(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.

Dr. John Rumble, Jr., is in the Office of Standard Reference Data at the U.S. National Bureau of Standards. He is responsible for the Material Data Program in that office. His duties include identifying needs for evaluated data on materials properties, setting priorities, establishing projects, and overseeing the dissemination of the results. Dr. Rumble received a Ph.D. in chemistry from Indiana University. Prior to coming to NBS, he was at the International Atomic Energy Agency in Vienna and before that at the University of Colorado. He has also worked in industry as a chemist. Dr. Rumble has been involved in making computer access to materials data a reality. He has helped organize several workshops and has authored many papers and books in this field.

Dr. David R. Lide, Jr., was educated at Carnegie Institute of Technology (B.S. in chemistry, 1949) and Harvard (Ph.D. in chemical physics, 1952). He joined the National Bureau of Standards in 1954 to establish a laboratory for research in microwave spectroscopy and molecular structure. From 1963 to 1968 he headed the Molecular Spectroscopy Section at NBS, which included research groups in ultraviolet, infrared, and microwave spectroscopies. Since 1969, he has been Director of the Standard Reference Data Program at NBS. Dr. Lide is the author of over 100 papers on molecular structure and spectroscopy, free radicals, molecular lasers, and various aspects of scientific information. As Secretary General of CODATA, the Committee on Data for Science and Technology of the International Council of Scientific Unions, Dr. Lide is active in the coordination of data programs and database development in many areas of the physical sciences, biological sciences, and geosciences.

Whatever the solution, however, the age of computerized chemical data will not really be here until the number of database increases greatly.

EVALUATION OF DATA IN DATABASES

Increased competition, new regulatory thrusts, and heightened consumer concerns have combined to create an awareness of the need for better technical decisions. This in turn has placed a premium on the quality of the data upon which these decisions are reached. Whether the decision is based on knowing the concentration of a pollutant in the ppb range or in knowing thermodynamic data for predicting the yield of an industrial chemical process, the better the data, the better the decision.

In the past 20 years, several significant efforts to improve the quality of chemical data have been started. The best known of these, perhaps, is the Standard Reference Data (SRD) program at NBS whose mission is to ensure a supply of evaluated data to the scientific community. A more complete description of the program has been previously published.[3] The program accomplishes its goals in three ways: support of ongoing data centers; funding of short-term data evaluation projects; cooperative work with other groups in the government and private sectors.

Table I. NBS Standard Reference Database Series[a]

Present Titles
NBS Chemical Thermodynamics Database
NBS/EPA/NIH/MSDC Mass Spectra Database
NBS Crystal Data Identification File
Thermophysical Properties of Organic Fluid Mixtures
Thermophysical Properties of Helium
Thermophysical Properties of Fluids
To Be Released in 1985/1986
Activity Coefficients of Aqueous Electrolytes
Electron Stopping Power of Materials
X-Ray Attenuation Coefficients
JANAF Thermochemical Tables

[a] Additional evaluated databases are being planned. Also, a number of cooperative activities with other groups have been started such as the DIPPR program (Design Institute of Physical Property Data) with the AIChE, which is evaluating data for the most important industrial compounds. Other cooperative programs include the Alloy Phase Diagram Program with the American Society of Metals and Mass Spectral Data with John Wiley.

Chemical data has been a prime emphasis of the data program, and many important compilations of evaluated data have been published.[4,5] While the main thrust of the effort has centered on properties of pure compounds, in recent years significant work has been done on the properties of fluid mixtures and solid materials such as alloys, ceramics, and polymers, which are of great interest industrially.

The basic thrust of the SRD program is to examine all research results in a given area and evaluate them with respect to their quality, accuracy, and reliability. This provides the users with increased confidence in the data. Since in most cases the data user is not an expert in the experimental technique employed for generation of the data, such quality indicators greatly add to the value of research results. Because the data evaluations are increasingly being released in machine-readable forms, NBS has established the NBS Standard Reference Database series, which presently numbers six titles [with several more to be released in 1985 (Table I)].

In 1982, a major workshop was held to discuss the need for quality indicators to be associated with all databases. This meeting was sponsored jointly by NBS and the Chemical Manufacturers Association. The report[6] recommends that all databases have quality indicators and outlines several ways this can be accomplished.

