public policy areas; and (3) it can relate sets of data files to those of similar focus that have been described previously. If those functions are enough to justify its continued publication, then the newsletter can go on for some time in the future. However, it is important to raise the possibility that the print medium is now an out-dated mode of communication in the field of data reference. On-line systems at individual archives and projects whereby archives exchange data descriptor tapes certainly hold the promise of a much improved data reference system. Networking and other developments, such as improved cataloguing of machine-readable data files, also offer interesting possibilities.

It is hard to imagine that one system will dominate the field of data reference in the future, and therefore one important function that organizations like IASSIST can play is to recommend ways in which new and old reference systems can be effectively integrated in the future.

ON-LINE REFERENCE TOOLS FOR THE HARD SCIENCES

Gordon H. Wood
Canada Institute for Scientific and Technical Information
National Research Council of Canada
Ottawa, Ontario

I Introduction

It is one thing to know or suspect that collections of machine-readable numerical data pertinent to one's discipline or problem may exist; it is something else to find that data, gain access, and use it profitably in one's research. The purpose of this paper is to review briefly the methods presently used to generate and access numeric data bases relevant to the so-called "hard sciences." Special emphasis will be given to the areas of on-line data retrieval and manipulation--areas where the state of the art in the "hard sciences" is generally conceded to be ahead of that in the "soft sciences." The reader wishing an inventory and description of the many scientific/technical data bases that are available worldwide is referred to the references at the end of the paper.

For the sake of clarity, it is useful to define a few terms as they will be used in this paper.

A. (Scientific/Technical) Numeric Data Base

An ordered collection of numbers whose values:

1. correspond to various properties, parameters or attributes of elements, substances or systems;
2. are critically evaluated by experts prior to their being included in the data base.

B. Numeric Data Base System

A numeric data base system consists of one or more machine-readable scientific/technical numeric data bases as defined above plus:

1. programs for searching, retrieving and organizing the data according to user selected criteria and, usually,
2. programs to manipulate the data.

(In general the latter property is what sets a numeric data system apart from a simple handbook or compendium. For example, a search routine may retrieve data giving the co-ordinate positions of the atoms in a given crystal. A simple command permits the user to calculate the various interatomic distances and the angles between the bonds joining the atoms. Another command generates a two dimensional drawing of the crystal projected along any desired axis or plane.)

C. Numeric Data Base System

A numeric data base network consists of one or more interconnected data base systems to which access is gained from a variety of remote locations by appropriate communication links.

II Numeric Data Base Creation

Experience has shown that an essential factor in the long term acceptance and success of a numeric data base system is the existence of an associated data evaluation center--a place where the data relevant to a particular data base are processed, both initially and in a continuing sense, for inclusion in the file. Ideally such a center should be located in an active research laboratory environment and be assured of long-term, stable financial and human resources. In practice, data evaluation centers are often co-ordinated and partly funded by national bodies established for that purpose such as the National Standard Reference Data System in the U.S.A. and the Science Research Council in the U.K. A framework for international co-operation in provided in part by the Committee of Data for Science and Technology (CODATA) and sub-groups of major international scientific organizations such as the International Unions of Pure and Applied Chemistry and of Pure and Applied Physics.

The functions of a data center may be summarized as follows:

A. Search the world literature, both published and non-published.

B. Retrieve and index papers and reports within the area of interest.

C. Extract the numerical data.

D. Check and evaluate the data with respect to accuracy, overall quality of the work, consistency with previously published values, etc.

E. Cast the data into the desired format and merge into the file.

It is perhaps instructive to briefly consider the rationale behind some of the attributes and functions desired for a data center. Long-term support with respect to funding and personnel is important because new data are continuously being generated, technical advances tend to make old data obsolete (e.g. too imprecise, too inaccurate, or too narrow in scope) and key personnel, being subject to the vagaries of the human condition, never last forever.

Critical evaluation filters out inaccurate or poorly documented data. The distilled product of evaluation is a compact, reliable data base which is tractable to handle and provides a real benefit to the researcher who perhaps has neither the resources, time, nor inclination to search the vast open literature himself. Clearly a data base system is no better than the data on which it is founded. If researchers lack confidence in the quality and reliability of the data, they are not likely to use a data base system no matter how sophisticated or elegant the accompanying software may be.

Having a data center located in an active research environment helps to assure that the workers responsible for compiling and evaluating the data stay at the fore-front of their field with respect to theory and experiment. The credibility of the data base is not likely to exceed the scientific credibility of those who produce it.

