J. Chem. Inf. Comput. Sci. 1980, 20, 132-136

Numerical Data Indexing†

JOHN W. MURDOCK

Informatics, Inc., Rockville, Maryland 20852

Received February 8, 1980

A review and analysis are given of programs aimed at developing methodologies for indexing the numerical data that occur in the literature used by scientists and engineers. While the cost-effectiveness of data indexing has not been demonstrated, an increasing application of this technique seems inevitable in view of the growth of information and data. Recommendations focus on the principal issues requiring consideration in any future data indexing efforts and emphasize the need for improved guidelines for indexes, more cooperative programs, and, ultimately, increased standardization.

Numerical data are produced at great expense, and often in great volume. Most of these data will be analyzed only for the single purpose for which they were originally collected, even though the data have potential value for other uses. The difficulty in finding specific data, even that reported in the scientific literature, is one of the major reasons for the lack of multiple uses of data. Numerical data indexing is a method that attempts to overcome this limited use of a created resource.

Another major problem arises, however, if a researcher finds many data values for one given event in the literature, but uses only one or two of the values. A high probability exists that such data are not the best that have been reported—they may even be wrong. This is true of refereed literature as well as for nonrefereed literature such as technical reports. This can best be illustrated by the curves compiled by the Center for Information and Numerical Data Analysis and Synthesis (CINDAS) at Purdue University (Figure 1). The curves in Figure 1 show the wide range of values available from 200 articles, each article being designated by the number on the individual curve. Consider, for example, the plight of a scientist who has chosen from one article a value for the thermal conductivity of copper over a specific temperature range. Thus, one of the advantages of numerical data indexing is to make it easy for a researcher to find several papers containing numerical data of the same event so that judgment can be excercised on the accuracy of a value.

Numerical data indexing in this paper is considered to be an extension of subject indexing. It is, at a minimum, an attempt to identify, through an index, those publications that contain numerical data. It is also an attempt at a higher level of indexing to help the researcher identify the nature of these data in the publications. The cost differential, however, between indexing that merely identifies that data exist in an article and indexing that describes the data is considerable.

WHAT ARE DATA?

The title of this paper, "Numerical Data Indexing", implies that it is numbers that are being discussed, but there is considerable debate on what constitutes data. Numerical data, as discussed in this paper, generally states the magnitude of some quantity that characterizes a property of a system, a compound, or a material that is measured under specific conditions. For readers who are interested in the broader problem of data indexing, CODATA Bulletin 16[13] presents a detailed discussion on the nature of data found in the scientific literature. [All reference numbers relate to the Bibliography at the end of the paper. Not all papers in the Bibliography are mentioned in the text.]

BRIEF HISTORY OF NUMERICAL DATA INDEXING

The first activity identified that directly related to numerical data indexing as being considered in this paper was a recommendation in 1967 by the National Advisory Committee on Research in Geological Sciences to index geological data contained both in files and in other sources on a national basis, in Canada. Next, in the early 1970s, the National Science Foundation (NSF) began a program to investigate the availability of numerical data. An ICSU AB/CODATA Joint Working Group held meetings in 1973 on data flagging and tagging and published their findings and recommendations in CODATA Bulletin 19, June 1976.[14]

The National Science Foundation awarded a grant to the American Institute of Physics (AIP) in February of 1974, held a conference with the National Bureau of Standards and the Atomic Energy Commission in May of 1974, and awarded grants to the Chemical Abstracts Service (CAS) in June of 1975 and to Informatics in 1976 to review what had been accomplished and to make recommendations related to the development of numerical data indexing.

The major activity in Europe was an effort by the International Nuclear Information System (INIS) of the International Atomic Energy Agency to improve data indexing. The INIS work was first reported at the CODATA Conference in June–July 1976,[6] and was updated in a consultant's report by Vassil Gadjokov in December 1977. In May 1976, NSF gave a grant to the National Aeronautics and Space Administration (NASA) for an experiment to determine user reaction to data indexing. In this experiment, the Denver Research Institute added data indexes to the subject index in two subject categories of several consecutive issues of the *Scientific and Technical Aerospace Reports* (STAR) and *International Aerospace Abstracts* (IAA). The researchers then observed whether users reacted differently to abstracts with the enhanced indexes. There was a measurable difference in user reaction that favored the numerical index, but as of now there has been no follow-up activity.

