SCIENTIFIC NUMERIC DATABASES - A NEW RESEARCH TOOL

LES BASES DE DONNEES NUMERIQUES - UN NOUVEL INSTRUMENT DE RECHERCHE

Gordon H. Wood
National Research Council of Canada
Bldg M-55, Montreal Road
Ottawa, Ontario K1A 0S2

ABSTRACT

Scientific numeric databases (SND), consisting of a machine-readable collection of critically evaluated scientific numeric data and computer programs allowing one to search, retrieve and manipulate those data, are impressive new research tools. To make these emerging tools available in Canada, CISTI has initiated a new program, entitled Scientific Numeric Databases, which is described in some detail. Special emphasis is given to the SNDs held by CISTI and the national CISTI-based network on which they are being inplemented. Finally, an overview is given of the present situation in Canada with respect to the availability, generation and dissemination of this important information resource.

RESUME

Les bases de données numériques en sciences (BDNS), qui sont une collection de données numériques évaluées dans les domaines scientifiques et en logiciels permettant de chercher, repérer et manipuler ces données, représentent un nouvel outil de recherche fort impressionnant. Afin de rendre disponible ce nouvel instrument de recherche au Canada, l'ICIST vient d'établir un nouveau programme, intilulé Bases de données numériques, lequel est décrit sommairement. Une attention particulière est donnée aux BDM de l'ICIST et au réseau télématique de l'ICIST sur lequel elles sont implantées. Pour terminer, l'auteur donne une vue d'ensemble sur la disponibilité, la production et la diffusion de cet important réservoir d'information au Canada.

SCIENTIFIC NUMERIC DATABASES

I INTRODUCTION

The spectrometer, the oscilloscope, the electron microscope -- these are but a small sub-set of the large number of instruments that scientists have used for years in making measurements and generating data. In the last decade or so, computers in their various incarnations (main frame, mini and micro) have been employed to automate measurements and analyze data thereby greatly facilitating the generation of numeric data. On the other hand, when those same scientists have needed to obtain numeric data that may be available in the world's literature the computer was of little use and they have had to manually search sources such as handbooks, data compilations, journals and the "grapevine". The advent of the scientific numeric database promises to redress this situation by making evaluated numeric data as close as the keyboard of the nearest computer terminal.

After some basic terminology is defined the scope of the scientific numeric database activities at the Canada Institute for Scientific and Technical Information (CISTI) is outlined. Following this is a more detailed description of the online numeric database services offered by CISTI and a brief survey of some of the scientific numeric database activity in Canada external to CISTI and the National Research Council (NRC).

II DEFINITIONS

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

A. Scientific Numeric Database

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 database.

B. Scientific Numeric Database System

One or more scientific numeric databases as described 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 it is the latter property which makes a scientific numeric database system much more than the electronic equivalent of thumbing through a handbook or compendium. For example, a search routine in one database system at CISTI retrieves data giving the co-ordinate positions of all the atoms in a given crystal. A simple command subsequently permits the user to calulate the various interatomic distances and the angles between the bonds joining the atoms. Another series of commands readily generates a two-dimensional drawing of the crystal projected along any desired axis or plane.

C. Scientific Numeric Database Network

One or more interconnected database systems to which access is gained from a variety of remote locations by appropriate communication links. Given modern telecommunication capabilities, a network may be local, national, international or intercontinental in extent. The scientific numeric database online service being developed at CISTI is an example of a national network wherein several database systems reside on a computer in Ottawa and scientists across the country communicate via the DATAPAC packet switching system.

