(10) Freeman, J. E.; Katz, R. M. "Information Marketing". *Annu. Rev. Inf. Sci. Technol.* **1978**, *13*, 37–59.

(11) Martin, Y. C. "A Practitioner's Perspective of The Role of Quantitative Structure–Activity Analysis in Medicinal Chemistry". *J. Med. Chem.* **1981**, *24*, 229–237.

(12) Brown, H. D.; Costlow, M.; Cutler, F. A., Jr.; Demott, A. N.; Gall, W. B.; Jacobus, D. P.; Miller, C. J. "The Computer-Based Chemical Structure Information Systems of Merck Sharp and Dohme Research Laboratories". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 5–10.

(13) Howe, W. J.; Hagadone, T. R. "Molecular Substructure Searching: Computer Graphics and Query Entry Methodology". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 8–15.

(14) Eggers, I. M. R.; Gall, W. B.; Cutler, F. A., Jr.; Brown, H. D. "Use of Proprietary Biological and Chemical Data at Merck & Co., Inc.". In "Retrieval of Medicinal Chemical Information". *ACS Symp. Ser.* **1978**, *No. 84*, 85–106.

(15) Gund, P.; Andose, J. D.; Rhodes, J. B.; Smith, G. M. "Three-Dimensional Molecular Modeling and Drug Design". *Science (Washington, D.C.)* **1980**, *208*, 1425–1431.

(16) Bernd, C. L. "Chemical Case Report Forms Design—A Key to Clinical Trial Success". *Drug Inf. J.* **1984**, *18*, 3–8.

(17) Schwartz, S.; George, R.; Owens, J.; Edson, J. "A Clinical Database Management System". *Drug Inf. J.* **1984**, *18*, 33–42.

(18) Herman, R. L., Ed. "Drug-Event Association: Perspectives, Methods, and Uses". *Drug Inf. J.* **1984**, *18*, 195–352.

(19) "Corporate Strategies". *Business Week* **1984**, *November 26*, 114–118.

(20) "Merck World 5"; Merck & Co., Inc.: P.O. Box 2000, Rahway, NJ, 1984; No. 5, December, p 5.

(21) Souder, W. E. "Encouraging Entrepreneurship in the Large Corporations". *Res. Manage.* **1981**, *24* (3), 18–22.

(22) Branscomb, L. M. "Improving R&D Productivity: The Federal Role". *Science (Washington, D.C.)* **1983**, *222*, 133–135.

(23) Baillie, A. S. "Management of Risk and Uncertainty". *Res. Manage.* **1980**, *23* (2), 20–24.

(24) Ozog, S. "On the Value of Information". *J. Am. Soc. Inf. Sci.* **1979**, *30*, 310–315.

(25) Branscomb, L. M. "Information: The Ultimate Frontier". *Science (Washington, D.C.)* **1979**, *203*, 143–147.

(26) Carlson, W. M. "Information Is Not a Manageable Resource". *Inf. Manager* **1980**, *2* (2), 6.

(27) Horton, F. W. "Information is a Manageable Resource". *Inf. Rec. Manage.* **1981**, *15* (1), 9.

The Chemical Information System and Spectral Databases[†]

STEPHEN R. HELLER[‡]

Silver Spring, Maryland 20902

Received January 1, 1985

From 1970 to 1984, the U.S. Government cooperated with various organizations in the support of the development, maintenance, and distribution of a computer-based chemical information system of spectral and other numeric databases, known as the NIH/EPA Chemical Information System (CIS). This presentation discusses the history of the project and related activities in the area of numeric database activities and summarizes the current state of the project.

INTRODUCTION

A major activity in modern chemistry is the identification of chemical substances from laboratory measurements made on these substances. It was this activity that lead the National Institutes of Health (NIH), in the early 1970s, to initiate an informal project for the identification of chemicals using a computer-based mass spectral database and associated search software.[1] It rapidly became evident that, in addition to mass spectral data, there were other capabilities that were needed in a modern chemical laboratory. These included other spectral and numeric data, coupled with chemical structure search, manipulation, and retrieval capabilities. Having available an excellent computer facility at NIH, which included a large PDP-10 time-sharing interactive computer, some databases, and the interest of a core group of scientists desiring to explore new areas, a chemical information system began to germinate.

In the early 1970s the main interest of the chemical information community was with bibliographic databases. Thus, the efforts of the group at NIH, later joined by scientists at the National Bureau of Standards Office of Standard Reference Data (NBS, OSRD), the Food and Drug Administration (FDA), and others, began to explore a new and potentially promising area. In late 1973 the Environmental Protection Agency (EPA) initiated an expansion of its support activities to EPA laboratories and state and local government laboratories by joining this budding informal cooperative effort to develop a computer system for support of its environmental legislative mandates.

Dr. Stephen Heller received a B.S. in chemistry in 1963 at the State University of New York—Stony Brook and a Ph.D. in organic chemistry in 1967 at Georgetown University. He is presently a research leader at the U.S. Department of Agriculture, Agricultural Research Service, in Beltsville, MD. Previously, he has served as a chemist at the U.S. EPA, as a Senior Staff Fellow at the National Institutes of Health (1970–1973), and as a chemist for the U,S. Army (1967–1969). He also served for 6 months with the U.S. House Subcommittee on Health and the Environment (1979–1980) and as a Lady Davis Visiting Professor of Chemistry at the Hebrew University, Jerusalem (1981). He has published over 100 papers and books during the past 20 years.

This paper will describe the development of the spectral databases of the chemical information system, called the NIH/EPA Chemical Information System (CIS), of some related spectral database activities, and of some recent events regarding this project. Further details on the development of

[†] This article does not reflect the official views of the EPA.

[‡] Address correspondence to the author at the USDA, ARS, BARC-West, Beltsville, MD 20705.

0095-2338/85/1625-0224$01.50/0 © 1985 American Chemical Society

the system can be found in a number of articles published over the past decade.[2]

BACKGROUND

One area of modern chemistry and toxicology that is receiving considerable attention is the identification of chemical substances and their properties from laboratory measurements made on these substances. Whatever measurement technique is used, the task generally devolves into one of recognizing a "fingerprint" given by the unknown substance in, for example, a mass spectrometer, when thousands of "fingerprints", e.g., mass spectra, of known compounds are available. In the past decade, a computer-based system has been developed to aid the chemist in this task of identification and to provide numeric toxicological and environmental data on chemicals identified. This paper will describe the results of that effort and the future areas that we feel need study and attention.