NSF has been silent on data indexing since 1976. The only current numerical data indexing work with which the author is familiar is that being done by INIS.

Numerical data indexing was discussed in the early 70s under the concepts of *Data Flags* and *Data Tags*. The Data Flag is an indicator of the presence of numerical data in an article, while the Data Tag is a more descriptive indexing activity which characterizes the numerical data in greater depth. While these terms are useful in distinguishing the

† Presented before the Division of Chemical information, Symposium on "Techniques and Problems in Retrieval of Numerical Data", 178th National Meeting of the American Chemical Society, Washington, D.C., Sept 12, 1979.

0095-2338/80/1620-0132$01.00/0

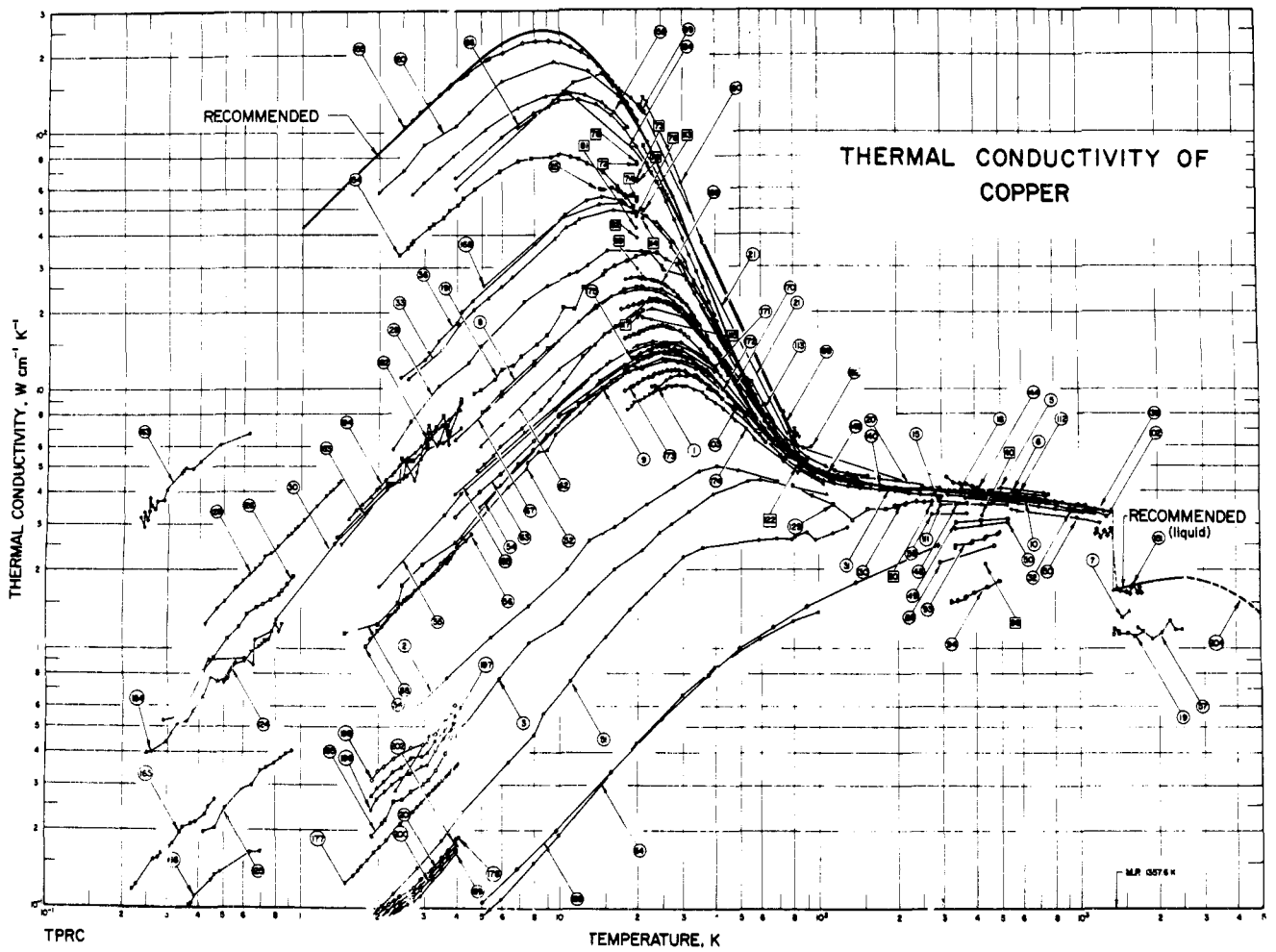

Figure 1. Thermal conductivity of copper from C. Y. Ho, R. W. Powell, and P. I. Liley, "Thermal Conductivity of the Elements: A Comprehensive Review", part of the *Journal of Physical and Chemical Reference Data*, Vol. III, Suppl., 1974.

nature of the work done and reported in abstracting literature, the term "numerical indexing", as used in this paper, subsumes both of those terms under the overall heading of numerical data indexing.

ECONOMICS

Researchers on numerical data indexing, and managers of abstracting services, quickly realized that costs of the numerical data index would be roughly equivalent to a subject index. A significant problem then is whether purchasers would be willing to pay the higher cost for an index journal, even though the addition of numerical data index terms seemed worthwhile. Given the cost of abstracting and indexing publications today, it is difficult to imagine customers being able or willing to pay nearly twice the current price in order to obtain abstracts accompanied by numerical data indexes. One possibility for minimizing cost was to include a minimum statement identifying the existence of numerical data, with the regular subject abstract, i.e., a single symbol such as the letter "n" or "d" indicating the absence or presence of data. Another cost-cutting suggestion encouraged abstracting services to merely add a phrase such as "numerical data included", to the abstract. Naturally it would be desirable to provide more than just this designation so that the reader of the abstract would have some idea of the nature of the numerical content of the indexed article, but any added complexity increases the overall indexing cost inordinately. One approach to obtain better numerical indexes at a reasonable cost was to have the author, under the guidance of an editor, prepare an index or data statement about the content of the article. This approach was explored by the American Institute of Physics.[19,20]

