Scientific numeric databases

Wood, Gordon H.

Publisher's version / Version de l'éditeur:
https://doi.org/10.4224/40000406
The application of new technologies to improve the delivery of aerospace and defence information, pp. 5-1-5-6, 1983-12

AGARD

ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT

7 RUE ANCELLE 92200 NEUILLY SUR SEINE FRANCE

Paper Reprinted from
Conference Proceedings No. 357

THE APPLICATION OF NEW TECHNOLOGIES TO IMPROVE THE DELIVERY OF AEROSPACE AND DEFENCE INFORMATION

NORTH ATLANTIC TREATY ORGANIZATION

Scientific Numeric Databases

Gordon H. Wood
Canada Institute for Scientific and Technical Information
National Research Council of Canada
Montreal Road
Ottawa, Canada
K1A 0S2

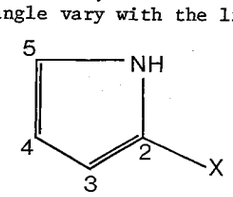
Summary

Scientific numeric databases (SND) are powerful, relatively new research tools for the scientific and technical community. This paper illustrates their use by some practical examples, describes the attributes and capabilities of such databases and gives a survey of the activity in this field. SND permit the direct location, retrieval and the subsequent analysis/manipulation of evaluated numeric data. Advances in telecommunications and increases in the number and types of SND produced greatly enhance the likelihood of relevant data being readily available. Although only 5% of all databases accessible online in North America and Europe are SND, they are growing in importance and acceptance as more databases are developed and scientists and engineers become aware of their potential. The National Research Council of Canada is active not only in the production of SND, involving some international collaboration, but also in their dissemination by means of a nation-wide online packet-switched network.

I Introduction

To appreciate the power and usefulness of scientific numeric databases (SND) as research tools for scientists and engineers it is helpful to consider some fairly typical questions which might be posed to scientific information personnel:

1. Data are needed on as many compounds as possible in which a pyrrole (C_4H_5N) ring appears. Any ligand (X) may be substituted for one of the hydrogen atoms attached to one of the carbon atoms 2 or 5. In particular, how does that carbon-nitrogen bond length and the carbon(2)-nitrogen-carbon(5) angle vary with the ligand X?



2. The analytical laboratory has an infrared spectrum of an unknown substance. To help identify that substance, is there any way they could compare its spectrum to the spectra of a large number of known compounds?

3. A measurement of the specific heat of gadolinium oxychloride at 720K has just been completed. How does that measured value compare with the best published values at corresponding temperatures?

Questions 2 and 3 could be answered by manually searching appropriate atlases of spectra and handbooks respectively, but answered more efficiently by the use of an SND. A scientific numeric database is the only practical means of tackling question 1. These questions will be considered in more detail in Section III where it will be shown how SND's can be used to solve these and related problems in a cost-effective manner.

Whereas most scientists and engineers are familiar with the computer as a means of performing calculations, automating measurements or searching large bibliographic databases, relatively few are familiar with the computer as a means of access to evaluated scientific data from the world's literature. By gaining some insight into the capabilities and benefits of SND's, information specialists can serve as catalysts in bringing about a profitable interaction between scientists, engineers and scientific numeric databases.

Section II defines some basic terminology and gives some perspective on SND's in the context of databases in general. Section III illustrates some of the capabilities and benefits of SND's. Section IV is a brief scan of some recent worldwide advances in the areas of telecommunications and database development. Section V, a survey of some of the SND activity in Canada, concludes the paper.

II Definitions

The classification scheme used in the "Directory of Online Databases"[1] is adapted here to describe the various types of databases in existence.

A. Reference/Source Databases

Reference databases are those which refer or point a user to another source, often a document, for more details or the complete text. This group may be further broken into two categories: Bibliographic (containing primarily citations to published information like journal articles, reports, patents, dissertations, conference proceedings and books) and Referral (containing primarily references to non-published information sources like organizations, individuals, audiovisual materials and non-print media).

Source databases, containing complete data or the full text of the original source information, are conveniently categorized as Numeric, Textual-Numeric and Full-Text. Numeric databases contain original and/or statistically manipulated representations of data; Textual-Numeric databases contain a mixture of numeric data and related textual information; Full-Text databases contain records of the complete text.

B. Scientific Numeric Databases

Primarily source databases, scientific numeric databases are 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.

Good scientific numeric databases are therefore much more than mere compilations of numbers. The important, and expensive, function of review and evaluation, which is not often found in reference databases, serves to make the data more reliable than those found in the open literature and more useful because of the rationalization of factors like uncertainty statements and units of measurement.

