MARTHA E. WILLIAMS
Director
Information Retrieval Research Laboratory
Coordinated Science Laboratory
University of Illinois
at Urbana-Champaign

Future Directions for Machine-Readable Data Bases and Their Use

Prior to discussing my views on the future directions of machine-readable data bases and their use, it is appropriate to indicate the point of departure. The history of the use of machine-readable data bases by the public commenced in the late 1960s and has progressed from a small-scale batch-searching activity, where services were largely restricted to SDI and operators were delighted if a system could be made to be self-supporting, to the current large-scale on-line retrospective and SDI service, where individual organizations are not only "for profit" but are making profits and operating with budgets in the tens of millions of dollars per year.

Data Bases and Their On-line Use

The importance of data bases within this information-oriented society can be measured in terms of their number, size, diversity, and volume of use. Prior to 1970 there were not more than a few dozen publicly available data bases, and combined they contained fewer than 30 million records.[1] Based on data collected for the directory *Computer-Readable Bibliographic Data Bases* and its updates (which contain data for the years 1975 and 1977), there was an increase in the number of bibliographic and natural language data bases, from 301 in 1975 to 362 in 1977, and an increase in the number of records contained in them from 50 million to 71 million. Of even more significance is the fact that the number of on-line searches of those data bases doubled between 1975 and 1977, from 1 million to 2 million.[2] The 1978 data indicate that there are 528 bibliographic and natural language data bases. In 1978 there were 2.67 million on-line searches. Also during 1978, a number of new data bases were brought up

on-line. BRS brought up an additional seven, SDC fourteen, and Lockheed twenty-one data bases.[3] The size, number and diversity of data bases are increasing, the use of data bases is increasing, and their use by new and different types of clientele is increasing. And the biggest increase in use will take place when end-users themselves are able to do a significant portion of the on-line searching without the aid of intermediaries.

Problems Due to Variety

Despite all the optimism, there are problems that result from the tremendous variety and variability that one finds in data bases. Data bases vary with respect to content: the subjects are different from data base to data base. There is a tremendous range of subject material in the more than 500 bibliographic and natural language data bases. Data bases vary with respect to format: each data base producer has his own format and very few of them conform to a standard. Data bases vary with respect to chronological coverage: some are less than a year old and some have been in existence more than ten years. Data bases vary with respect to relationships that may exist between them. For example, one data base may contain access keys which link it to one or more other data bases. The CASIA data base of Chemical Abstracts Service, for example, provides links to the CACON data base. One is an index, the other contains citations, and they are linked by CAS numbers. (These two data bases have just recently been combined into a new data base called CA Search.) Among the data bases which contain ties to one another are some of the Predicasts Inc. files and some of the BioScience Information Service files. Data bases vary with respect to vocabularies: some have controlled or semicontrolled vocabularies and most include free language terms. Titles are available for searching in almost all bibliographic data bases.

Data bases vary with respect to the systems used for searching them. There are many different systems available for on-line and batch-searching of data bases. There are also many different services offered through various on-line and batch systems. All of the on-line systems differ from each other with respect to access protocols, command languages, system responses and messages, system features, and even data element labels or tags.

Data bases vary with respect to the way they are loaded in different systems. Different on-line vendors will load the same data base in different ways. Lockheed, SDC, and BRS, for example, do not mount the same data base in the same way. One vendor may combine corporate information terms with subject terms; others will keep them separate. One may combine geographic location information with subject information; others

will not. As a result, one cannot search the same data base in exactly the same way in two different systems and get the same results. Data bases vary with respect to the features and functions found in the different systems, and in the techniques one can use for searching them; and they vary regarding output.

A user, or the user's representative (intermediary), has to contend with and accommodate this variability in data bases and systems in doing on-line searching. This leads to a problem. As the number and variety of data bases and systems increase, user confusion increases and the need for intermediaries trained to cope with the variables increases.

