]), coding (Kilgour[14]), and compression (Heaps[15]). The IRIS (Alsberg[10]) employs compression of data associated with land parcels. A compression ratio of 4:1 reduces both response time and storage costs by 4:1. After compression (hash coding), the current file includes 25 million bytes of compressed data. The IRIS system can accomodate more than a billion bytes of compressed data. Data compression as applied to bibliographic data bases functions somewhat differently than in the case of numerical data bases. This is true because the use of natural language data makes it impossible to take advantage of certain relationships that may occur between various sets of numerical data. It is true also because of the frequency-of-occurrence characteristics that have been observed in bibliographic data bases; for instance, and in accordance with the Zipf law, 50% of all the different terms in a data base occur with a frequency of one. In analyzing the *Chemical Titles* (CT) and CAIN data bases, Heaps[15] found that the 127 most frequent terms in CT accounted for 44% of all occurrences of title words. He described a scheme for assigning 8-bit codes to only the 127 most frequent terms, thereby achieving a 63% compression ratio. Heaps also described a scheme for coding text by use of equifrequent variable length test fragments rather than words.

Another solution to the problem of large files is computer systems or networks with massive storage capability. This is the tack taken by the Lawrence Livermore Laboratories where they use both data cells and a terabit (10^{12} bits) IBM photodigital store in a network of computers called OCTOPUS. Although the amount of storage sounds big, they have found that "Even a trillion bits eventually runs out, particularily when utilized by over a thousand users, seven days a week, for five years through the action of programs executing at over one million instructions per second" (Fletcher[16]).

Solutions to the problem of large data bases have been sought in many ways such as the use of compression, clustering, subsetting, and mass storage devices. However, even when these techniques and devices are used, the searching of large files still is costly. The need for sharing of resources and network use of data bases is definitely indicated. There still remain many other related problems which are less susceptible to ready solution because the answers will not come from technology. The really difficult problems are the psychological, sociological, political, and legal problems associated with shared use of files or exchange and transfer of information.[17]

* A sequential file search scheme has recently been developed within the Information Retrieval Research Laboratory at the University of Illinois which significantly lowers time and cost. Experimental results indicate a sequential search of 100,000 records in less than 4 seconds.

LITERATURE CITED

(1) Giering, R., "Search Strategies and User Interface," *J. Chem. Inf. Comput. Sci.*, **15**, 6 (1975).

(2) Summit, R. K., "Lockheed Experience in Processing Large Data Bases for Its Commercial Information Retrieval Service," *J. Chem. Inf. Comput. Sci.*, **15**, 40 (1975).

(3) Hummel, D., "A Comparative Report on an On-Line Retrieval Service Employing Two Distinct Software Systems," *J. Chem. Inf. Comput. Sci.*, **15**, 24 (1975).

(4) Orceyre, M., "Data Security," *J. Chem. Inf. Comput. Sci.*, **15**, 11 (1975).

(5) Rule, D. F., "Character Sets," *J. Chem. Inf. Comput. Sci.*, **15**, 31 (1975).

(6) Vasta, B. M., *J. Chem. Inf. Comput. Sci.*, to be published.

(7) Skolnik, H., and Snyder, J., "Input/Output Considerations for Large Data Bases," *J. Chem. Inf. Comput. Sci.*, **15**, 28 (1975).

(8) Huffenberger, M. A., and Wigington, R. L., "Chemical Abstracts Service Approach to Management of Large Data Bases," *J. Chem. Inf. Comput. Sci.*, **15**, 43 (1975).

(9) Weinstock, M., *J. Chem. Inf. Comput. Sci.*, to be published.

(10) Alsberg, P. A., "The Management of a Large Data Base in IRIS," *J. Chem. Inf., Comput. Sci.*, **15**, 23 (1975).

(11) Cuadra, C. A., "SDC Experiences with Large Data Bases," *J. Chem. Inf. Comput. Sci.*, **15**, 48 (1975).

(12) European Association of Scientific Information Dissemination Centres. Working Group/D. Retrospective Search Systems. "Preliminary Report on Retrospective Search Systems," NEWSIDIC, No. 7 (Jan 1973), pp 9–17.

(13) Lefkowitz, D., "The Large Data Base File Structure Dilemma," *J. Chem. Inf., Comput. Sci.*, **15**, 14 (1975).

(14) Kilgour, F., presentation at Conference

(15) Heaps, H. S., "Data Compression of Large Document Data Bases," *J. Chem. Inf. Comput. Sci.*, **15**, 32 (1975).

(16) Fletcher, J., "Large Data Base at the Lawrence Livermore Laboratory," *J. Chem. Inf. Comput. Sci.*, **15**, 19 (1975).

(17) Williams, M. E., "Use of Machine-Readable Data Bases," *Annu. Rev. Inf. Sci. Technol.*, **9**, 221–284 (1974).