The development of the first readily available spectral search, the Mass Spectral Search System (MSSS),[1] led over the past decade to the step-by-step building of a large computer-based chemical information system. The system, which was developed in cooperation with many groups in the U.S. and elsewhere, is called the NIH/EPA Chemical Information System (CIS).[2] The NIH/EPA Chemical Information System permits "fingerprint recognition" in a variety of efficient and inexpensive ways and is used very heavily in this manner by scientists all over the world. Furthermore, the CIS contains considerable toxicological information and data so that the properties of chemicals can also be made readily available.

From its inception, with the first publicly available component of the CIS, the MSSS, the CIS has always resided on private–sector, commercial computers. From 1972 to 1984, six separate vendors were used, with each new vendor providing lower hardware costs than the previous vendor. In addition, the majority of the work performed in developing the CIS was, after the initial U.S. Government staff programming, all done by grant or contract, with universities and private companies for the software development, system maintenance, database updates, and user support. The role of the U.S. Government staff was primarily in managing and technically directing the grants and contractor staff, which numbered at its peak some 50 professionals.

The quantity of data associated with chemistry has been expanding in recent decades, but until the advent of third-generation (integrated-circuit) computers, handling and using this vast amount of information has presented insuperable problems. With modern computer technology, the cost of computerization has come down, while accessibility of computers has increased through the use of computer networks—accessible over standard telephone lines. In these circumstances, the development of a highly interactive, disk-oriented chemical information system of numerical data was inevitable. The strength of such a system is its ready availability and low cost to many Government agencies' laboratories, contractors, grantees, scientific collaborators, and the general public. Related to the vast quantity of data that is available is the issue of quality. While most of the numeric spectroscopic and toxicological data that are in the CIS come from published sources, the peer review process of such data is not sufficient. Since few journals and publications are concerned with the quality of data itself (accuracy, precision, completeness, comparability, purity of the samples, experimental conditions of spectrometers, or handling of animals), much reported data are not of good or even defined quality. While the U.S. Congress has established an Office of Standard Reference Data (OSRD) at the National Bureau of Standards (NBS), lack of funding has prevented this office, along with others, from producing large amounts of quality numeric data.

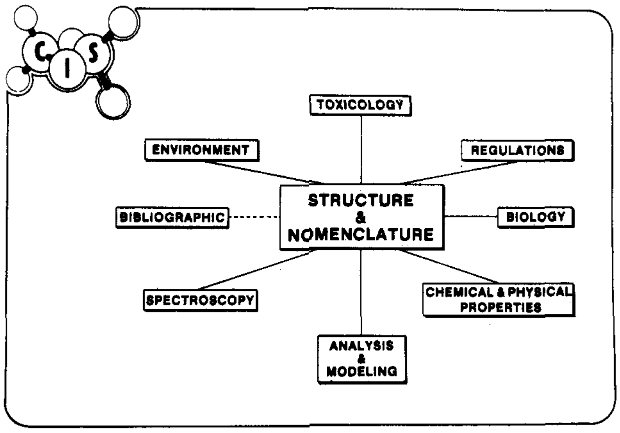


Figure 1. Overview of the NIH/EPA CIS.

The NIH/EPA CIS consists of a collection of chemical databases, together with a battery of computer programs for interactively searching. In addition, the CIS has a data locator or referral capability as well as a data analysis software system. It can be thought of, then, as having five main areas: (1) numerical chemical/physical property databases; (2) toxicology and regulatory databases; (3) structure and nomenclature search system; (4) database referral/chemical locator function; (5) data analysis software.

The numeric chemical and physical property databases that are part of the CIS include files of mass spectra,[1] carbon-13 nuclear magnetic resonance,[3] and X-ray diffraction data for crystals[4] and powders.[5] The toxicology and regulatory databases of the CIS include mammalian acute toxicity data,[6] hazardous chemical data,[2] water pollutant data,[2] aquatic and dermal toxicity data,[7] U.S. Federal Register and Code of Federal Regulations citations,[2] and Clinical Toxicology of Commercial Products.[8] The analytical programs include a family of statistical analysis and mathematical modeling algorithms[9] and chemical modeling and energy minimization of conformational structures.[10]

The center or "hub" of the CIS is the Structure and Nomenclature Search System (SANSS),[11] which allows the user to search through databases of structures (such as those associated with collections of mass spectra) for occurrences of a specific structure or substructure.

The entire CIS structure can be viewed, as seen in Figure 1, primarily as a system of independent numerical databases, most of which are linked together through the SANSS "hub", with the Chemical Abstracts Service Registry Number (CAS RN) as the unique universal chemical identifier for each compound. The use of the CAS RN to "tag" all CIS files was codified in EPA regulation 2800.2 in 1975.[12] With the passage of TSCA in 1976, the use of the CAS RN was extended to the TSCA inventory; this establishes the link between regulatory data and scientific data, both within CIS and in the literature.

LINKING NUMERIC DATABASES THROUGH CHEMICAL STRUCTURES

Structure and Nomenclature Search System (SANSS). All the compounds in the files of CIS have been assigned a Registry Number by CAS. The Registry Number is a unique identifier for that compound and may be used to retrieve from the CAS Master Registry of about 7 000 000 entries all the synonyms that the CAS has identified for that particular compound—these synonyms being names that have been used for the compound in addition to the name(s) used in the CAS *Eighth*, *Ninth*, or *Tenth Collective Index*. Further, the Registry Number can be used to locate in the CAS files the connection table for the compound's structure. This is a two-dimensional record of all atoms in the molecule together with the atoms to which each is bonded and the nature of the bonds. The connection table is the basis of the substructure search component of SANSS.

The purpose of the SANSS is to permit a search for a user-defined structure or substructure through databases of CIS. If a substructure is found to be in a CIS database, then, armed with its CAS Registry Number, the user can access that file and locate the compound and hence retrieve whatever data are available for it.

There are a number of ways to search the CIS Unified Database.[11] The main ones are (1) name/fragment name search (NPROBE), (2) nucleus/ring search (RPROBE), (3) fragment search (FPROBE), (4) structure code search (SPROBE), (5) molecular weight, molecular formula, partial formula, (6) total atom-by-atom, bond-by-bond search (SUBSS), and (7) total or full structure search (IDENT).

While structure searching is very important and cannot be replaced by other methods (such as fragment searching, linear notations, or name searching), the ability to search for a chemical by name or partial name (NPROBE) is still very useful in many cases and should be used first, if possible. In particular, if one wishes to search for a drug or pesticide, many of which have simple and short trivial names, a name search is likely to be the best method because such compounds are often complex cyclic structures, difficult to draw.