The Problem Regarding Retrieval Steps

Overall, there are various levels of retrieval goals. Initially, there is a need to retrieve source information; in other words, to determine what data base has the information the user wants. Following that is the need to retrieve the information or data itself. At a somewhat higher level is the desire to retrieve facts. And even higher than that is the goal to retrieve knowledge, and eventually to eliminate uncertainty.

Parallel with these retrieval goals are several retrieval steps. First it is necessary to identify the source locator or directory that contains pointers to secondary sources. Then it is necessary to identify and locate the secondary system that contains the required information, for example, BIOSIS Previews, CA Search, or COMPENDEX. Following this step, the secondary system must be queried. Finally, it is necessary to locate the primary source, obtain or access the primary document, read the appropriate portions, and assimilate the facts and data needed to satisfy the original purposes of the search and retrieval operation. These steps are all discrete, and most investigators (end-users) find it difficult to carry out all the steps without seeking outside help. However, if information retrieval is ever going to become really widespread, it will be necessary for end-users to do their own searching. And if end-users are to do their own searching, the discreteness of all of these steps must become much less apparent. In other words, what is really needed is a transparent information system, a means of reducing the discreteness of the retrieval steps so that the user can proceed directly from entering a query to the end-point of the search, retrieving the desired information, facts, and data from primary documents, without going through all of the intervening steps.

The Problem Regarding Multiplicity of Names

Another area of confusion for end-users, or even for their intermediaries, is that of distinguishing the various names for the component systems

and entities involved in on-line searching. Searchers should be able to recognize the distinctions among various communications networks; parent organizations that may have an on-line search system; information service organizations; the names of the services provided, the software packages, the computer operating systems, and the data bases; and the coded names of the data bases within specific systems (different vendors assign different names to the same data base). The name of an on-line vendor organization is *not* identical with all of these entities. The multiplicity of entities involved in on-line systems and the multiplicity of names for those entities contributes to the confusion in using on-line systems.

The Problems of Subject Access

One of the biggest problem areas of on-line searching is that of subject access. Anyone who produces data bases or is knowledgeable in their use knows what these problems are. The use of both controlled and uncontrolled terms in most of the data bases requires that the user employ both in his query. In addition, within any given data base, vocabularies may change every four to five years or less. Thus, a user searching several years of a data base must know how the vocabulary has changed over the years. The terms one would use today to describe a concept may not retrieve the relevant items from data base issues of ten years ago.

Another subject access problem is that of homography; words that are spelled alike but have different meanings in different contexts will retrieve unwanted items. There is also the problem of synonymy. A user must specify all of the synonyms appropriate for a particular term or concept that might have been used by an author in order to retrieve all the items that relate to the concept. If items in a data base have been well indexed using a controlled vocabulary, the problem of synonym specification can be greatly reduced. Unfortunately, most data bases have natural language titles and abstracts that do not contain highly controlled terms.

Another aspect of the subject access problem is the problem of chemical nomenclature. This is a very significant one because of the large number of chemicals—approximately 5 to 7 million. An individual chemical can be named in many ways, all of which are legitimate and correct. For example, in the course of analyzing chemical data bases we identified 27 different types of nomenclature schemes used in 165 machine-readable chemical data bases.[4] This means that there are *at least* twenty-seven different ways in which a given chemical can be identified. In addition to the controlled ways of naming chemicals, there are many kinds of trivial names or company-assigned names given to chemicals and chemical products to further expand the nomenclature problem. An indication of the extent of the problem can be seen when looking for synonyms in the

Chemline data base of the National Library of Medicine; as many as 110 different names have been found for the same chemical entity.