The increased expense of printing the added information in an abstract bulletin leads to the consideration that indexing should be developed around a highly symbolic structured approach. However, in order for indexers and users to understand the symbolic nature of the index, there would be an added cost of training that might exceed the increase in printing costs. CODATA, in Bulletin No. 19,[14] stated, however, that the fineness of grid, i.e., the depth of indexing, was one of the fundamental problems in numerical data indexing. In my opinion, a highly structured symbolic index seems to require a statement well beyond the capability of many abstractors and the willingness of users to learn the system.

STANDARDIZATION

It now appears that standardization of abstracting and indexing practices is being accepted internationally after many years of resistance. The major resistance to standards resulted from the fact that standardization had come after many different services, over several decades, had developed their own systems. Thus, standardization required the surrendering of techniques and procedures developed over many years and the accepting of changes in publications which lead to the need to educate both abstractors and users. It would seem desirable, therefore, in order to avoid later conflicts that standarization of numerical indexing practices be one of the first areas of consideration.

Variances, for example, are to be found in the choice of symbols that represent physical properties and characteristics

of measurements. For example, in the fourth interim list of IUPAC flag codes, "MP" stands for mechanical properties; in the CAS data tagging list, "MP" stands for melting point. Thus, a basic problem exists as to whether to change the vocabulary and the symbols that are well established in a discipline to terms that are multidisciplinary or that can be used as cross-disciplines. It appears at this time that the tendency of a vocabulary to develop among workers in a specific discipline will dominate attempts to standardize across several or all disciplines.