Some feeling for the relative abundance of scientific numeric databases may be gained from the following table (based on entries in reference 1):

Database Type	Science/Engineering		Other		Total	
	1982	1983	1982	1983	1982	1983
Source	67	86	630	977	697	1063
Reference	160	198	346	472	506	670
					1203	1733

Thus, in 1983, source databases in the science and engineering disciplines represent only about 5% of the total, down somewhat from 5.6% in 1982. Not all of these science or engineering source databases would strictly qualify as scientific numeric databases nor, of course, are databases listed that are under development but not publicly available.

C. Scientific Numeric Database Systems

To avoid confusion, the term scientific numeric database system should be used to describe a set of one or more scientific numeric databases combined with a suite of computer programs enabling the scientist or engineer to search the database(s), retrieve items of interest and manipulate those items in a variety of ways. Using a scientific numeric database system is therefore much more than electronically flipping through a handbook to find a specific entry as the next section will illustrate.

III Capabilities and Benefits

The utility and cost effectiveness of scientific numeric database systems (SNDS) can probably best be illustrated by considering the three questions posed in the Introduction.

To the best of the author's knowledge there is no practical way to solve question 1 using bibliographic reference databases or hardcopy reference tools apart from an incredibly exhaustive literature search and a great number of manual calculations. With the aid of an SNDS like the Cambridge Crystallographic Database[2], however, such a problem may be solved in one or two hours in a straightforward, systematic way. One need only describe the chemical connectivity of the pyrrole fragment and ask the system to check that connectivity against the connectivity of all the compounds in its collection. The output of that search, a listing of all the compounds containing a pyrrole fragment, may subsequently be operated upon by a built-in program to execute the required geometric analyses.

Question 2 could be solved in a brute-force way, of course, by manually searching through compilations and atlases of infrared spectra and looking for one that resembles that of the unknown at hand. An improvement on that would be a manual system such as, for example, the "Spec-Finder" marketed by the Sadtler Research Laboratories[3]. In this system, the spectrum is divided into 27 intervals and the strongest peak or band, if any, in each interval is coded. An index ordered by the strongest peak overall then points the user to the spectra in

that firm's spectra collections most closely matching the unknown. A further improvement is gained by the use of an SNDS like FIRST-1[4], SPIR[5], IRGO[6] or IRIS[3] all of which use a database of spectra compiled by th American Society for Testing and Materials. In these systems, the spectrum is coded in terms of its peaks, bands and no-band regions and entered into the computer to form an electronic "mask". This "mask" is then automatically compared with the large number of spectra (about 140,000) in the database and the user is presented with a list of the target compounds having spectra most closely resembling that of the unknown. In most cases the user must still consult a hardcopy of the known spectrum for detailed comparison. Nonetheless a considerable amount of time has been saved and the user is confident that the search has been as exhaustive as currently feasible.

Question 3 could be addressed by consulting handbooks of thermophysical data subject to the usual constraints of actually finding the most recent volume. Even if that constraint is overcome and a value for the compound of interest is found, the probability is quite high that the cited value will not be at the temperature of interest nor in the appropriate set of units. By invoking an SNDS like FACT[7], TBANK[8], or THERMODATA[9], for example, a user has immediate access to a fairly exhaustive set of the most recent data. More than that, the user will be able to ask the system to interpolate between the values of the specific heat at various temperatures to give an estimate at the temperature of interest in the appropriate energy units.

A. Capabilities

Against the background provided by this set of examples, it is now useful to sketch the range of functions that SNDS's can perform.

1. Retrieve items quickly, exhaustively and accurately from large collections of data. Retrieve along lines of thought for which compilers could not have foreseen the need for an index. Retrieve types of information too detailed and tedious for the human mind to readily handle (eg. the connectivity search described earlier).

2. Manipulate and analyze the data in a variety of ways, for instance:
   a) fit curves to quantify relationships
   b) interpolate or extrapolate to facilitate comparison of new measurements
   c) generate graphs of trends or make statistical comparions
   d) produce plots of molecular geometry

3. Simulate experiments with mathematical models, exploring processes like chemical reactions theoretically, thereby often obviating the need to perform the actual experiments or build prototype equipment.

4. Formulate new ideas from observations and statistical inferences on the data themselves. The Cambridge Crystallographic Database is a large body of reliable, basic data and authors have used it to gain information on the effects of substituents, on chemical reactivity, on molecular flexibility and intermolecular forces (see eg. 10, 11 & 12).

B. Benefits

In a paper given before this group in 1981, V. Hampel[13] addressed most of the economic advantages accruing from the use of an SND or SNDS. For the sake of completeness it is useful to reiterate a few of them here.