Another subject access problem results from the fact that terms are used differently in different data bases. A term having the same meaning in different data bases will be assigned different values. For example, a term such as *acid* would occur tens of thousands of times annually in CA Search, and only a few thousand times in BIOSIS Previews. The same term might occur fewer than a hundred times in COMPENDEX. It is obvious, then, that if a question containing the term *acid* is asked of all three data bases, the term cannot be used the same way in each. In one case the query will retrieve too much material, and in another case it may not retrieve enough. In one case the term could stand alone, thereby retrieving every item that contained it. In the case where it is a high-frequency term, it would need to be used in conjunction with other terms to reduce the number of "hits."

Yet another problem related to subject access is that of subject codes. These are found in various data bases and are often data base specific. Thus, a data base-specific code used in one data base certainly cannot be used as a search term in another data base, because the code does not occur in that data base. Consequently, if a question is to be run against multiple data bases, the search terms and strategy must vary to obtain optimal results. Subject codes that do not exist in a data base can be used as search terms with the net effect of wasted machine time and money.

Standards in the On-Line World

It seems that the way to solve the problem of variablity in the on-line data base world would be standardization. Standardization of on-line retrieval services would involve many different components, including data bases, subject access or analysis, the recording media, the systems, command languages, software, communications systems, and hardware. Moreover, each of these components would have to be analyzed at the subcomponent level.

Standards associated with data bases would need to deal with the following:

1. *content*—identifying which data base elements should be included in a particular search service;
2. *data representation* for each element or item within a record—indicating the character code and character set used;
3. the *form* of terms contained in a record—indicating whether they are abbreviated, coded, or fully spelled out;
4. *format*—indicating such things as the spacing and sequencing of data elements within a record;

5. *representation on a recording medium*—including the physical characteristics, physical format, and logical format; and
6. *subject access or subject analysis*—covering such things as classification schemes, natural language (which is obviously not standardized and never will be), key words, etc.

Some data bases have index terms; some do not. Some have controlled terms; some have semicontrolled ones. These all differ from data base to data base and would need to be standardized, if standards are to be used as the solution to the problems we are facing.

There would need to be standards regarding the systems for searching, file-loading techniques, file names, data element identifiers, and system vocabularies. These all differ from system to system. There would have to be standards for software—the software for search and retrieval, command languages, system features, protocols, and system techniques. Both responses and messages from systems would have to be standardized. Communications systems would need to be standardized—including access procedures and protocols, passwords, and system designations. These differ from one communications network to another.

Even if it were possible to develop standards for all the subcomponents mentioned, they would take a long time to develop and once achieved, their implementation is usually voluntary; thus, it is unlikely that development and implementation of standards for all of these items will ever be accomplished to simplify the problems of on-line retrieval. This does not mean that standardization efforts should not be continued. Certainly, wherever standards can be achieved, problems can be alleviated to a certain extent. The goal of achieving standards in all of the necessary areas, however, is unrealistic.

Alternative to Standards—A Transparent System

Since it is unlikely that all of the necessary standards will be developed and since it is even less likely that, if developed, they would be implemented, an alternative to standardization is needed. Such an alternative could be the development of a "transparent system." The likelihood of a single transparent system is remote, because there are too many organizations with vested interests; that is, too many organizations have invested sizable sums of money in developing systems, data bases, etc., and the chances of their changing without a demonstrable economic benefit are unlikely. The possibility of a distributed, integrated transparent system, however, is not unreal.

We at the Information Retrieval Research Laboratory of the University of Illinois are currently conducting a research project, with National

Science Foundation funding, that involves designing a transparent information system. We are determining what the components of a transparent system should be. We are also determining who is doing research and development on various elements that could be involved in such a system. We are considering alternative system architectures by weighing the alternatives of centralization vs. distribution of the various components in light of economic considerations, update requirements, etc.

A transparent system involves a variety of types of users, computers, terminals, operating networks, software, communications networks, and data bases. The various classes of users should include schoolchildren to sophisticated researchers. We must also consider the various classes of computers—maxi, mini, and micro—for various system components; the use of dumb or intelligent terminals for input and output; the operating network; software requirements; communications networks; and data bases of all types, whether they contain references, numeric data, full text or facts. We are also including derivative data bases; that is, data bases that contain descriptive information about other data bases—data bases that include term frequencies and word patterns, etc., or data bases that include quality indicators or value judgments associated with items contained within a data base.