RECENT RESEARCH IN NUMERICAL DATA INDEXING

The year 1976 appeared to be the zenith for numerical data indexing. Three projects were funded by the National Science Foundation and one by the International Nuclear Information System of the International Atomic Energy Agency. The NSF-funded research was with the American Institute of Physics, Chemical Abstracts Service, and the Denver Research Institute, the latter in cooperation with NASA. The first two programs, the ones with the AIP and with CAS, were directed at obtaining a better understanding of the preparation of data indexes and of their use in the abstracting services. The latter program, done by the Denver Research Institute in cooperation with NASA, was designed to determine the impact that data statements would have on users, the precision that would satisfy users, and the change in volume of use that would result in aerospace literature. The study by the American Institute of Physics centered on author-prepared abstracts to be published in the primary literature. The program at Chemical Abstracts Service emphasized the addition of tags to index entries in the sections of Chemical Abstracts on energy. Incidentally, these additions were added to the tape service only. The work at the International Nuclear Information System emphasized the development of data statements that would be consistent with the INIS publication and group services. A detailed report of these four studies is given in the publication "Current Knowledge on Numerical Data Indexing and Possible Future Development",[21] released as a final report by the National Science Foundation in April 1978.

In December of 1977, Vassil Gadjokov developed a report on practical steps toward improving the identification of data sources for the International Nuclear Information System of the International Atomic Energy Agency. After struggling with the differences of six classification schemes and the complexity of data indexing and its expense, the author had a strong inclination to abandon the study. He continued his work, however, and put together a series of compromises for INIS that served as constraints on eventual decisions for data indexing that would make the scheme practical. The author suggested that whatever data indexing scheme might be adopted, it should not change the general pattern of document indexing in the system. It should not modify the basic structure of the INIS record; in particular, it should not prescribe either an increase of the numerical data content of the record or an introduction of data sets into it. It should not replace in-depth indexing to be carried out either in parallel or at a later date by data analysis centers and their specific subject fields. It should not lead to an increase of INIS input cost substantially above a certain reasonable figure of, say, 5%, and, finally, it should not preclude the adoption of future international standards in the field of data flagging and tagging.

The author developed the opinion that INIS was not prepared to make a major effort as yet, and that the cost of deeper data indexing could become very expensive with consistency varying from country to country, while benefits would remain controversial. A pilot project in a limited subject area that could be started immediately was recommended for INIS. If one or two national INIS centers would volunteer to carry out the experiment, a pilot project in data indexing would be proposed to include two narrow subject fields: (a) in an area with well-organized data analysis center activities, e.g., nuclear properties and reaction, data centers operating in such an area may help in evaluating the test; and (b) in an applied area where few or no data analysis centers operate, e.g., applied life sciences. In this second area results may be more difficult to evaluate but they are expected to be of interest and of some use to scientists in developing countries. It was reported that the technical proposal that had been discussed with representatives from two national INIS centers is now being implemented.

FUTURE DIRECTIONS

There is only minor direct evidence that a need exists for numerical data indexing. However, the following organizations have sensed an urgent need in the science community to have better mechanisms for locating numerical data: the American Institute of Physics, the Chemical Abstracts Service, the International Nuclear Information System, CODATA, UNESCO, ICSU/AB, the Numerical Data Advisory Board, the National Academy of Sciences, the National Bureau of Standards, the National Aeronautics and Space Administration, the National Science Foundation, and the Canada Centre for Geoscience Data. Numerical data indexing is one means of fulfilling that need—particularly with respect to reports in published literature. In spite of the current sense of frustration due to the high costs that would be involved, the work in standardization yet to be done, and the training of abstractors and users, it appears that eventually there will be an increase in the use of numerical data indexing. This conclusion is based on the premise that while the cost of numerical data indexing is high, those costs associated with technological efforts to obtain data a second time through duplicate research is much higher. There is also the problem that much of the data now being collected are from experiments that will not be repeated for years to come, such as some of the space shots.

It appears particularly important, at this time, to arrive at a quantifiable benefits measure if the increased costs associated with upgrading indexing to include numerical data is to occur. The useful quantifiable value of the market place appears to be rather unsatisfactory in the case of numerical data indexing since information services are not necessarily evaluated on the basis of market response. The tradition of free information and broad government subsidization of technical information have, to a great extent, eliminated the market place as a method of economic control. Perhaps a more satisfactory quantifiable benefit can be obtained from measuring the time saved by users of abstracting services where numerical data indexing has occurred.