III COMPONENTS OF THE SCIENTIFIC NUMERIC DATABASE ACTIVITY AT CISTI

The provision of machine readable scientific numeric data is consistent with CISTI's mandate to provide scientific and technical information to Canadians. Emphasis will, of course, be given to those databases in the areas of the basic and applied natural sciences, such as physics, chemistry, biology, nmeterials, astronomy, energy and engineering, with which CISTI is concerned. For description purposes, it is convenient to subdivide this embryonic CISTI activity into four parts:

A. Access

The generation of scientific numeric databases is an expensive operation. As a result, the fees charged by organizations offering these databases to the scientific community must necessarily be substantial, often placing them financially out of reach of a private individual or a small laboratory. Thus, one of the roles played by CISTI is to ascertain which databases are needed in Canada and gain access to those databases on behalf of Canadians. In some situations it may require that CISTI pay the necessary subventions and acquire the database and all its updates and modifications for mounting in Canada; in other situations it may prove expedient for CISTI to negotiate direct access for Canadians and serve as a central co-ordinating agency between the database supplier and the Canadian users.

B. Referral

Closely related to the provision of access, the function of this component is to maintain a current world-wide scientific numeric database inventory so that Canadian scientists may determine the existence of files relevant to their needs. Depending on factors such as cost and breadth of interest in a given file, CISTI will either assist the user in gaining personal access to that file or attempt to acquire it for general Canadian use.

C. Generation

For very good scientific, technical and economic reasons it is important that Canada produce indigenous scientific numeric databases as well as import them from other countries. To this end CISTI endeavours to consult and co-operate with Canadian scientists interested in producing appropriate databases. Database production being an expensive procedure, as just mentioned, it is good stewardship of Canada's resources to minimize unnecessary duplication of effort and encourage the pooling of capabilities and interests.

One example of such an effort is the Metal Data File (Calvert, 1979) being produced at the National Research Council as a joint project of the Division of Chemistry and CISTI. Another example, though still in the planning stage, is a proposal whereby CISTI will assist a Canadian university professor who has an opportunity to co-produce a database with a group in the Federal Republic of Germany.

As this activity matures, it is planned that CISTI will not only co-operate in file production but prosecute original research into search systems and storage techniques.

D. Dissemination

Databases held by CISTI are made available to Canadians via three modes: online access, (discussed more fully in Section IV), sub-lease of database magnetic tapes and customized searches by NRC personnel. In the sub-lease mode, a user obtains from CISTI a copy of the desired parts of the database and, optionally, the searching software for mounting on a nearby, typically institutional, computer. 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 most from this means of dissemination would be those who tend to work almost exclusively in one discipline and for whom the leasing charges plus the storage and usage costs of a local computer would be more favourable than online charges. Update tapes are made available at incremental cost as CISTI acquires them and the user must sign

an agreement limiting the use of the tapes to his own institution.

Customized searches are of interest to the scientist who perhaps does not have access to a computer terminal or the inclination to invest the time needed to learn the protocols and procedures necessary to personally use a database system. In this case, a researcher may formulate his question in quite general terms and submit it to CISTI where, for a modest fee, someone skilled both in the scientific discipline and the use of the database will attempt to solve the problem.

To publicize the availibility of these databases it is planned to conduct demonstrations at relevant scientific conferences and hold workshops at strategic centers. The educational and practical background of most potential users, combined with didactic manuals, will obviate the need for extensive training in many of the database systems. Nonetheless, training will be available as needs dictate.

IV THE ONLINE SCIENTIFIC NUMERIC DATABASE SERVICE AT CISTI

It is anticipated that the online network will become the primary means of database dissemination. This section reviews the mechanics, the design philosophy of the system, its present status and plans for the future.

A. Technical Details

As mentioned earlier, the online database service is in the form of a national network allowing anyone in Canada with access to a computer terminal and a telephone to use the database system mounted on the NRC computer in Ottawa. To enter the network, a user need only dial his local DATAPAC node (the packet switching network operated by the Trans Canada Telephone System with nodes currently in 57 Canadian cities), indicate the speed at which his terminal can operate and give the address code of the NRC computer. Access to a given database system is then accomplished by executing a brief LOGON procedure and issuing an appropriate database selection command. After verifying that the user is authorized to use that specific database, the computer responds in the official language chosen by the user and prompts the user to begin the session.