The direct savings in time of the scientist, engineer or information specialist are clear. Consider some examples:

1) the time and money expended in needlessly measuring some property of a substance which is already known
2) the effort involved in finding data of interest
3) the labour involved in re-formatting such data for further use or manipulation, not to mention the exposure to errors involved in keyboarding from hardcopy to magnetic form.

Other factors less easy to quantify but nonetheless of considerable value are the timeliness of the data (both new and corrected) and the assurance that the data are generally more reliable than those available in the open literature or in compilations.

In summary, SNDS can maximize the proportion of time spent by the scientist or engineer in creative activities and, in fact, may serve as powerful tools in those activities.

IV Recent Developments

An exhaustive review being beyond the scope of this paper, this section attempts only to scan some of the relevant advances in telecommunications and database development in order to give an indication of the state-of-the-art of these information delivery systems.

A. Telecommunications

The perfection and proliferation of public, low cost, packet-switched networks (eg. TYMNET, TELENET, GE GEISCO in the U.S.A.; UKPSS in the UK; EURONET-DIANE in Europe; DATAPAC in Canada) has made feasible the accessing of SNDS by at least two distinct schemes: centralized or "star" (host computer at one node of the network, users may access from any other node); gateway (several hosts at various nodes but a gateway computer selects and interfaces with the desired host on behalf of the user). The speed and integrity of these networks in general is such that continental boundaries are becoming transparent -- the computer on another continent often responds just as well as the one next door.

Typical examples of centralized networks are the Chemical Information System (CIS)[14] in the U.S.A., the DARC Pluridata System (DPDS)[15] in France, the Information System Karlsrube (INKA)[16] in the Federal Republic of Germany and Scientific Numeric Databases Services (CAN/SND)[5] in Canada. Common to these services is the goal of "one-stop shopping" where users can anticipate having all their database needs met simply by moving from database to database within one host computer.

A practical example of a gateway network is the Intelligent Gateway Network (iNet) trial underway in Canada[17]. On the basis of a personal profile maintained in the gateway computer, users may access a wide variety of hosts simply by entering the name or mnemonic for the desired host. All the user need do is remember one telephone number, one account code and one password; the gateway computer makes the necessary telephone connections and supplies the protocols applicable to the selected host. One of the conclusions of the Materials Data Workshop[18] held in November, 1982 in Tennessee, U.S.A. was that an intelligent gateway should be used to make the various hosts and databases comprising the proposed material properties data system easily and widely available. Ideally, the gateway computer could automatically select the database appropriate to the user's query as well as execute the basic connecting functions just described.

B. Databases

A fairly recent inventory of SND in the physical-chemical disciplines has been compiled by Hilsenrath[19]. The categories employed in that publication serve as a useful framework for listing some of the databases released or nearing release since the compilation was completed.

1. Identification of Unknown Substances

In the area of infrared spectra, a new component of CIS[14] called Infrared Search System (IRSS) features fully digitized spectra, with the potential option of graphical reproduction at the user's terminal, rather than the so-called "fingerprint" representations used in the systems mentioned earlier[3-6]. The trade-off to be considered, of course, is the much smaller number of spectra (approximately 3000-5000) currently available in IRSS.

Two new crystallographic databases will soon be available. The Inorganic Crystal Structure Database (ICSD)[16, 28] and the Metal Data File (MDF)[5,25], which have bibliographic and structural data on the substances suggested by their names, are complementary to the Cambridge Crystallographic Database of organic and organometallics mentioned earlier.

2. Properties of Pure Substances and Mixtures

The database being compiled as part of the program of the Design Institute for Physical Property Research (DIPPR)[20,21] will contain thermodynamic and physical property data of industrially important compounds. Current information indicates that members of DIPPR will have priority access to both hardcopy and magnetic forms. Public release of the first 200 compounds is anticipated in 1984.

Data on thermophysical, thermoradiative, electronic, electrical, dielectric, optical and magnetic properties of all materials of scientific and technical interest are being assembled into a Material Properties Numerical Data System[22] by the Center for Information and Numerical Data Analysis and Synthesis (CINDAS) at Purdue University. Plans call for it to be online and interactive but no release date has been announced.

THERMO, the NBS Chemical Thermodynamic Database which has recently been released on the CIS[14], contains the recommended values for selected thermodynamic properties of about 15,000 inorganic and simple organic (one- and two-carbon atoms) substances. This is an excellent example of a SND in which the evaluation function has been strongly emphasized.