In addition to costs, some of the issues which will have to be considered in developing a numerical data program are:

- the availability of expertise
- user acceptance
- standardization
- development of data indexing techniques
- the relationship of numerical indexing to other data access programs
- the education of users
- continuing research to improve indexing methods

• the role of the private, professional, and public sectors and their interrelationships

RECOMMENDATIONS

The author, as a result of preparing the publication, "Current Knowledge on Data Indexing and Possible Future Developments", for the National Science Foundation, developed ten recommendations with respect to evolving a numerical data indexing program. These recommendations are presented here without further comment:

1. The indexing of numerical data should be developed as an extension to current abstracting and indexing practices.

2. Organizations that support abstracting and indexing activities should support further studies on the production of numerical data indexing and abstracting by professional indexers and abstractors and on factors related to author-produced numerical data abstracts.

3. Organizations that are responsible for the preparation of reports and published literature should consider the requirement to have authors prepare the numerical statement and index as part of their responsibilities in the preparation of reports.

4. Government organizations responsible for the preparation of reports which contain bibliographic data sheets or technical report standard title pages, such as the DOD Form DD-1473, NTIS-35, the HEW Form OE-6000(ERIC), and similar forms, should modify the existing forms and instructions to include information on the data content of the report.

5. ANSI Z-39 committees, ASIDIC, NFAIS, EUSIDIC, ICSU/AB, and other such groups should consider modifying guidelines and standards to include the indexing and abstracting of numerical data in their existing formats and instructions. These modifications should also be reflected in similar instructions and guidelines prepared for the use of computerized information services.

6. Organizations that are considering data indexing should not wait until standards are developed but should proceed in a manner that would allow for changes in format and content. It is recommended, however, that each organization that starts data indexing should participate, where feasible, in standards-making activities so that lessons learned are disseminated rapidly.

7. Further research is recommended on computer-aided vocabulary switching.

8. Research is recommended to determine the value and use of mechanisms to indicate the presence and location of numerical data in full text that is in machine-readable form.

9. Professional societies, trade associations, and scientific unions should form committees to determine the numerical data practices and needs of their members. These organizations should also consider increasing the number of articles in their journals and the number of technical sessions at their meetings on numerical data being developed and used by their members, and on the data practices in their area of specialty.

10. Organizations that perform or support work that requires the collection data, its analysis, and presentation in reports and literature should develop guidelines and procedures that would permit effective comparisons and correlations of the data that are included in their reports with data from other sources.