SPECTRAL DATABASES

Mass Spectral Search System (MSSS). The first database of the CIS was a collection of 8124 mass spectra provided to NIH by Professor Biemann of MIT. With this database in hand, a collection of programs to search and manipulate the data and information in this database were developed over a period of years.[1] It was the learning experience from this first database that began to drive all future activities of the NIH/EPA CIS project. For example, it was clear from almost the start that the database was not large enough to be as useful as tool as desired. Furthermore the multicopies of spectra, coupled with different names for the same materials, led to confusion, and probably to a reinforcement of the notion of "garbage in–garbage out". The lack of a measure of the quality of the mass spectral data was also an issue. Lastly, the need for full-time staff to obtain, enter, edit, and prepare a final product of the mass spectral database was the most critical problem, as solving the technical matters are usually simpler. Thus, a collaboration with the Mass Spectrometry Data Centre (MSDC) in Aldermaston (and later Nottingham), England, was initiated, through the NBS OSRD. For a number of years the MSDC disseminated the mass spectral database and programs to search the file, which was given the name the Mass Spectral Search System (MSSS).[1] However, changes occurred in the U.K. government policy, coupled with decision of EPA management to assume a greater role in the development and dissemination of mass spectral data [which occurred at the time the EPA laboratories and Research Office (ORD) had just started their efforts to use mass spectrometry as the main tool for pollutant and toxic chemical identification].

From 1974 to 1980 EPA, NIH, and NBS funded and collaborated in the development of a quality index (QI) for the mass spectral database, which was comprised of some 10 data quality indicators (DQI's), which were originally developed by McLafferty and co-workers.[13] Unfortunately, there has been some minor changes in the DQI's between the McLafferty database and the NBS database,[14] and thus, the QI's in the two databases today differ slightly.

From 1971 to early 1985, the mass spectrometry database grew from 8124 mass spectra to 42 261. In addition, thousands of duplicate spectra and spectra of labeled chemicals were archived to a separate database, which now consists of over 70 000 spectra. Coupled with this activity, Chemical Abstracts Service (CAS) Registry Numbers were obtained for all chemicals in the database, providing unique identifiers for each spectrum. Associated with the addition of CAS Registry Numbers were the CAS standard index names, molecular formula for all chemicals in the standard Hill notation form, and a quality index for each spectrum. When a new spectrum is received, even if it is a duplicate, it will replace a spectrum in the database if its quality index is higher than that of the existing spectrum in the database. Thus, a "living" database has been created, which contains both new and replaced spectra at every yearly update.

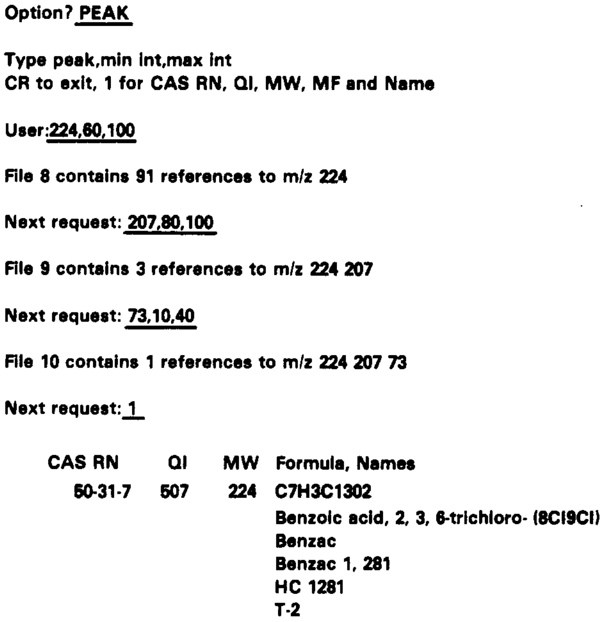
The mass spectral database is now being made available through the NBS OSRD[15] as well as having been combined with the McLafferty mass spectral database, available from John Wiley & Sons.[16] In addition, the NBS OSRD, through the U.S. Government Printing Office, has made available a hard-copy version of the database. These books, called the *EPA/NIH Mass Spectral Data Base*, were first published as a four-volume set, with an index. Over time another two volumes were published, and the entire set now consists of six volumes and an index that covers all six volumes.[15]

The mass spectrometry database is now under the joint responsibility of the NBS OSRD and the EPA Office of Research and Development, Environmental Monitoring, and Support Laboratory in Cincinnati, OH.[17] Data are still being received from the MSDC, as well as contributions from the scientific community and from an EPA contract for new mass spectral data, which is part of the ORD EMSL activities.

In addition to the EI mass spectral database, there have been some efforts in the areas of other types of mass spectral data, both chemical ionization (CI) and fast ion bombardment (FAB) spectra. At present, the CI database consists of less than 2000 spectra, and the FAB activities at MSDC[18] are still in their formative stages.

In addition to the value of electron-impact mass spectra, the high level of interest in chemical ionization mass spectrometry has led to a need for a reliable file of gas-phase proton affinities. No database of this sort has previously been assembled, and the task of gathering and evaluating all published gas-phase proton affinities was completed by Rosenstock and co-workers at NBS. This file,[19] which has about 400 critically evaluated gas-phase proton affinities drawn from the open literature, can be searched on the basis of compound type or the proton affinity value.

Searches through the MSSS database can be carried out in a number of ways. With the mass spectrum of an unknown substance in hand, the search can be conducted interactively, as is shown in Figure 2. In this search the user finds that 91 database spectra have a peak (minimum intensity 60%, maximum intensity 100%) at an m/z value of 224. When this subset is examined for spectra containing a peak at m/z 207 with intensity of between 80 and 100%, only three spectra are found. The entering of a third peak, at an m/z value of 73 (with an intensity between 10 and 40%), narrows the search down to just one answer, which is then printed out. In the example shown, the answer "2,3,6 trichloro benzoic acid" is shown with a number of synonyms used in naming this chemical, as well as other identifying information. If there still had been a large number of answers after entering the three peaks used in this example, the search could have been reduced further to a manageable number of spectra by entering further peaks. In addition, the database can be examined for all occurrences of a specific molecular weight or a partial or complete molecular formula. Combinations of these properties can also be used in searches. Thus, all compounds containing, for example, five chlorines and whose mass spectra have a base



```
Option? PEAK

Type peak,min int,max int
CR to exit, 1 for CAS RN, QI, MW, MF and Name

User:224,60,100

File 8 contains 91 references to m/z 224

Next request: 207,80,100

File 9 contains 3 references to m/z 224 207

Next request: 73,10,40

File 10 contains 1 references to m/z 224 207 73

Next request: 1

    CAS RN      QI     MW   Formula, Names
      50-31-7   507    224  C7H3C1302
                            Benzoic acid, 2, 3, 6-trichloro- (8CI9CI)
                            Benzac
                            Benzac 1, 281
                            HC 1281
                            T-2
```

Figure 2. Peak search in the CIS MSSS.

peak at a particular m/z value can be identified.