3) Properties of Materials

Source databases dealing with properties of materials of engineering interest, especially mechanical properties, have tended to remain small in size and limited in accessibility. A recent review by Westbrook[21], for example, reported that over 40 such databases have been identified, none of which was publicly available. One major impediment is that the properties of engineering materials are more complex to quantify than those of pure substances; universally accepted standard test methods, materials

properties definitions and appraisal procedures are still pending. These and other problems were discussed at the Materials Data Workshop[18] mentioned earlier. As an indication that some progress is being made, however, three databases dealing with steel and plastics were announced there as being publicly available; a fourth database, on metals properties, has been recently announced.

Measured Properties of Steel[23] contains mechanical properties, long term creep rupture data, fatigue behaviour, deformation properties and physical characteristics on some 300 grades of steel. References to the relevant literature are also stored. Standard Properties of Steel[23] contains standard values of the chemical composition as well as the mechanical, technological and physical characteristics of approximately 1000 grades of steel.

A textual-numeric database system with property or attribute data but limited analysis capability, POLYPROBE[24] contains information about various characteristics of commercially available plastics. POLYPROBE is meant to be used chiefly as an expedient means of locating suitable materials rather than as a research tool in predicting performance or response.

Metals Datafile/I[25] contains mechanical and physical properties data like tensile strength, yield point, shear and impact strength, hardness, fatigue life, density, specific heat, melting temperature and conductivity as well as composition, specification and designation information. Bibliographic references may also be retrieved. Like POLYPROBE, this database appears to be primarily a means of locating, rather than studying, the material of interest.

V Survey of Activity in Canada

Through the CAN/SND office of the Canada Institute for Scientific and Technical Information, the National Research Council (NRC) encourages and supports SND use and development in Canada. A national online network has been established, several SND are nearing completion and some interesting database management system applications are being developed.

As mentioned in the previous section, the DATAPAC network of the TransCanada Telephone System permits anyone in Canada with a terminal and a telecommunication link to access the SND mounted on the NRC computer in Ottawa. The current access cost of (Cdn) $10.50 per hour is the same for virtually all users because network charges are distance independent and there are sufficient network nodes that most users need only make local (ie. time independent) telephone calls. With added charges for computation, typical costs are (Cdn) $30-$90 per connect hour. Publicly available since November 1981, the CAN/SND online service currently has 37 institutional accounts. Users may interact with the database of their choice in either English or French. Two databases, infrared spectra (SPIR) and the Cambridge Crystallographic Database[2] (CRYSTOR), are available at the time of writing. Others planned for the near future include the MDF and ICSD mentioned earlier as well as the Crystal Data Identification File[26], the Powder Diffraction File[26] and the Facility for Analysis of Chemical Thermodynamics (FACT)[7].

The MDF, which contains crystallographic and bibliographic data for metallic structures determined by diffraction methods, is being developed in the Chemistry Division of the National Research Council[27]. When completed, the file will contain about 5600 entries for structures determined from 1913 to the present. An additional 4000 entries, covering 1975 to the present, describe those metals and alloys for which sufficient data are available to assign them to known structure groups. Another 800-1000 entries will be added annually as updates. Initial plans call for the National Research Council to make the MDF available online in Canada and by magnetic tape lease worldwide except in the Federal Republic of Germany where the Fachinformationszentrum Energie, Physik, Mathematik[16] (FIZ) will be responsible.

The ICSD, containing the same types of data as the MDF but for inorganic compounds, is being developed through German-Canadian co-operation[28]. The University of Bonn produces the major part of the file with support from the FIZ. McMaster University, under sponsorship of the National Research Council, produces most of the remainder. With 17,000 entries at present, the ICSD is about 80% complete. Another 1200 entries are anticipated each year as updates. FIZ will be making the ICSD available online from Karlsruhe or by magnetic tape lease in all parts of the world except Canada where the National Research Council will be responsible.

To complement ongoing database efforts, work is also proceeding on the development of search, retrieval and analysis software. The National Research Council Information System (NIS)[29] has been adapted for interactive use of the Crystal Data Identification File[30] as well as the MDF and ICSD. In prototype is an application of NIS to the connectivity files of the Cambridge Crystallographic Database. Preliminary results indicate a saving in execution time by factors of 10-20 over times required by the connectivity searching program (CONNSER) supplied by the Cambridge Crystallographic Data Centre.

VI Conclusion

Given their capabilities and benefits, scientific numeric databases are cost-effective research tools for scientists and engineers. Developments in telecommunications and increases in the number and types of databases available, not to mention the impact of personal computers, are all combining to make these tools readily accessible on a worldwide scale.