ON-LINE SYSTEMS AND PERSONAL COMPUTERS

Creating databases with useful information is only half the story; getting them to the users is the other half. Over the past 6 years, the EPA/NIH Chemical Information System (CIS) has shown that the concept of a comprehensive on-line chemical data system is not only viable but also can be designed to meet a wide variety of user needs. Elsewhere in this issue, the CIS is discussed in some detail. Recently, the U.S. Government has withdrawn as the prime sponsor of the CIS, and at least two private on-line vendors have begun offering systems at approximately the same level of service. In addition, a major handbook publisher has announced the on-line availability of thousands of handbook tables starting in 1985.

This activity signals the start of the third stage in the delivery of computerized chemical data, which promises to be very exciting. The first stage was concerned with the development of computer software packages to handle chemical

structure and nomenclature in a way that allows chemists to identify compounds and reactions by automated techniques. At least four such systems are available now, including CAS Online, SANSS (Structure and Nomenclature Search System), Molecular Design, and DARC (Questel). Others are being developed. The second stage involved making their search systems available for public on-line access. The Chemical Information System (CIS), Questel, and CAS Online are the most notable examples.

The next step is to test the commercial viability of these remote-access systems and, in particular, to determine their competitiveness with systems under the control of individual users or their organizations. Incredible progress on personal computers has been and is continuing to be made. It is now possible to have at one's desk a system with all the computing power necessary for intricate searches and large calculations. Mass storage is now available in the amount needed (hundreds of megabytes). The major hurdles to handling large databases and search systems or personal computers appear to be more administrative and procedural rather than technical.

A central on-line service can maintain tens, even hundreds, of databases routinely, make necessary updates, link them together, and amortize the costs among hundreds of users. The same work is necessary for a user with a personal computer, and it is not clear that the average chemist will want to invest the time and energy needed to learn to do this work and then do it routinely, when he has the alternative of turning to an on-line system. In addition, there is the problem of being aware of all databases available and making the necessary business arrangements with each database owner.

There is a useful analogy with the procedures for maintaining collections of reference books. It is not easy, in most cases, to work out the optimum mix between private collections in each individual office, group collections, and central research libraries. Similar problems will occur with computerized collections, where on-line services offer the potential of bringing the collective capabilities (i.e. the equivalent of the central research library) into an individual's office. The problems for academic chemists with tight budget constraints are particularly sensitive.

From the point of view of the database builder, on-line services appear to offer substantial benefits. Marketing responsibility is lessened, currency and updating concerns are minimized, and business relationships are simplified. Naturally, the next decade will see a variety of approaches and services, all complicated by real economic questions such as who pays for the required investments. The only safe prediction is that computers will make it easier to get at the chemical information that is needed. The scenario by which that will happen is not yet clear.

DATABASES IN INSTRUMENTS

Two of the most pervasive processes in chemistry are the determination of what a substance is and what is happening to it. Qualitative and quantitative analysis have been important fields of chemistry from the beginning, and over the last 25 years incredible advances in chemical instrumentation have revolutionized routine and nonroutine analysis. Instrumentation has also revolutionized our way of monitoring changes to chemical systems, whether for purposes of determining reaction rates, identifying reaction pathways and intermediates, or measuring the effect of chemicals on the environment.

Already databases have made big impacts in analytical instrumentation, especially in the areas of mass spectra, infrared spectra, and X-ray powder diffraction data. In these areas, databases with information on tens of thousands of compounds have been integrated directly into instruments that allow for routine determination of the spectra or diffraction patterns of an unknown substance, followed by matching to the database for identification purposes. Several other instrumental techniques soon will have the same capability, including UV spectra, GC retention times, ESCA, and single-crystal diffraction.

In all the existing uses, a powerful software package that mimics the decision processes for interpreting the experimental values has been developed. In reality, these systems are "expert systems" (to be discussed later), which incorporate the knowledge of expert chemists in these areas of science.

One fascinating aspect is the potential for using these instruments to monitor the course of chemical reactions by monitoring the spectra or other properties of a system. Real-time chemistry becomes a possibility, in which nano- and picosecond changes of concentration can be detected and controlled by changes in experimental parameters such as temperature, acidity, etc. Short-lived reaction intermediates can be detected more easily and reaction pathways controlled.

Many researchers are already pushing ahead in these areas, but the role of databases has not yet been fully recognized. In addition to making the databases more complete, the associated software will have to be improved to speed up matching and to improve response time.