A transparent information system would require directories, and a directory of directories in order to send a user to the appropriate subject-area directory or to gain information about the various files in a particular subject area. A transparent system would contain applications programs of various types—not just search and retrieval software, but statistical packages, modeling packages, and various other kinds of programs needed to manipulate data found within data bases. And it would contain a variety of "transparency aids." These transparency aids consist of converters, selectors, evaluators, analyzers, and routers. In order to provide insight to transparency aids, I will briefly discuss the first two classes of aids—converters and selectors.

Converters

Converters are needed in many areas. They are needed, for example, for access protocols. Currently there are a variety of types of access protocols to gain entrance into various networks and to send a user to the appropriate on-line service. Access protocol converters are needed to convert system A's protocol to system B's and vice versa, or for converting both A and B to a common protocol or standard. This does not mean that system A, which might be Tymshare, and system B, which might be Telenet, would have to make internal changes. Someone outside of those systems

could convert both of them to a standard. Thus, neither A nor B would lose the investment they made in their existing system.

Converters are also needed for the *language* of the access protocols; they can be used to convert the native language of the system to a foreign language. This would enable a speaker of German to use an English language-based system in German instead of English. There is a need for command language converters, whether they convert langauge A to language B, vice versa, or both to a standard. And there is a need for converters for the *language* of the command language, again from native language to a foreign language.

Converters are needed for converting the controlled language of one data base to that of another and vice versa. Converters are needed to transform natural language terms in a data base to controlled language terms. There need to be converters for system responses and messages; again, to convert the system messages and responses of system A to those of system B, vice versa, or both to a standard. And again, native to foreign language conversions are needed for system responses and messages. This is not infeasible. Currently, work is being done at the National Bureau of Standards on their network access machine which actually does convert access protocols and even dials up the target system.[5] The native to foreign language conversion problem is being handled in several places right now. The Canada Institute for Scientific and Technical Information (CISTI) permits use of either English or French protocols to access the CAN/OLE system. Similar work is being done by SDC for the use of the ORBIT system in Canada and other French-speaking countries. The command language conversion problem is being worked on by a number of people.[6] MIT began work on a common command language five or six years ago.[7] The language, called CONIT, is operational on four different on-line systems. Euronet is also working on the problem of a common or standard command language for use within the DIANE (Direct Information Access Network for Europe) system over the Euronet communication network.[8] The problem of the language used for commands, system responses, and messages has also been addressed by Euronet in Europe, SDC in the United States, and CISTI in Canada.

The problem of converting a data base's controlled vocabulary to that of another data base is under study at Battelle Columbus Laboratories.[9] The problem of converting a natural language to a controlled language has been worked on by the Robot System. The more difficult problem of converting free native-language text to free foreign-language text is being worked on by the Commission of the European Communities.[10] They are developing an autotranslation system for interconversion of at least four languages. All of these research and development efforts will result in the development of converters that are transparency aids.

Selectors

Another class of transparency aid is that of selectors. Selectors are needed for a variety of purposes. A selector could select classes of data bases appropriate to user characteristics and to the user's query. As the number and types of data bases increase, automatic selectors are needed to help a user determine which data base to use for a particular query. Data base selectors can be based on and include data such as term frequency, relative frequency of terms within a data base, user-assigned values, growth rates of the vocabularies, and variant forms of terms. Work has been done on automatic data base selection at the University of Illinois, at BRS, and at SDC. The University of Illinois work was carried out within the Information Retrieval Research Laboratory of the Coordinated Science Laboratory, with National Science Foundation funding.[11] Our research commenced in 1977 and was intended to determine the feasibility of an automatic data base selector (DBS). The work has been completed and the feasibility proven.