In contrast to these interactive searches, which are of little appeal to those with large numbers of searches to carry out, there are available two batch-type searches that accept the complete spectrum of the unknown substance and examine all spectra in the file sequentially to find the best fits. These are the KB (forward search) and PBM (reverse search) search algorithms. Spectra can be entered from a teletype; but in a more powerful approach, a user's data system can be connected to the network and the unknown spectra down-loaded into the network computer for searching.

Once an identification has been made and the name and CAS Registry Number of the database compound are reported to the user, the database spectrum can be listed or, if a CRT terminal is being used, plotted, to facilitate direct comparison of the unknown and standard spectra.

Carbon-13 Nuclear Magnetic Resonance (CNMR) Spectral Search System. The CNMR database consists of over 10 000 CNMR spectra. Until funding was no longer provided by the U.S. Government, every compound has a CAS Registry Number, and exact duplicate spectra had been removed from the file. The CNMR file is still small but is growing and should benefit from recent international agreements to the effect that many major compilations of CNMR data will, in the future, be pooled.[20]

In addition to the CNMR database administrated in The Netherlands, and being developed by Professor Mills at the University of Manchester, there is a second CNMR database. This database has been developed by Bremser and co-workers at BASF in Germany.[21] This database consists of over 50 000 spectra of some 40 000 different compounds. The database uses chemical names from the literature and generally does not contain CAS Registry Numbers. The database uses its own substructure code. The database is made available on-line by INKA in Germany, which is now part of the CAS STN Network.[22]

Searching through the CIS CNMR database, as in the case of the MSSS, can be interactive or not.[6] In the interactive search, a user enters a shift, with an acceptable deviation. The algorithm reports the number of file spectra fitting this criterion. The names of the compounds whose spectra have been retrieved can be listed, or alternatively, the list can be reduced by the entry of a second chemical shift. A search for spectra of compounds having a specific molecular formula can also be carried out, but there is no capability for searching on molecular weight—a parameter considered to be of little relevance in CNMR spectroscopy.

If an interactive search is not appropriate to the problem at hand, a batch type of search through the database is available by the techniques described by Clerc et al.[23] To institute such a search, the user enters all the chemical shifts from the unknown and starts the search. The entire unknown spectrum is compared to every entry in the file, and the best fits are noted and reported to the user. This program searches for the absence of peaks in a given region as well as for the presence of peaks and thus has the capability of finding those compounds that are structurally similar to the material that gave the unknown spectrum.

When a search is completed, the user is provided with the CAS Registry Numbers of compounds whose spectra fit the input data. The names of the compounds in question are also given. If more information is required, the complete entry for a given CAS Registry Number can be retrieved. This includes a numbered structural formula, the name, molecular formula, and Registry Number of the compound, experimental data pertaining to the spectrum, and the entire spectrum, together with single-frequency off-resonance decoupled multiplicities and, for 60% of the spectra, relative line intensities and assignments.

In addition to the library aspects of CNMR, an interface between CNMR and SANSS (the structure search component of the CIS) has been written.[24] This allows a user to define a substructure and then examine the chemical shifts associated with particular carbon atoms of interest. The shift data are neatly plotted out to the user, with appropriate standard deviations for the data, and should be quite helpful in structure elucidation problems.

Infrared Search System (IRSS). The most recent addition to the spectroscopic databases of CIS is the infrared (IR) database. This IR database of complete spectra (i.e., the entire IR curve from about 4000 to 100 cm^{-1}) contained some 5500 vapor- and liquid-phase Fourier-transform and grating spectra as of late 1983.[25] The IRSS search capabilities are analogous to those of MSSS and CNMR. An IR peak or band can be searched by entering a wavenumber, with a range (owing to solvent effects), as is also the case in CNMR. Searches can also be performed for molecular weight, and the spectrum printout on a vector terminal (such as a Tektronix) contains some 700–1000 points, which comprise an excellent plot of the spectrum.

Plans that called for the IRSS database to be increased to some 15 000 spectra in the next years by using high-quality Coblentz Society grating spectra, which have been evaluated, and by obtaining additional spectra through scientific collaboration with industry, government laboratories, and academia have been cancelled owing to the current status of the CIS.

In addition to the CIS IR database activities, there have been a number of other activities in IR database development taking place. For a summary of existing IR databases, the reader is referred elsewhere.[26] The most well-known of the new IR database activities is the Sadtler IR database project.[27] In addition, the Aldrich Chemical Co. and the Nicolet Instrument Co. are working jointly and have produced a high-quality file of liquid-phase FT-IR spectra, which are available in both computer-readable and book form.[28]

X-ray Crystallographic Search System (CRYST). This is a series of search programs working against the Cambridge Crystal File[4] a database of some 33 000 compounds for which full atomic coordinate data are available and over 35 000 bibliographic entries dealing with published crystallographic data, mainly for organic compounds. The entry for each compound contains the compound name, its molecular weight and CAS Registry Number, the space group in which it crystallizes, and the parameters of the unit cell of the crystal, as well as the atomic coordinate data. The file may be searched on the basis of any of these parameters.

Almost all the compounds in this file have been registered by CAS, and these data are currently being merged into the CRYST system along with the connection tables for the structures. This database is therefore searchable on a structural or substructural basis, as are all the other files of the CIS.

Once an entry of interest in the Cambridge X-ray file has been located by one of the search programs, its "crystal sequence number" can be used to retrieve the appropriate literature reference, structure, or atomic coordinate data.

The database used in the X-ray crystallographic search system described above possesses complete literature references to all entries in the file.[4] This information has been made the basis of a system for searching by author, title word, etc. the literature pertaining to the X-ray diffraction study of organic molecules. It is possible to search for papers by a specific author or authors, and papers that appeared in given years in given journals may also be retrieved. Additionally, papers may be located on the basis of specific words appearing in their titles.