III Dissemination of Data Bases and Data Base Systems

The information contained in numeric data bases is made available to the "hard science" community in three major modes or "packages" which roughly parallel the definitions given earlier. Naturally, variations and permutations of these modes are also possible.

A. Subscription to a Network

For a fee, which may consist of some combination of annual subscription, connect time and characters transmitted charges, a user connects appropriately to a node of a network and is given access to all of the data base systems for which he has paid. Most applications do not require a very "smart" terminal and the user needs to master only a few simple commands to operate on-line. Because of the small capital investment and the wide variety of information potentially available, this mode of dissemination is most likely to suit the worker who has a considerable range of research interests which tend to fluctuate both in intensity and focus. Usually a network will, of course, support batch and quasi "batch on-line" tasks as well.

Two such networks already functioning are the Chemical Information System in the United States, which has eleven data bases presently available with eight more under test [1], and the Direct Information

Access Network for Europe. A similar network on a smaller scale, which exists now in embryonic form, is being assembled in Canada by the National Research Council. Initially, five data bases are planned: three are crystallographic, covering the areas of metals, organics and inorganics; the fourth is thermochemical; the fifth, already functional, is a program for comparing an unknown infra red spectrum against a collection of some 100,000 or more spectra of known compounds.

B. Lease or purchase of a data base system

In this mode, a user obtains a tape copy of the complete data base and its relevant software for mounting on a nearby, typically institutional, computing system. Assuming the software is adequate for his needs and compatible with the local computer, the user need not have extensive computer expertise nor a "smart" terminal. Researchers benefiting from this means of dissemination would be those who tend to work almost exclusively in one discipline and for whom the storage and usage costs of a local computer would be more favourable than on-line charges. Update tapes would normally be supplied periodically by the data base supplier and the user would agree to distribute the tapes no further than his own institution.

C. Lease or purchase of all or part of a data base

With the proliferation of small yet powerful computers into many laboratories, the option of obtaining a tape copy of an entire data base, or a selected sub-set of it, for private use is becoming more popular. In this distribution mode, the user either requests an actual tape or copies what he needs on-line and proceeds locally form there using his own customized search, retrieval and manipulation routines. Such a user, of course, needs not only to support a mini-computer and its ancillary equipment but needs considerable programming expertise as well. Again, update tapes would be made available from the supplier and restrictions concerning extra institutional use would usually apply.

An example of a user for whom this option might be of value would be a lecturer in thermochemistry who leases the data covering a few hundred compounds of interest and generates an appropriate suite of programs. The pedagogical value is clear. His students are able to tackle practical problems, rather than artificially contrived ones, without tedious calculations causing them to lose sight of the chemical principles involved.

IV Data Bases of Numeric Data Bases

The media employed in cataloguing and marketing numeric data bases for the hard sciences are basically the same as those covering numeric data bases in general. First, there are compendia and directories [2,3,4] which cover a broad range of data bases but provide limited detailed information about any one in particular. Second, there are a number of publications [5,6,7], devoted to on-line non-bibliographical data base methods, technology and management, which periodically review the current world-wide inventory of numeric data bases. Third, there are diffuse or indirect sources which, while less systematic than those just mentioned, often provide the most useful information. In this category are found scientific conferences, journals of the various learned societies, and the very effective "grapevine" formed by scientists and information specialists.

V Conclusion

The future of on-line numeric data bases in the hard sciences appears very promising indeed. With computer technology improving so rapidly because of pressures from many sectors, it is likely the greatest impediment to the growth of numeric data bases as reference tools will be only the lack of imagination and interaction in the supplier/user component of the activity.

References

1. See, e.g.,. Heller S.R. & Milne G.W.A. American Laboratory 12, 33 (1980).
2. Computer-Readable Data Bases, A Directory and Data Source Book. Martha E. Williams, *et al*, eds. American Society for Information Science, (1979).
3. Directory of Online Data Bases. Ruth N. Landau, *et al*, eds. Cuadra Associates, Santa Monica (Issued semi-annually, quarterly updates.)
4. Eusidic Database Guide. Alex Tomberg, ed. Learned Information, New York (1978).
5. On-Line Review, Learned Information, New York. (Published quarterly.)
6. Online and Database (2 journals). J.K. Pemberton, ed. Online, Inc., Weston, Ct. (Published quarterly.)
7. CODATA Bulletin. CODATA Secretariat, Paris (Published irregularly.)