COMPUTERIZED CHEMICAL ENGINEERING

A major revolution in the design of chemical plants has taken place in the last 10–20 years. Computer models for chemical equipment design and process operation have been developed that permit sophisticated analysis of various alternatives and optimization of the design. The results have led to savings in pilot plant investments and identifying of innovative processes. As in every other type of computer modeling, the results are only as reliable as the data used. This means the chemical thermodynamics and other databases used to support the models must be as complete and accurate as possible.

The modeling software has been developed both by individual companies for in-house use only and by small groups, either university related or industry spin-offs, who market their software. Consequently, a large number of databases have been built, often for the sole purpose of supporting the process software. The result is expensive, since many groups repeat the database-building exercise, possibly inaccurate and incomplete, since different groups have access to different data sources, and inconsistent, since data sources do not always agree.

Steps are being taken to improve this situation. For example, a series of three workshops on thermodynamic databases will be sponsored by the International Union of Pure and Applied Chemistry (IUPAC) and the Committee on Data for Science and Technology (CODATA):[7] Second International IUPAC Workshop on Vapor-Liquid Equilibria in 1-Alkanol + *n*-Alkane Mixtures, 5–7 September 1985; First CODATA Symposium on Chemical Thermodynamic and Thermophysical Properties Data Bases, 9–10 September 1985; Second CODATA Symposium on Critical Evaluation and Prediction of Phase Equilibria in Multicomponent Systems, 11–13 September 1985. CODATA and IUPAC intend these meetings to stimulate relevant activities by promoting international cooperation among interested parties. The goals are "for establishing the best possible data correlation and prediction methodologies, and for the homogenization of phase equilibrium and related property data base content." The result will be better databases to support the chemical engineering design models mentioned above.

Only through such concentrated cooperative activities will definitive databases be available that can be relied upon for technical decisions. These three workshops are part of con-

tinuing efforts that will help make the needed databases available.

DATA TRENDS AND PREDICTIVE MODELING

The progress of science has always depended on the use of experimental data to deduce concepts and develop quantitative theories. The new dimension that computers can bring to this already well-organized process lies in the ability of the computer to manipulate and handle large quantities of facts and data. We will concern ourselves here with only two facets of this—extracting trends in large data sets and creating predictive models for complex phenomena that have many degrees of freedom. We will base our discussions on work done in two NBS data activities.

Large amounts of data have long been available in many fields of science, beginning with ancient records of astronomical observations and tidal charts. It has always been a challenge to digest these data and extract conceptual understanding from them. To take one field as an example, structural determinations have by now been made on approximately 100 000 crystalline compounds. These include full structural determinations, powder diffraction patterns, and cell parameters.

The NBS Crystal Data Center presently has about 60% of these data in a database, the NBS Crystal Data Identification File, which was created by combining data from several other databases. Using a sophisticated data-evaluation program called NBS*AIDS, the data center has been able to examine data for each compound for consistency and accuracy and has highlighted a number of problems, such as refinement of a structure using the wrong crystal system.[8]

Crystallographers have established a classification system based on the possible symmetry of atoms in a crystal, leading to the theoretical possibility of exactly 230 different space groups. The space group frequency for about 30 000 organic compounds has been determined with the NBS Crystal Data Identification File.[9] It has been found that 75% of the compounds reported are classified in one of only five space groups. In contrast, 29 space groups have only one entry and 35 space groups none at all. As Mighell et al.[9] have pointed out, "it may be possible to develop theories which would explain why certain space groups are rare or uninhabited, or one may be able to correlate the molecular shape, physical properties, etc. with the probability that the compound crystallizes in a given space group."

It is also possible that those compounds that represent the only occurrence of a space group may be chemically interesting or even unique, or may only represent an experimental mistake. While it certainly is possible to study such data trends by paper and pencil techniques, the existence of a computer database clearly makes work feasible that otherwise would not be done.

Another common application of computerized databases is related to extending property data into regimes where experimental measurements are missing. Predictive modeling is very important because it is obviously impossible to make every needed measurement. For complex phenomena, it is often not easy to extract the underlying physical basis for the observed measurements and express them in compact, easy-to-use form. This is especially true for properties of complex mixtures in liquid-phase and liquid–solid systems. Examples of this include thermophysical properties of fluids and corrosion phenomena.