Once a paper of interest has been identified, all the crystallographic information in that paper can be examined because the crystal sequence serial number associated with the paper can be used in the crystallographic search system to retrieve that information. Alternatively, the CAS Registry Number of any particular compound can be used to retrieve any data of interest on that compound from other files of the CIS.

X-ray Crystal Data Search System (XTAL). The National Bureau of Standards (NBS) has collected a file of data pertaining to some 60 000 crystalline materials, including those in the Cambridge file described above.[29] The data in the NBS file include the cell parameters, the number of molecules in the unit cell, the measured and calculated densities of the crystal, and two determinative ratios, such as A/B and A/C (but no coordinate data). Every compound in the file is identified by its name, molecular formula, and CAS Registry Number (if available), and the file can be structurally searched by the CIS structure and nomenclature search system as is described below.

Searches through this database for crystals with specific space groups or densities have been developed, and it is possible to locate crystals with reduced cells of given dimensions. It is hoped that this will prove to be a very rapid method of identifying compounds from the readily measured crystal properties.

X-ray Powder Diffraction Search Match (PDSM) System. A collection of powder diffraction patterns proves to be a very effective means by which to identify materials, and indeed, one of the very earliest search systems in chemical analysis was based upon such data by Hanawalt[30] over 40 years ago. The importance of these data in TSCA can be seen by examining the TSCA Inventory regulations for treatment of confidential chemicals.[31] Section 710.7 of these regulations indicates that EPA intends to rely on powder diffraction data to assure the validity and seriousness of a manufacturer's request for treating information on a chemical as confidential.

The database of over 44 000 powder diffraction patterns, which was used in the CIS[32] until 1982, is in fact a direct descendant of that with which Hanawalt carried out his pioneering work. A problem that arises in connection with this particular component stems from the fact that powders are frequently mixtures of different crystalline phases, and so, the patterns that are obtained experimentally are often combinations of one or more file entries. A reverse-searching program,[5] which examines the experimental data to see if each entry from the file is contained in it, has been written after the general approach of Abramson[33] and seems to cope with this particular difficulty. A subtraction routine to help in identifying mixtures has also been implemented.

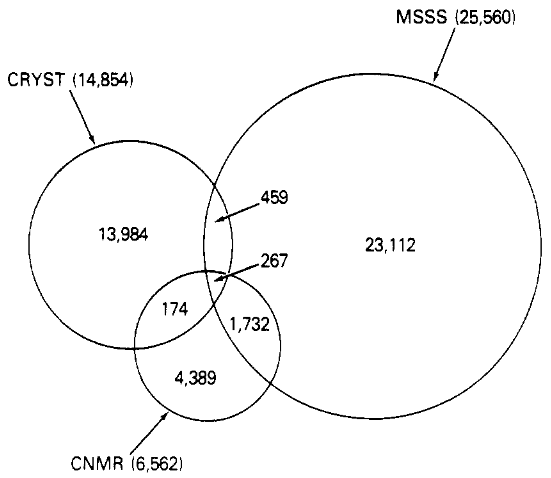


Figure 3. Overlap of some spectral databases in the NIH/EPA CIS.

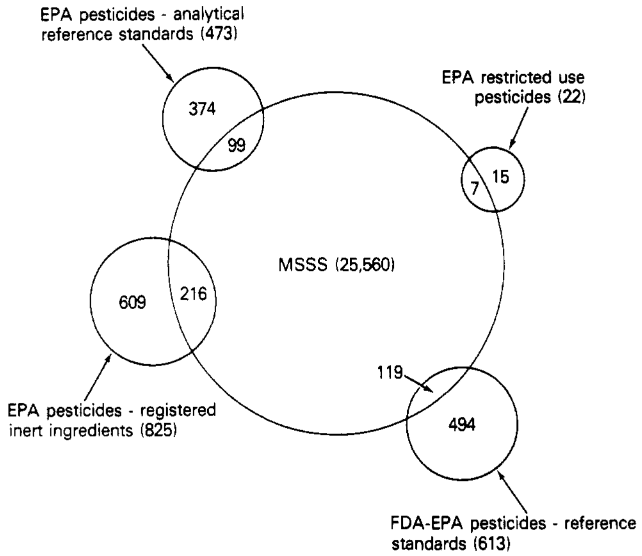

EPA pesticides - analytical reference standards (473)
374
99
EPA restricted use pesticides (22)
7
15
MSSS (25,560)
609
216
119
494
EPA pesticides - registered inert ingredients (825)
FDA-EPA pesticides - reference standards (613)

Figure 4. Mass spectral data of some lists of pesticides.

OBSERVATIONS ON NUMERIC DATABASE COLLECTIONS

One very interesting observation regarding the spectral database activities, other than their small sizes, is the amazing lack of overlap of having multispectral data on the same chemicals. A number of years ago a study was undertaken to see what the overlaps were between the CIS spectral databases, as well as the overlap of spectral data with other sources of information. The results of these studies, which used the INTERSECT command of SANSS, are shown in Figures 3–5. The first of these, Figure 3, shows the overlap of mass spectral, CNMR spectral data, and crystallographic data. About 1% of the mass spectra also have CNMR and crystal structures in the files. The overlap between CNMR and mass spectral data is better, from the point of view of the CNMR database, reaching a value of some 40%. In the area of mass spectra of pesticides, it is interesting to note in Figure 4 how few spectra there are in the database of registered pesticides, and EPA/FDA reference standards. Similarly, in Figure 5 one can see the overlap between the mass spectral database and files of drugs, commercial chemicals, and laboratory chemicals and a file of toxicology data (RTECS) is less than

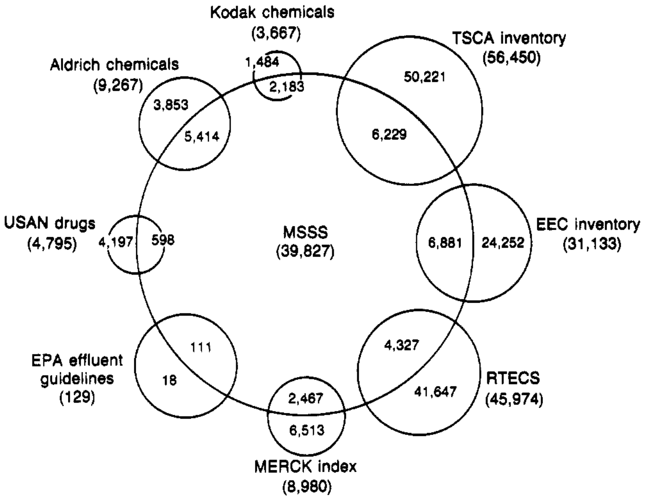


Figure 5. Overlap of mass spectral database and some SANSS lists of chemicals.

one would find desirable and useful.