The University of Illinois's DBS includes normalizers for several variables found within the data bases. Procedurally, in order to build a test model selector we used the inverted files for twenty data bases from Lockheed and SDC and merged them, keeping one record for every unique term found in any data base. Within each term record we recorded information about the data base in which it was found, the frequency with which it occurred, and an indication of the kind of term it was—a word from a title, a word from an abstract, a controlled term, or an uncontrolled term.

Two other organizations that have developed aids to data base selection are SDC and BRS. SDC has developed a Data Base Index (DBI) and BRS has developed the CROS data base. DBI is restricted to data bases at SDC, and CROS is restricted to data bases at BRS. SDC's DBI operates on a single term at a time and produces a sequential ranking indicating which data bases contain the term in question; no indication of distance ranking is given; that is, if a given term occurred 1000 times in data base 1, 980 times in data base 2, and only 5 times in data base 3, there is no way of knowing that the distance between data base 2 and 3 is so great that the latter should probably not be searched. If terms are combined, the resulting list of data bases indicates that the combined terms occurred in the same data base—though not necessarily in the same record. Also, when terms are combined in DBI, the resulting list of data bases is unranked, so the user will not know which data base is the most likely candidate for search.

Bibliographic Retrieval Services' CROS operates on single or combined terms, so long as they are contained in a single search statement. The list of data bases produced as a result of a CROS search is alphabetically arranged by BRS data base mnemonic, and the postings value for the

statement, including logical combinations, is indicated next to each mnemonic. The user can then select the data bases with the highest numbers of references for the search. The search must then be run against the data bases chosen by the user. A limitation of CROS is that the values provided are for the on-line portion of the data base only; it does not reflect the off-line backfiles that BRS has for a given data base.

DBI and CROS are both operational and usable on publicly available systems. The University of Illinois's DBS is a test system and so is not publicly available (although the algorithms are available as they were developed with public funds). DBS was designed to determine which data bases are most likely to provide references in answer to a user's query and to provide results in the form of a histogram. The resultant list of data bases is in ranked order and the distance between data bases is indicated by the histogram. Neither DBI nor CROS accounts for variable factors associated with data bases; thus, results are based on postings alone. DBS, on the other hand, utilizes a mathematical model that operates on the term records and takes into account the number of years' worth of a data base, relative frequency of the term within the data base, relative frequency of the term across data bases, and the value of a term type (title, abstract, controlled or uncontrolled) within a data base as indicated by the data base producer.

A data base selector, to be of the most value to a user, should include all data bases, whether they are on-line or batch, and it should factor in the variables that account for the different ways in which the same file may be mounted in different systems. Such a selector capability would probably have to exist outside of the on-line systems, as it is unlikely that an individual on-line system would wish to promote data bases that it does not offer.

A data base selector is only one type of selector that should be included in a total transparent retrieval system. Automatic selectors could be developed for selecting search service organizations (on-line vendors or batch-search operators), communications networks, command languages (if users have preferences), terms to be used in query expansion, applications packages for operating on retrieved data, and output formats.

Converters and selectors are two types of transparency aids that would be used in a transparent system. In addition, there would be automatic routers,[12] evaluators, and analyzers. I have discussed only converters and selectors to illustrate what a transparency aid is.

Conclusion

The purpose of this paper has been to explain the current status of data bases, to discuss some of the current limitations in the on-line use of

data bases, and to indicate future directions. In the future, we will certainly see the development of systems and components that will simplify the retrieval process. A variety of automated aids are being and will be developed to carry out many of the activities that are now done by search intermediaries. Many of the conversion, translation, selection, evaluation, and analysis activities which are carried out by searchers can be done or assisted through automation. I have referred to these as "transparency aids" and have described some of them in the context of a transparent system. Whether or not a total integrated, distributed transparent system will be developed is uncertain, but the development of many of the separate components is assured. Many of them have already been developed, others are in the research phase, and others are not yet on the drawing boards. Many changes are underway in this dynamic field, but if an integrated transparent system is developed, the changes will not be apparent to the user. What the user will see is a greatly improved and easy-to-use system.