Over the past few years, the NBS Fluid Mixtures Data Center has developed a model for the prediction of the density, viscosity, and thermal conductivity of nonpolar fluid mixtures over a broad PVT range. The model is based on the extended corresponding states approach and covers molecular weight ranges up to C_{20}.[10,11] The motivation for this work was the need to have *reliable* data for fluid mixtures at arbitrary temperature and pressure and for any mixture composition. What the computer brought to the modeling effort was the ability to calculate the properties over a wide range of parameters, compare them to experimental measurements, and assess the reliability of the calculated data. The average percent deviation for both viscosity and thermal conductivity over the range of validity was observed to be less than 8%.

The resulting model has been made available as part of the NBS Standard Reference Database Series and is commonly referred to as TRAPP.[10] This is an example of a database in the form of an interactive program which calculates reference data at desired conditions with a well-defined and well-tested model. The alternative would be mammoth tables of numbers covering all parameter ranges, an impossible task when dealing with many-component mixtures. We expect that more such databases will be developed in the future, especially for complicated multiparameter phenomena that are difficult to describe in analytic form.

EXPERT SYSTEMS

One gleam in the eyes of those interested in the use of computers to distribute chemical information is the development of *expert systems*. An expert system can be defined as a method for capturing the decision-making skills of an expert and expressing these in the medium of computer software. The crux of such software is the decision-making process and the ability to let people make use of someone else's process, especially when that someone else is preeminent in his field of science.

Much has been written recently about expert systems, and we will not go into any detail about them. It should be pointed out that much of the early practical development was in the field of chemistry; the DENDRAL Project in fact was one of the first large-scale successful examples of an expert system. Recently, Dessy[12] has edited a series of articles in *Analytical Chemistry* on expert systems that provides a good introduction to the subject for chemists.

There are, however, some significant points to be made with respect to expert systems and databases. The first is that expert systems go far beyond databases in their transmittal of chemical information. In an expert system, the emphasis is on how to use data to reach some type of technical decision. For example, a database can easily contain the infrared spectra of thousands of compounds. An expert system will contain the steps of the process to determine if the spectrum of an unknown substance is "identical" with a spectrum in the database, so that one can "positively" know the identity of the substance. Consequently, for some types of chemical applications of expert systems, reliable databases are essential.

Obviously, expert systems can make use of a database easier. The question remains whether they are necessary. Here, the issues are murky for a variety of reasons. First, it is clear that some database use will always go on without any need of an expert system; e.g., "what compounds boil at 100 °C." To the extent that databases are electronic handbooks, these uses are well understood. To the extent that the questions asked usually represent just one step in solving a chemistry-related problem, the matter is less clear. It is here that expert systems will provide the greatest benefit.

From our experience at NBS with databases, we have concluded that tables of property data or their electronic counterparts often are not enough. There is a need for software that allows intelligent and correct use of a database. In analytical chemistry, this software already has taken the form of primitive expert systems. In other areas, such as chemical process design and chemical synthesis on the computer, expert systems are just being developed.

One particularly interesting problem associated with expert

systems, that of a *good* user interface, also occurs with database use. In neither case have user interfaces been built that really are easy to use. Progress has to be made in this area.

There is no doubt that the impact of expert systems on chemistry will be huge in the future. Part of the impact clearly will be in the use of chemical databases.

SOLID MATERIALS

With the exception of diffraction data, very little work has been done on databases of the properties of solid materials. Yet data on the characterization of surfaces, catalytic properties, corrosion properties, and other areas of solid-state chemistry are very important. The Standard Reference Data program at NBS has begun work on databases of ESCA information, as well as chemical-stability diagrams for corrosion-prediction purposes. However, much more effort is needed in this area. Two examples can indicate the importance of these data. The reliability of microelectronic circuits or "chips" depends on the reactivity of the exposed surfaces to chemicals in their environment. Such microchemistry is very important, yet no organized activity to collect these data is under way. The same type of microchemistry is also important in catalysis, yet again these data are not being systematically collected, evaluated, and made available in databases.

SUMMARY AND CONCLUSIONS

We have outlined in the discussion above some of the future directions of chemical databases. Clearly, many developments will take place, and it is difficult in every case to pinpoint exactly how fast progress will be made. It is equally clear that we can anticipate chemical databases as a dynamic area of development which will eventually lead to the computer revolution in chemical information that has been predicted. Hopefully, the reader will take up some of these challenges and help make computer access to chemical information an everyday reality.