Why are these overlaps so small? It seems the answer is probably quite simple. Whether it be mass spectral or CNMR or crystal structures, scientists obtain these data for their own specific research needs. Often one studies a particular class of chemicals by a certain method. Furthermore, scientists measure the acute toxicology of a chemical on materials for which spectral data is of little or no use.

RECENT EVENTS

In 1981 NIH began to reduce its support of the system, in terms of both financial and staff support. Over a period of about 1 year, culminating on 1 April 1982, a staff of two full-time scientists went to one, new part-time coordinator. At the EPA, with a new administration appearing in the spring of 1983, a management decision was made to phase out all EPA CIS management and coordination activities. Much of these EPA management activities have been documented elsewhere, and only highlights will be mentioned here. In early 1984, EPA convened two panels, one government and one from the private sector, to review the system and recommend what EPA should do. "The Government panel concluded that the CIS should be turned over to the private sector. The private panel suggested several alternatives, but generally concluded that government should continue to manage the system, while its implementation should go to the private sector."[34] "Ignoring, or at least short-circuiting, recommendations in [these] two panel reports assessing the system, EPA officials...decided to hand over operation of the database to private contractors without providing any interim federal support to ease the transfer."[35] After this was done, "critics of EPA, including members of the panels, accused the agency of using the reports as a 'smoke screen' and ignoring their recommendations."[35]

Other systems have been transferred from the government to the private sector. However, the usual method is to provide for an orderly transition, with public notice, comment, and thoughtful contractual efforts. Such was the case recently with the SCAN database from the Department of Health and Human Services, Administration on Aging.[36] In this case there seems to be no criticism of the transfer. In the case of the CIS, the situation is different, as evidenced by the following comment. "We feel that the CIS should continue to exist and that it should be kept intact. But it may be too late; not having been updated now for nearly a year. Costs are down to $250,000, but all the files need massive updating, and revenue has fallen off as users became disenchanted with data files without current data. Updating the system alone, after nearly twelve months, would be a very costly exercise. It seems inconceivable to us that the whole system could not be privitized, intact, and with guarantees concerning continuing availability and that, as a result, the CIS would be a better-used system."[37] A month later the same author continued as follows: "The EPA does not come out of the affair very well; it is the EPA that had, effectively, frozen the system and its updates since November 1983, giving anyone taking over the system the headache (and bank account ache) of having to start operations with a difficult and expensive one-year update. As we chronicled last month, it appears that the EPA dithered and embarked on in-fighting for nearly a year with the frozen system before precipitously announcing on 1 October (1984) that the system would not be supported after 31 October. A changeover to the private sector there may have been, but there is little else the EPA could have done to have made the changeover even less smooth."[38]

With the serious government budget and staffing problems, and the many other critical priorities at NIH, NBS, EPA, and elsewhere, it seems likely this decision will, for the foreseeable future, remain. There had been some outcry from the scientific community, but it has not generated sufficient strength to date to have any effect. The EPA decided to leave any database activities to either other agencies or individual offices within the EPA. Thus, a decade of cooperation and coordination between parts of EPA, other government agencies, and organizations in the the U.S. and abroad came to an end. The two EPA staff members working on the project either left EPA or were transferred to other activities. The coordination and dissemination of spectral database activities thus returned to being a cottage industry.

CONCLUSIONS

The NIH/EPA Chemical Information System (CIS) was an experiment in the collection and on-line dissemination of scientific numeric databases. The CIS began with spectral databases and later added toxicological, environmental, and regulatory databases, coupled with a structure search system, SANSS, which connected the various pieces. Owing to the ongoing cost of the project, the lack of management support, and the general trend of cutbacks in the U.S. Government budget, the CIS is no longer an existing activity. Parts of the system are being made available by commercial companies,[39,40] but with the low revenue ($50 000–$60 000 per month) from users, it is not clear how long even these parts of the CIS will be available. In fact, at the beginning of 1985, a number of the CIS components had been removed from the on-line system for financial reasons. New components were also expected to be added, but on a business income basis not a scientific basis. While this makes sense for the private sector, it is not clear the scientific community benefits by no longer having valuable but not commercially viable data available.

While numeric databases are heavily used in economics, finance, business, and advertising, their use in science is still quite limited. Much of this is no doubt due to the lack of numeric databases; and without the necessary publicity and marketing, even these few databases have not attracted a large community of users. Another factor that, in this author's opinion, has kept the number of scientific numeric databases so small, and little used, is the lack of interest and skill to generate and use these databases. Databanking activities are not generally viewed as attractive activities for young, bright, and energetic scientists. However, many librarians or information scientists, not trained as X-ray crystallographers, toxicologists, or the like, seem to feel insecure in performing these scientific numeric databank activities.

In spite of these problems, interest and use in scientific numeric databases will grow in the latter part of this decade.

While the end users of bibliographic databases readily settled for computer listings of probable or possible relevant references to their queries, users are becoming more sophisticated and demanding. If there is a heat of formation for bicyclo-[2.1.0]aziridine, a user wants to know its value and the accuracy of the number. If the experimental value is not available, the user wants an accurate estimation of the value. Gone will be days when an end user is satisfied with a list of 50–100 references to a bibliographic search for "physical properties of bicyclic nitrogen compounds with a three-membered ring".

To arrive at this ideal of providing timely and accurate answers to highly specific questions will take a vast cooperative effort between government, academia, the publishers of scientific journals, and industry, not just in the U.S. but worldwide. Mechanisms such as the U.S. National Bureau of Standards Office of Reference Data (OSRD) and international groups such as CODATA and IUPAC exist today as a potential nucleus for such an activity of the 1980s. Undoubtedly other groups, organizations, and even governments will become involved as time, interest, and need grows.