Two levels of operation are possible. The new or casual user may choose an interactive approach in which he is benignly led "by the hand" and prompted to enter the required information; the expert user may choose to formulate his request in the succinct form of a command followed by a series of parameters.

B. Design Philosophy

Central to the design of the service is the commitment that the service should respond primarily to the needs of the scientist. Updates to the databases will therefore be added as quickly as practicable subject to the proviso that database quality is not sacrificed for mere quantity. Proposed enhancements and modifications will be assessed more on their scientific utility than their computer system elegance.

C. Present Status

Two database systems are currently mounted on the NRC Computer. SPIR (Search Program for Infrared Spectra) has been available from CISTI since 1978; CRYSTOR (Cambridge Crystallographic Database) is available in prototype form with full public release planned for the summer of 1981.

1. SPIR. The primary function of SPIR is to enable a user to search a collection of some 143,000 infrared spectra and attempt to locate the spectrum which most closely matches the spectrum of his unknown. After prompting the researcher for the data required to describe the unknown, the computer searches the database and prints out serial numbers and brief descriptions of the twenty spectra in the file most similar to that of the unknown. A figure of merit proportional to the closeness of the match is calculated for each spectrum found and the listing is ranked accordingly. To complete the identification the user must consult the spectral atlases indicated by the retrieved serial numbers and make a visual comparison of the known and unknown spectra.

2. CRYSTOR. Originating in the Crystallographic Data Center in Cambridge, England, CRYSTOR contains data on all organic and organometallic compounds whose crystal structures have been determined by x-ray or neutron diffraction studies since 1935. Growing at an annual rate of about 3000 new entries, the database currently covers about 26,000 distinct compounds.

For convenience in searching, the database is divided into three sub-files: Bibliographic (compound name, chemical formula, chemical class, literature reference, etc.), Connectivity (an alphanumeric representation of a two-dimensional diagram showing the manner in which the atoms of a given compound are bonded together) and Data (unit cell and symmetry information, atomic co-ordinates, bond lengths, accuracy indicators, etc.)

A typical query consists of posing suitable questions to the Bibliographic and/or Connectivity files. The user may stop at this point, obtaining a listing of the bibliographic entries corresponding to the "hits" resulting from his questions or he may carry on and retrieve the crystallographic structural data for those "hits" from the Data file for subsequent processing by two programs called GEOM78 and PLUTO78. GEOM78 acts on the

retrieved structural data to produce tables of fragment geometry and geometric parameters; PLUTO78 acts on the retrieved structural data to produce a wide variety of plots of molecular diagrams.

D. Future Additions

Generally, databases will be added to the online service in response to demonstrated needs as resources permit. The addition of two crystallographic files, complementary to CRYSTOR, is planned in the next two years. One, the Metal Data File mentioned in Section III, contains crystal structure data on about 5000 metals and metallic compounds. It currently is operational on a mini-computer at NRC but is not ready for public release. The other file, the Inorganic Crystal Structure Database being produced in the Federal Republic of Germany with Canadian participation, contains crystal structure data on about 6000 inorganic compounds.

Other possible additions include databases of mass spectrograms, powder diffraction patterns and unit crystal data. As will be seen in the next section, there already exists in Canada good accessibility to thermodynamic/thermochemical databases. It is unlikely therefore that databases in this field will be acquired by CISTI.

V SCIENTIFIC NUMERIC DATABASE SYSTEMS IN CANADA EXTERNAL TO NRC

Although the number of machine-readable collections of scientific and engineering data extant in Canada is relatively large, the number of functional scientific numeric database systems is not. An exhaustive survey being beyond the scope of this paper, Table 1 gives only a brief summary of five database systems known to be available in Canada. All are functional in a local sense; most are available online now or will be in the near future.