REFERENCES AND NOTES

(1) Hampel, V. E.; Bollinger, W. A.; Gaynor, C. A.; Oldani, J. J. "An Online Directory of Databases for Material Properties"; Lawrence Livermore National Laboratory: Livermore, CA, 1984; UCRL-90276 Rev. 1.
(2) Rumble, J. R., Jr. "Why Can't We Access More Numeric Data via Computers". In "Proceedings of the Fifth National Online Meeting"; Williams, M. E.; Hogan, T. H., Eds.; Learned Information Inc.: Medford, NJ, 1984; p 325.
(3) Lide, D. R., Jr. "Critical Data for Critical Needs". *Science (Washington, D.C.)* **1981**, *212*, 1334-40.
(4) "Standard Reference Data Publications 1964-1980"; Sherwood, G. B., Ed.; U.S. Department of Commerce: Washington, DC, 1981; NBS Special Publ. 612.
(5) "Standard Reference Data Publications 1981-1982 Supplement"; Sherwood, G. B., Ed.; U.S. Department of Commerce: Washington, DC, 1983.
(6) "Workshop on Data Quality Indicators—Summary Report and Recommendations"; Chemical Manufacturers Assoc.: Washington, DC, 1982.
(7) Further information on these meetings can be obtained by contacting: Dr. Henry Kehiaian, Universite Paris VII-CNRS, Institut de Topologie et de Dynamique des Systems, 1 rue Guy de la Brosse, 75005 Paris, France.
(8) Himes, V. L.; Mighell, A. D. "A Matrix Method for Lattice Symmetry Determination". *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1982**, *A38*, 748-749.
(9) Mighell, A. D.; Himes, V. L.; Rodgers, J. R. "Space Group Frequencies for Organic Compounds". *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1983**, *A39*, 737-740.
(10) Ely, J. F.; Hanley, H. J. M. "Prediction of Transport Properties. I. Viscosity of Fluids and Mixtures". *Ind. Eng. Chem. Fundam.* **1981**, *20*, 323.
(11) Ely, J. F.; Hanley, H. J. M. "Prediction of Transport Properties. II. Thermal Conductivity of Fluids and Mixtures". *Ind. Eng. Chem. Fundam.* **1983**, *22*, 90.
(12) Dessy, R. E., Ed. "Expert Systems Part I". *Anal. Chem.* **1984**, *56*, 1200A-1212A.
(13) Dessy, R. E., Ed. "Expert Systems Part II". *Anal. Chem.* **1984**, *56*, 1312A-1332A.

Data Base Development and Search Algorithms for Automated Infrared Spectral Identification

S. R. LOWRY,* D. A. HUPPLER, and C. R. ANDERSON

Nicolet Instrument Corporation, Madison, Wisconsin 53711

Received February 19, 1985

Specifications and sampling methods for infrared spectral data acquisition are presented. Two spectral search algorithms and some of their special features are described. The relationship between high-quality Fourier-transform infrared reference spectra and good search results is also discussed, and some other applications of large reference libraries are presented.

INTRODUCTION

Infrared spectroscopy has long been the method of choice for qualitative analysis of organic materials. The unique fingerprinting and identification ability provided by an infrared spectrum results from the fact that the peaks in the spectrum correspond to vibrational modes that are characteristic of the complete molecule and to other modes that are directly related to the fundamental vibrations of specific functional groups. This combination of group frequencies and the "fingerprint" region in infrared spectra has made the comparison of an unknown spectrum to a standard spectrum from a reference material a commonly accepted method for compound confirmation, not only in the laboratory but also in a court of law.

Spectroscopists have tried to improve on visual comparison techniques since the early days of infrared spectroscopy. The first methods for automatically retrieving reference spectra that were similar to an unknown involved encoding punched cards with the locations of the major peaks in a spectrum. One early system actually used a series of notches and holes whereby when a needle was inserted into the hole signifying a specific peak location in the molecule, only those cards with spectra containing the peak were captured. This manual technique was replaced by the automatic card sorting machines from the early days of computers. Both of these sorting methods resulted in a set of cards from those compounds containing the specified spectral features.[1,2]

The first computerized infrared spectral data base of significant size was the ASTM spectral file. This was basically a digitized form of the original punch cards used in the card-sorting methods. This file consists of over 100 000 infrared spectra in a binary format. In a binary format the spectrum is broken into a series of equally spaced intervals.

0095-2338/85/1625-0235$01.50/0