It now appears that the only organization in the U.S., and perhaps anywhere, that has management support for chemical information in the area of nonbibliographic or numeric databases is the Chemical Abstracts Service. CAS has, over the past few years, developed CAS Online for the dissemination of the CAS 7 million chemical structure database and more recently for the dissemination of the CAS bibliographic files of literature citations. CAS has indicated over the recent years that it would like to also provide scientific data. It has not yet taken any or all of the CIS databases, apparently for three reasons. First, it is moving slowly and carefully in developing the dissemination of its own CAS products. Second, the economics of the CIS spectral and other databases are such that it would take a considerable investment of funds to first bring the databases up to date (EPA having not funded most of the databases for over 1.5 years) and maintain the databases at the level the scientific user community desires. Lastly, the management and staffing of a scientific spectral and other numeric database activity requires considerably different staff than that which CAS now has. There is a need for evaluation (quality control), editing, validation, and other activities in the area of numeric databases, which do not exist in bibliographic databases. For example, in a bibliographic database, one can use an author's abstract, independent of the validity of the work, and the abstract is correct. Also, the abstract never has to be corrected. Finding mass spectral or NMR data requires a considerable intellectual effort or considerable expenses and also requires a commitment to correct or replace such data if it is found in error or if better data is found.

The very nature of developing and maintaining high-quality scientific numeric databases will probably require administrative oversight by the information community, coupled with strong technical direction and involvement by the expert scientific staffs in their respect fields of endeavor in the development of data-quality indicators or equivalent tags for the data.

However, none of these developments can proceed without a considerable amount of funding. While one can, in a sense, set up an abstracting operation in his garage and not worry about what is abstracted or the errors in the articles abstracted, this is not the case for numeric database activities. having only a number, or a set of values with tolerances, requires one to "stick his neck out" quite a bit further. For example, in *Chemical Abstracts* and *Biosis*, the word "Massachusetts" is spelled some 10 ways, as is "conformation"; and yet, these databases (via Lockheed, SDC, and others) have yet to be corrected, but continue to be used. In contrast, it is very unlikely that a database of pH values where one finds a pH of 27 would be trusted a second time. "It is regrettable that in a country where considerable expenditures of time and dollars are made for scientific research we have been unable to produce and maintain, on an economically stable basis, scientific numeric databases that could be shared by many researchers."[41]

Numeric databases in the physical and chemical sciences hopefully will develop in the 1980s much the way bibliographic databases developed in the 1970s; but, there will be considerable differences in the size, cost, and required quality-assurance and quality-control efforts. All in all, it should be a highly stimulating time in which the information and scientific communities must begin to work much more closely than they have in the past. It is a time for us all to look forward to.

ACKNOWLEDGMENT

While many people in the U.S. and elsewhere have been part of the cooperative CIS project, amongst those in the U.S. Government who have worked most to develop the project are G. Milne, H. Fales, M. Yaguda, R. Potenzone, L. Dusold, L. Gevantman, D. Lide, R. Feldmann, and R. Striker.

REFERENCES AND NOTES

(1) Heller, S. R. *Anal. Chem.* **1972**, *44* 1951. Heller, S. R.; Feldmann, R. J.; Fales, H. M.; Milne, G. W. A. *J. Chem. Doc.* **1973**, *13*, 130. Heller, R. S.; Milne, G. W. A.; Feldmann, R. J; Heller, S. R. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 176. Heller, S. R.; Heller, r. S.; Martinsen, D. P. *Adv. Mass Spectrom.*, **1980**, *8B*, 1578. Heller, S. R. Chem.-Kemi, **1984**, *January*, 15–16.

(2) Heller, S. R.; Milne, G. W. A.; Feldmann, R. J. *Science (Washington, D.C.)* **1977**, *195*, 253. Milne, G. W. A.; Heller, S. R.; Potenzone, R., Jr. *Science (Washington, D.C.)* **1982**, *215*, 371.

(3) Dalrymple, D. L.; Wilkins, C. L.; Milne, G. W. A.; Heller, S. R. *Org. Mag. Reson.* **1978**, *11*, 535.

(4) Kennard, O.; Watson, D. G.; Town, W. G. *J. Chem. Doc.* **1972**, *12*, 14.

(5) Marquart, R. G.; Katsnelson, I.; Milne, G. W. A.; Heller, S. R.; Johnson, G. G., Jr.; Jenkins, R. J. *Appl. Cryst.* **1979**, *12*, 629.

(6) McGill, J. R.; Heller, S. R.; Milne, G. W. A. *J. Environ. Pathol. Toxicol.* **1978**, *2*, 539.

(7) EPA data that are part of the EPA/OTS SPHERE project. For further information contact Ms. Paula Miles, EPA, OTS, TS-793, Washington, DC 20460.

(8) Gosselin, R. E.; Hodge, H. C.; Smith, R. P.; Gleason, M. N. "*Clinical Toxicology of Commercial Products*", 5th ed.; Williams and Wilkins: Baltimore, MD, 1984.

(9) Knott, G. D.; Shrager, R. I. *Assoc. Comput. Mach., SIGGRAPH Notes* **1972**, *6*, 138.

(10) Potenzone, R.; Cavicchi, E.; Weintraub, H. J. R.; Hopfinger, A. J. *Comput. Chem.*, **1977**, *1*, 187.

(11) Milne, G. W. A.; Heller, S. R.; Fein, A. E.; Fres, E. F.; Marquart, R. G.; McGill, J. A.; Miller, J. A.; Spiers, D. S. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 185. Heller, S. R.; Milne, G. W. A.; Feldmann, R. J. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 232.

(12) EPA Order 2800.2, issued May 27, 1975.

(13) Speck, D. D.; Venhataraghavan, R.; McLafferty, F. W. *Org. Magn. Reson.* **1978**, *13*, 208.

(14) Milne, G. W. A.; Budde, W. L.; Heller, S. R.; Martinsen, D. P.; Oldham, R. G. *Org. Mass Spectrom.* **1982**, *17*, 547.

(15) The NIH/EPA/MSDC database is available for lease on computer from the U.S. National Bureau of Standards (NBS), Office of Standard Reference Data, Gaithersburg, MD 20899 (telephone 301-921-2104). The database is also available in printed form (currently six volumes and an index volume; a seventh volume is being considered and should be available in 1986). Please contact NBS for details.

(16) John Wiley & Sons, Electronic Publishing Division, 605 Third Avenue, New York, NY 10158.

(17) W. L. Budde, private communication.

(18) Mass Spectrometry Data Centre, The University, Nottingham NG7 2RD, England.

(19) Hartmann, K.; Lias, S.; Ausloss, P. J.; Rosenstock, H. M. Publication NBSIR 76-1061, July 1976.