Rapid development in this area has been impeded by at least two factors. First, database development work is not universally regarded as a worthy scientific endeavour. Hence, only well established scientists can enjoy the luxury of pursuing the activity, often as a sideline, while younger workers are discouraged from becoming involved because such work is not perceived as a foundation on which a career can be built. Second, the field is sufficiently new and yet "unexciting" that support from the usual granting agencies is difficult to obtain. No doubt, as time passes and this activity matures and gains scientific stature, these hindrances will be overcome and the general level of Canadian input into scientific numeric databases will rise significantly.

VI SUMMARY

SCIENTIFIC NUMERIC DATABASES

Scientific numeric databases as research tools are indeed an idea "whose time has come". Inhibiting factors such as cost, availability and user apprehension concerning the quality and utility of databases are steadily being overcome. Costs for computing, mass storage and telecommunications, for example, are consistently decreasing whereas capability and efficiency of these functions are consistently increasing. There is worldwide, multi-discipline activity in database development (e.g. United States, United Kingdom, France, Federal Republic of Germany and Japan) and many of those databases will be available internationally by various leasing and licensing arrangements. Apprehensions about the utility of databases will tend to fade as fabourable experience accumulates and the degree to which the average scientist is computer-oriented increases. Competition will tend to eliminate databases of questionable quality as will, hopefully, a heightened international awareness of the importance and worth of properly evaluated data (see eg. Rossmassler, 1980).

TABLE 1

Field	Database Name	Function	Organization
Chemical Engineering	CIMPP (Comprehensive Industrial Materials Property Package)	Calculations & Simulations for chemical engineering problems	University of Western Ontario London, Ontario (Shewchuk, 1979)
Chemical Thermodynamics	FACT (Facility for Analysis of Chemical Thermodynamics)	Thermochemical calculations, emphasis on inorganic compounds	Thermfact Ltée 447 Berwick Ave. Mont-Royal, Que. (Pelton, 1978)
Electrical Power Engineering	PSSDB (Power System Simulation Database)	Study steady state power flows and dynamic response to disturbances	Manitoba Hydro Winnipeg, Manitoba (Mr. L.Q. Chow)
Thermochemical	TBANK (Thermochemical Data Bank)	Calculate thermochemical properties of substances and compounds	University of Toronto Toronto, Ontario (Alcock, 1979)
Soil Engineering	Digital Terrain Modelling & Analysis	Predict potential soil erosion in an area given soil and topographical characteristics	Collins & Moon Ltd. 435 Stone Road West Guelph, Ont. (Collins, 1981)

REFERENCES

ALCOCK, C.B. "Some Aspects of the Development of a Prototype International Databank" presented at Symposium on the Industrial Use of Thermochemical Data, University of Surrey, U.K., 11-13 September 1979.

CALVERT, L.D. and YANG, Y. "The Metal Data File at NRCC" presented at a Symposium of the American Chemical Society Meeting, Washington, D.C., 9-14 September 1979.

COLLINS, S.H. and Moon, G.C. "Algorithms for Dense Digital Terrain Models", in Photogrammetric Engineering and Remote Sensing, Vol 47, No. 1 (January 1981), pp. 71-76.

PELTON, A.D. et al. "FACT (Facility for the Analysis of Chemical Thermodynamic Data Treatment Centre" in Application of Phase Diagrams in Metallurgy and Ceramics, (NBS-SP 496) National Bureau of Standards, Washington, D.C., 1978, pp. 1077-89.

ROSSMASSLER, S.A. and WATSON, D.G. eds. Data Handling for Science and Technology, North Holland, 1980.

SHEWCHUK, C.F. "A Materials Properties Package for Engineering Applications", Report of SACDA (Systems Analysis, Control and Design Activity), University of Western Ontario, London, Canada, 1979.

SHEWCHUK, C.F. "A Practical User-Oriented Computer Program for Evaporator Analysis", in AIChE Symposium Series, Vol 200, No. 76, 1980, pp. 143-58.