REFERENCES AND NOTES

(1) American National Standards Institute, Inc., ANSI, Guidelines for Format and Production of Scientic and Technical Reports, ANSI Z39.18-1974.
(2) Battelle Columbus Laboratories. Energy R&D Data Workshop. Springfield, VA: National Technical Information Service; Nov 1974. 30 pp (Summary report of meeting held at the National Bureau of Standards, Gaithersburg, MD, May 6 and 7, 1974).
(3) Burk, C. F., Jr. Computer-Based Storage and Retrieval of Geoscience Information: Bibliography 1970–72. Ottawa, Canada: Geological Survey of Canada; 1973. 38 pp (GSC Paper 73–14).
(4) Federal Council for Science and Technology, Committee on Scientific and Technical Information (COSATI), Guidelines to Format Standards for Scientific and Technical Reports Prepared by or for the Federal Government, Dec 1968.
(5) Freeman, James E., et al. Evaluation of Numerical Data Tagging and Flagging in a Real-World Aerospace Environment (paper presented at 5th International CODATA Conference, Boulder, CO, June 28–July 1, 1976).
(6) Gadjokov, V. Data Indexing in INIS: Problems and Approaches (paper presented at 5th International CODATA Conference, Boulder, CO, June 28–July 1, 1976).
(7) Gunn, K. L., and Burk, C. F., Jr. A Decentralized, Cooperative Indexing Project: Canadian Index to Geoscience Data: Proceedings of Canadian Association Information Scientists, May 1975, Quebec, pp 243–252.
(8) Hruska, J., and Burk, C. F., Jr. Computer-Based Storage and Retrieval of Geoscience Information: Bibliography 1946–69. Ottawa, Canada: Geological Survey of Canada; 1971. 52 pp (GSC Paper 71-40).
(9) International Council of Scientific Unions, Committee on Data for Science and Technology. ICSU CODATA. Paris, France: CODATA Secretariat; n.d. 32 pp (pamphlet).
(10) International Council of Scientific Unions, Committee on Data for Science and Technology. Geological Data Files: Survey of International Activity. 30 pp (CODATA Bulletin 8, Nov 1972).
(11) International Council of Scientific Unions, Committee on Data for Science and Technology. Guide for the Presentation in the Primary Literature of Numerical Data Derived From Experiments. 6 pp (CODATA Bulletin 9, Dec 1973).
(12) International Council of Scientific Unions, Committee on Data for Science and Technology. Energy Data: Accessing and/or Retrieval. 11 pp (CODATA Bulletin 12, Sept 1974).
(13) International Council of Scientific Unions, Committee on Data for Science and Technology. Study on the Problems of Accessibility and Dissemination of Data for Science and Technology. 32 pp (CODATA Bulletin 16, Oct 1975).
(14) International Council of Scientific Unions, Committee on Data for Science and Technology. Flagging and Tagging Data. 22 pp (CODATA Bulletin 19, June 1976).
(15) International Council of Scientific Unions, Committee on Data for Science and Technology. Statement of Long-Term Program. 16 pp (CODATA Special Report, July 1975).
(16) International Council of Scientific Unions, Committee on Data for Science and Technology. Feasibility Study of a World Data Referral Centre. 39 pp (CODATA Special Report, Sept 1975).
(17) International Council of Scientific Unions, Panel on World Data Centres (Geophysical and Solar). Third Consolidated Guide to International Data Exchange Through the World Data Centres. Washington, D.C.: National Academy of Sciences; Dec 1973. 72 pp.
(18) Kottenstette, James P. Data Flagging and Tagging Experiment. Denver, CO: Denver Research Institute; Dec 20, 1976. 8 pp (Working Paper No. 2, NASA Contract NASW 2992).
(19) Lerner, Rita G., et al. Data-Descriptive Records in the Physical Sciences, Final Report. New York: American Institute of Physics; July 1976. 95 pp.
(20) Lerner, Rita G. Data Tagging in Physics (paper presented at 5th International CODATA Conference, Boulder, CO, June 28–July 1976).
(21) Murdock, John W. Current Knowledge on Numerical Data Indexing and Possible Future Developments. Final Report under Grant No. DSI76-17294, National Science Foundation. April 1978. Available NTIS.
(22) National Advisory Committee on Research in the Geological Sciences (Canada). A National System for Storage and Retrieval of Geological Data in Canada. Ottawa, Canada: Geological Survey of Canada; 1967. pp 23–24.
(23) National Research Council (U.S.). Problems of Distribution and Marketing of Machine Readable Scientific Data Bases. 12 pp (report of meeting of Numerical Data Advisory Board Ad Hoc Panel, Washington, D.C., Dec 4, 1973).
(24) Niehoff, R. T. Development of an Integrated Energy Vocabulary and the Possibilities for On-Line Subject Switching. *J. Am. Soc. Inf. Sci.* **27**, 3–17 (1976).
(25) Study of Scientific and Technical Data Activities in the United States, Vol. 1. Springfield, VA: National Technical Information Service; 1968 (AD 670606).

(26) Tate, Fred A., and Zaye, David F. Data Tagging in Information-Accessing Services (paper presented at 5th International CODATA Conference, Boulder, CO, June 28-July 1, 1976.)
(27) U.S. National Oceanic and Atmospheric Administration. Marine Geology and Geophysics Data Services and Publications, Boulder, CO: Environmental Data Service; May 1976. 11 pp.
(28) U.S. National Oceanic and Atmospheric Administration. Solar-Geophysical Data: Explanation of Data Reports. Asheville, NC: National Climatic Center; Feb 1976. 83 pp (Report No. 378, Supplement.)
(29) U.S. National Oceanic and Atmospheric Administration. User's Guide to NODC's Data Services. Washington, D.C.: Government Printing Office; Feb 1974. 72 pp.
(30) Van Olphen, Hendrik. The Numerical Data Advisory Board. *Bull. Am. Soc. Inf. Sci.* 1, 8-9, 33 (1975).