(20) The NIC/NIH/EPA CNMR database is available for lease from Dr. Charles L. Citroen, Netherlands Information Combine (NIC), CID-TNO, P.O. Box 36, 2600 AA, Delft, The Netherlands.

(21) W. Bremser, private communication.

(22) STN-Karlsruhe, c/o Fachinformationszentrum, Energerie, Physik, Mathematik, D-7500 Karlsruhe 1, Postfach 2465, West Germany.

(23) Clerc, J. T.; Schwarzenbach, R.; Meili, J.; Koenitzer, H. *Org. Magn. Reson.* **1976**, *8*, 11.

(24) Zupan, J.; Heller, S. R.; Milne, G. W. A.; Miller, J. A. *Anal. Chim. Acta* **1978**, *103*, 141.

(25) M. Penca, private communication.
(26) Fisk, C. L.; Milne, G. W. A. *J. Chromatogr. Sci.* **1979**, *17*, 441–444.
(27) Sadtler Research Labs, Inc., 3314 Spring Garden Street, Philadelphia, PA 19104.
(28) Pouchert, C. J. "The Aldrich Library of FT-IR Spectra", edition I; Aldrich: Milwaukee, WI, 1984; Catalog Z12,700-0. For information on the IR database on magnetic tape, contact Dr. C. Anderson, Nicolet Analytical Instruments, 5225-1 Verona Road, Madison, WI 53711-4495.
(29) These data are available as NBS tape 9. Contact the National Technical Information Service (NTIS), Springfield, VA 22151, for details.
(30) Hanawalt, J. D.; Rinn, H. W.; Frevel, L. K. *Ind. Eng. Chem.* **1938**, *10*, 457.
(31) Environmental Protection Agency (EPA), Toxic Substances Control Act (TSCA) Inventory Reporting Requirements, Federal Register, 42, 247, Friday December 23, 1977, pp 64572–64596. In particular, see section 710.7 on pp 64579–64580.
(32) International Centre for Diffraction Data, 1601 Park Lane, Swarthmore, PA 19081.
(33) Abramson, F. P. *Anal. Chem.* **1975**, *47*, 45.
(34) *Ind. Chem. News* **1984**, *October*, 6.
(35) Fox, J. *Science (Washington, D.C.)* **1984**, *226*, 816.
(36) Halpin, P. *J. Am. Soc. Inf. Sci.* **1985**, *36*, 53–55.
(37) Collier, H. *Monitor* **1984**, *September*, 3–4.
(38) Collier, H. *Monitor* **1984**, *October*, 1–2.
(39) CIS Project Manager, ICI Inc., 1133 15th Street, NW, Washington DC 20005 (202-822-5200).
(40) CIS Operations Project Manager, CIS Inc., 7215 York Road, Baltimore, MD 21212 (301-821-5980).
(41) Williams, M. E. *Science (Washington, D.C.)* **1985**, *228*, 445.

Chemical and Spectral Databases: A Look into the Future

JOHN R. RUMBLE, JR., and DAVID R. LIDE, JR.*

Office of Standard Reference Data, National Bureau of Standards, Gaithersburg, Maryland 20899

Received January 28, 1985

Over 50 databases of chemical and spectral information are now available, and in the coming years many more will be built. We discuss some of the current trends in the use of these databases and how such databases might affect chemistry.

As reflected by the change in the title of this journal in 1975, from *The Journal of Chemical Documentation* to the *Journal of Chemical Information and Computer Sciences*, the last 25 years have seen radical changes with respect to collections of chemical data. Today, the impact of computers on the compilation, evaluation, and dissemination of chemical data is obvious to all chemists as the articles in this special issue demonstrate. There is every reason to believe that a steady state has not yet been reached and that the next 25 years will see equally important changes. Not only are computers continuing to improve, but our understanding of the building and use of chemical databases is also growing.

In this article we attempt to look into the future by discussing some trends that exist with present chemical databases. In some cases the ideas represent directions that the Standard Reference Data program at the National Bureau of Standards (NBS) and others are now pursuing; in other cases, we can claim only speculation. For the former, we will give some concrete examples; for the later, only our best guesses.

DEFINITION OF CHEMICAL AND SPECTRAL DATABASES

In this paper, we will be discussing databases of *factual* information related to chemistry, chemical compounds, and their spectra. By the term *factual*, we mean the numbers, text, and graphics that identify or describe compounds and their properties. It includes data such as the structural geometry of molecules and crystals and complex graphs such as phase diagrams. It excludes *bibliographic* databases, which only contain references and abstracts, and *full-text* databases, which are computer-searchable versions of original research publications. Naturally, some databases are hard to classify, but this distinction is reasonably sharp.

With the above in mind, let us look into the future and try to understand how chemists will be able to use the computer as their primary source of chemical data.

THE NUMBER OF DATABASES

When compared to the number of printed handbooks and compilations, the number of computer databases of chemical and spectral data is very small. Hampel et al.[1] have identified about 52, which most likely is an underestimate by two-thirds. While the number of printed data sources is uncountable, there do exist some measures. At NBS, a collection of handbooks and other data compilations in chemistry, physics, and material sciences is maintained that now numbers over 2500 titles. At least one-third of these are specifically in the field of chemistry. Also, many of the physics and material sciences titles are of great use to chemists. However the classification is done, it is clear that this collection is of the order of 20–30 times larger than the number of databases.

As another method, the *Journal of Physical and Chemical Reference Data*, sponsored jointly by NBS, the American Chemical Society, and the American Institute of Physics, has published 256 articles and seven book-length supplements since its inception in 1972. Each article contains a significant compilation of evaluated data in chemistry and physics. Yet of these many valuable compilations, only two are now available as computer-readable databases.

From these examples, it is obvious that we are just beginning to make chemical data available via computer. Several barriers to building chemical databases have been identified:[2] scientists have little experience with database management systems; there is a temptation to reinvent existing database capabilities; most databases are built for individual use and are hard to adapt to more general use; users have not been involved enough in the design of databases; very few ways for accessing databases are available to the chemical community at large; funding for database building rarely is available. Of these, perhaps the last two are most important—the lack of an outlet and the small amount of support available.

These barriers will be overcome in the future but not necessarily easily. A commitment must be made by data publishers to issue databases simultaneously with publications. In reality, this will not be as much of a problem as it might appear since electronic typesetting is now the norm. What will be needed is an investment in transforming the typesetting files into usable databases. Within the NBS Standard Reference Data program, a policy has been adopted to build databases as the primary step and then spin-off both publications and distributable databases from the master database.