References

1. D.M. Abels, K.R. Duzy, R.G. Alden, R.N. Cuadra and J. Wanger, eds., Directory of Online Databases, Spring 1983 edition, California, U.S.A., Cuadra Associates, Inc., p. 7

2. F.H. Allen et al, The Cambridge Crystallographic Data Centre: Computer-based Search, Retrieval, Analysis and Display of Information, Acta Cryst, Vol. B35, 1975, p. 2331

3. Sadtler Research Laboratories, 3316 Spring Garden St, Philadelphia, PA 19104, U.S.A.

4. DNA Systems Inc., P.O. Box 1424, Saginaw, Michigan 48605, U.S.A.

5. CISTI, National Research Council of Canada, Ottawa, Canada, K1A 0S2

6. Chemir Laboratories, 761 W. Kirkham, St. Louis, M063122, U.S.A.

7. Thermfact, 447 Berwick Ave., Mont-Royal, Québec, Canada, H3R 1Z8

8. Prof. C.B. Alcock, University of Toronto, Toronto, Ontario M5S 1A4, Canada

9. Themodata, B.U.S., B.P. 22, 38042 Saint-Martin-D'Heres, France

10. R. Taylor and O. Kennard, Crystallographic Evidence for the Existence of C-H··· O, C-H··· N and C-H··· Cl Hydrogen Bonds, J. Am. Chem. Soc., Vol. 104, 1982, 5063

11. E. Bye, W.B. Schweizer, J.D. Dunitz, Chemical Reaction Paths. 8. Stereoisomerization Path for Triphenylphosphine Oxide and Related Molecules: Indirect Observation of the Structure of the Transition State. J. Am. Chem. Soc., Vol 104, 1982, 5893

12. R.E. Rosenfield, Jr. & P. Murray-Rust, Analysis of Atomic Environment of Quaternary Ammonium Groups in Crystal Structures, Using Computerized Data Retrieval and Interactive Graphics: Modeling Acetylcholine - Receptor Interactions, J. Amer. Chem. Soc., Vol. 104, 1982, 5427; P. Murray-Rust et al, Intermolecular Interactions of the C-F Bond: The Crystallographic Environment of Fluorinated Carboxylic Acids and Related Structures, J. Amer. Chem. Soc., Vol. 105, 1983, p. 3206

13. V. Hampel, Fact Retrieval in the 1980's, AGARD Conference Proceedings No. 304, Technical Information Panel Specialists' Meeting, Munich, September 1981, p. 6-2→6-4

14. CIS User Support Group, 6565 Arlington Blvd, Falls Church, VA 22046, U.S.A.

15. DARC PLURIDATA SYSTEM, 25 Rue Jussieu, F-75005 Paris, France

16. Fachinformationszentrum Energie, Physik, Mathematik GmbH, Karlsruhe, D7514 Eggenstein-Leopoldshafen 2, F.R.G.

17. P. Wolters, The iNet Gateway Trial, AGARD, Technical Information Panel Symposium, Ottawa, September 1983, paper no. 2

18. J. Westbrook, Chairman Steering Committee, Report to be published

19. J. Hilsenrath, Summary of On-line or Interactive Physico-Chemical Numerical Data Systems, NBS Technical Note 1122, National Bureau of Standards, U.S.A., 1980

20. S.C. Stinson, Institute Gathers Chemical Engineering Data, Chemical and Engineering News, Vol. 61, 1983, p.34

21. J.H. Westbrook, Cooperation in Developing Computerized Materials Data Bases, CODATA'82 Conference, Poland (to be published)

22. CINDAS, Purdue University, West Lafayette, Indiana 47906, U.S.A.

23. Betriebsforschungsinstitut, VDEh-Institut für angewandte Forschung GmbH, Sohnstrasse 65, 4000 Dusseldorf, F.R.G.

24. International Plastics Selector, Inc., P.O. Box 26637, San Diego, California 92126, U.S.A.

25. Metals Information, American Society for Metals, Metals Park, Ohio, U.S.A.

26. JCPDS-International Centre for Diffraction Data, 1601 Park Lane, Swarthmore, PA 19081

27. L.D. Calvert and J.R. Rodgers, The Metal Data File, Computer Physics Communications, (to be published)

28. G. Bergerhoff et al, The Inorganic Crystal Structure Database, J. Chem. Inf. Comp. Sci. (accepted for publication)

29. National Research Council Computation Center, National Research Council, Ottawa, Canada K1A 0S2, Publication No. TSS-4035A-1, 1982

30. J.R. Rodgers and A.D. Mighell, Searching the NBS Crystal Data File, American Crystallographic Association Series II, Vol. II, Part I, 1983, p.15