REFERENCES

1. Williams, Martha E., et al. "Data Base Statistics for 1977" (Final report on NSF Grant No. SP 77-0986; Coordinated Science Laboratory Report No. T-76). Urbana-Champaign, University of Illinois, 1979.

2. Williams, Martha E., and Rouse, Sandra H., eds., comps. *Computer-Readable Bibliographic Data Bases: A Directory and Data Sourcebook.* Washington, D.C., American Society for Information Science, 1976. (Looseleaf updates published 1977 and 1978.)

3. Williams, Martha E., ed. *Computer-Readable Data Bases: A Directory and Data Sourcebook.* Washington, D.C., American Society for Information Science, 1979.

4. _____, and MacLaury, Keith. "Mapping of Chemical Data Bases Using a Relational Data Base Structure." *In* Ludena, E.V., et al. *Computers in Chemical Education and Research.* New York, Plenum, 1977, pp. 3-23; and Williams, Martha E., et al. "Data Base Mapping Model and Search Scheme to Facilitate Resource Sharing—Vol. 1. Mapping of Chemical Data Bases and Mapping of Data Base Elements Using a Relational Data Base Structure" (Final report on NSF Grant No. SIS 74-18558; Coordinated Science Laboratory Report No. T-56, Vol. 1). Urbana-Champaign, University of Illinois, March 1978. (PB 283 892/8G1)

5. Rosenthal, Robert. "A Review of Network Access Techniques with a Case Study: The Network Access Machine" (National Bureau of Standards Technical Note 917). Washington, D.C., USGPO, 1976. (PB 256 525/7G1)

6. Iljon, Ariane. "Scientific and Technical Data Bases in a Multilingual Society," *Online Review* 1:133-36, June 1977.

7. Marcus, R.S., and Reintjes, J.F. "Experiments and Analysis on a Computer Interface to an Information-Retrieval Network" (Report on NSF Grant No. IST-76-82117; Laboratory for Information Decision Systems Report No. LIDS-R-900). Cambridge, Mass., MIT Press, 1979.

8. Negus, A.E. *Study to Determine the Feasibility of a Standardised Command Set for EURONET: Final Report on a Study Carried Out for the Commission of the European Communities, DG XIII*. London, INSPEC, Oct. 1976.

9. Niehoff, Robert T., and Kwasny, Stan C. "The Role of Automated Subject Switching in a Distributed Information Network," *Online Review* 3:181-94, June 1979.

10. Rolling, L.N. "The Second Birth of Machine Translation, A Timely Event for Data Base Suppliers and Users" (Paper presented at the Seventh Cranfield International Conference on Mechanised Information Storage and Retrieval Systems). Cranfield, England, July 1979.

11. Williams, Martha E., and Preece, Scott E. "Data Base Selector for Network Use: A Feasibility Study." *In* Bernard M. Fry and Clayton A. Shepherd, comps. *Information Management in the 1980's: Proceedings of the ASIS Annual Meeting*. White Plains, N.Y., Knowledge Industry, 1977, vol. 14, C13-D6, fiche 10; and Williams, Martha E. "Automatic Database Selection and Overlap of Terms Among Major Databases" (Paper presented at the Seventh Cranfield International Conference on Mechanised Information Storage and Retrieval Systems). Cranfield, England, July 1979.

12. Hampel, V., et al. "An Integrated Information System for Energy Storage." Livermore, Calif., Lawrence Livermore Laboratory, 1978. (UCRL-80349); and Goldstein, Charles M., and Ford, William H. "The User-Cordial Interface," *Online Review* 2:269-75, Sept. 1978.