Special Features of NBS's Omnidata System Applicable to the Retrieval, Analysis, and Dissemination of Chemical Data†

BETTIJOYCE BREEN MOLINO

National Bureau of Standards, Washington, D.C. 20234

Received February 8, 1980

Omnidata is an interactive, general-purpose system for data retrieval, data analysis, and file maintenance, developed at NBS. The system allows individuals with little background in computers to search and analyze data files and prepare reports. In addition to the "typical" searching, reporting, sorting, and updating, there are roughly 30 modules providing statistical and graphical analysis, data manipulation, and file management. Many are specifically designed and have unique features to aid the chemist in the retrieval, analysis, and dissemination of data. Some of these are discussed and illustrated on files of chemical data.

Omnidata is a general-purpose system for data retrieval, data analysis, and data file maintenance. The system has been designed so that persons with little or no knowledge of computers are able to search computerized data files, do analyses on these files, and prepare ad hoc or periodic reports. Although designed with the novice in view, the system is of use to the computer professional and data-base administrator as well. Numerous utility modules provide these individuals with tools for maintaining the integrity of those data bases under their control. For the management staff, the system can provide answers—sometimes within minutes, often within the hour—to questions requiring computer processing of stored data.

Most of the existing data management systems have adequate and roughly comparable search and arithmetic capability, file definition features, and more or less flexible report generators. None of these, however, has nearly enough data analysis and data manipulation facilities for handling the numerical and alphanumeric data files in an active scientific data analysis center or in any large commercial endeavor. Using, therefore, NBS experience in designing general-purpose programs and in using a large variety of time-shared computer systems, Omnidata was designed and programmed as a modular interactive data analysis and retrieval system. It consists of 45 unique modules in addition to a main supervisory program. Program modularity has long been a hallmark of truly efficient computer programming and systems design. Indeed, in most systems the modularity is not necessarily seen by the user. Our system is quite different in this respect. The Omnidata system is as modular to the user as it is to the computer. Each operation is specific and distinct, and the user calls the required modules in turn to achieve his desired solution. In each module the user is asked to supply the requisite particulars to achieve the result required. After each module has done its work, the user has an opportunity to check the results before going on to the next operation. Such interaction with the data file is facilitated by requiring the user to perform each operation separately, and Omnidata has a number of interesting and useful ways of assisting the user in looking at the data in the file.

The various modules available provide facile tools for searching, reporting, plotting, and other graphical analysis, arithmetic operations in general, statistical analysis, file partitioning and subsequent sequential analysis on subfiles, keyword indexing of bibliographic files, flagging, coding and decoding of data items, analysis of questionnaires and surveys, and a large variety of data management and validation routines of use to both the user of the data and the file builder, or the data-base administrator. These modules are coordinated through the supervisory program, OMNIDATA, which performs such functions as obtaining the user's identification and password, assigning the designated file, reading in the header records with information such as label names and pointers, length of each record, and number of records, and recording information of use. The entire system runs equally well in demand mode, from a deck of cards in the batch mode, or in a remote batch environment.

The Omnidata system runs on Univac 1100 series machines (Sperry Rand Corp.) running under Exec 8. It is presently operative on several machines, including an international network. The programs were written in XBASIC, an extension of the BASIC language as developed by Language and Systems Development, Inc., Silver Spring, Md. This language was chosen since it was the only interactive language available at the time the system was begun, not only from the user's viewpoint, but also for the programmer to be able to write programs easily, make changes, and compile and go. XBASIC has extensive string function capabilities, the ability to read and write direct access files, and a built-in chaining feature. All of these features provide the programmer with powerful tools and, yet, a simplicity for use and change.

The Omnidata system has been described in a 288-page user's manual by Joseph Hilsenrath and Bettijoyce Breen

† Presented before the Division of Chemical Information, Symposium on "Techniques and Problems in Retrieval of Numerical Data", 178th National Meeting of the American Chemical Society, Washington, D.C., Sept 12, 1979.