(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.

entitled OMNIDATA—An Interactive System for Data Retrieval, Statistical and Graphical Analysis, and Data-Base Management. This publication is NBS Handbook 125, dated September 1978, and is for sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

Omnidata is a self-contained package of programs which can be run after invoking the XBASIC compiler or in the absolute form. In either case, the system is conversational and obtains the needed information to perform the required operations through the user's answers to simple questions. One enters the system through the supervisory module, where the user is asked to supply his account number, his password, and the name of the file with which he wishes to work. Omnidata can then perform the preliminary operations as described above: checking account number and password, assigning the file, reading in the header records, recording information of usage, etc. After this, the user is asked to:

*Type a module name and/or instructions → indicating that everything is in order to perform the first operation on the given data. At this point, a module name may be supplied, in which case the supervisory module will automatically chain to that piece of coding to perform that operation. Other responses acceptable instead of or in addition to a module name include LABELS (to print out all the given labels in the file), MODULES (to print out all available modules), and such global commands as WIDTH setting the line width, MONITOR setting the command to issue an appropriate message after performing on so many records, LIMIT controlling the portion of the file operated on, and many others. Any or all of these global switches can be changed any time the supervisory module is in command, which is between each manipulation.

At this point several modules deserve mention because of their general utility in facilitating use of the file. The DISPLAY module provides a quick look at any or all of the data items in specified records in the current file. Intended primarily for just looking at the data elements in the file, it can also be used for reporting data when the formatting is not critical. There are two modes for displaying the data in this module. Using the first mode, all the data elements for the given record are displayed, along with the label for each. Only one record at a time is printed. In the second mode, only selected elements are printed, with the labels appearing at the top of each column. The user can specify the number of records to print in this fashion, the default being ten. Other features of this module include allowing the user to skip forward or back up to a given record number.

The BROWSE module allows the user to become familiar with how the data are entered in any given vector. The module reads through the file and prints out each unique entry for the data item specified by the user, quickly answering such questions as whether citizenship is designated by a number code or by listing the country, or whether sex is indicated by M and F or by 1 and 2. After 10 unique entries are printed, the user is asked if he wishes to see more. Also, if 100 records are read without finding any new entries, the user is given the option to continue the operation. BROWSE also keeps a running tally of the frequency of occurrence of each unique entry it finds, and the user can request it be printed out before leaving the module. Applications of this module specific to a chemical file may include seeing if a formula is entered with upper and lower case letters and with spaces between elements, or seeing how many significant digits a given field allows.

Actual frequency distributions and associated histograms for data entries in up to three data vectors at a time can be produced by the TALLY module. Frequency of occurrence and percentage of occurrence are always tabulated; cumulative frequencies and cumulative percentages are included in the printout when formatting allows. Options include tallying on a given number of characters or words, starting from either the left or the right of the field, or at a given word. For example, one could tally the last word—one word from the right—in a field containing formula to determine how many compounds in the file contained, say, water of hydration. Another option is that the module recognizes any specified separator. An example of how this could be useful would be in a field of attributes—color, state, form, etc. If each attribute were separated by a given symbol, the TALLY module could recognize this and tally each one individually. The module has many options for output, too. The tallies can be printed alphabetically, in ascending value, or in descending value. The histograms can be normalized, or exact values can be printed. For a long tally, the user can specify a given number of lines to be printed and, after examining these, determine whether or not it would be worthwhile to print any more. One very useful application of this module is a tally of all the space groups represented in a given crystal system and the corresponding frequency of occurrence of compounds in each one. This is often used in the identification and verification of new crystal compounds.

It is the SEARCH module which has many unique features to aid the chemist in the retrieval and analysis of data. Through this module, specific records can be selected or rejected based on whether or not they satisfy the given criteria. Capabilities include the use of relational and Boolean logic, searching for exact matches or partial string searches (anchored or unanchored), ignoring specific characters, locating only the first record(s) that satisfy given criteria, selecting records in which the entry for one vector equals that for another, locating records having blank entries in given vectors, and taking into account the order and distance between given fragments. Let me now give examples as to how some of these features are of special interest to the chemist.

First, consider the capability of the SEARCH module to select those records in which the data in one vector are identical with those in another. Should the data contain the values for both the measured density and the experimental density, it is possible to determine through this module those records for which the two values agree. This, in fact, has been useful in working with our crystal data file here at NBS. Upon examination of the subset of the above search, however, it is evident that in some of the cases where the two density values agree, it is because they are both blank. Another feature of the SEARCH module can be used to reject those records of the subset in which the density fields are blank. Indeed, searching for blank entries in any given vector is a valuable tool for the chemist to assess where in the file there may be "missing" data. It should be noted that, in the above example, the module could look for blank fields in the subfile created by searching for matching densities; it did not need to return to the original, complete file. At all times in the module the user has three files available: the one with which he entered the module (original), the current file, and the file of the previous search. Thus, should a search strategy turn out to be too stringent and provide no hits, one can return to the previous cut of the file. Also, the SEARCH module provides limited display capabilities so that the user is able to examine data between passes.

The majority of computerized files are maintained in all upper-case characters. This may present a problem for the chemist, say, in differentiating between CO as cobalt or CO as carbon monoxide. Many chemists, then, use the convention of inserting a space between each element in a formula. Thus, cobalt would be CO, carbon monoxide would be C O, water would be H2 O, etc. The SEARCH module in Omnidata has the feature that you can search for a specific character and specify that it must be followed by a graphic, a number, or an alphabetic character. In this fashion, by specifying a search C!a the user would indicate that he wanted a C followed by an alphabetic character. He would retrieve CO, but not C O, since the space is a graphic character. Likewise, H!ng (H followed by a number or a graphic) would obtain for the user H2 O but not HE.

Another useful feature of SEARCH is the ability to search for several substrings in a given data vector in sequence. Say that we have a field in our file containing a chemical equation in the format H2 + O1 → H1 + H1O1. Searching for just H2 and H1 would give us hits regardless of which side of the equation H2 and H1 were on, whether each was a reactant or a product. By searching in sequence, however, for H2, then →, then H1, we can limit our number of records retrieved to only those in which H2 is a reactant, H1 a product.

Anchored searches can also be useful. Just a search for the letters ION would find matches if those three letters appeared anywhere in the word. On the other hand, searching for ION* will ensure that those letters occur at the beginning of the word (ionic, ionization); *ION selects entries where those letters occur at the word's end (motion, notion). Of course, an exact match for the word "ION" can also be performed.

We have discussed just a few of the many features of SEARCH which can aid the chemist. Several more modules deserve at least a brief mention. The CROSSTAB module produces a two-dimensional array of frequencies of occurrence of entries under the first label as a function of entries under the second. The output can be produced in any or all of four modes: actual values, percentage of total in the file, percentage of total in the column, or percentage of total in the row; histograms can be obtained, and the matrix can be transposed.

We discussed earlier how one can use SEARCH to look for blanks in any given data vector. If places are scattered throughout the files where data are missing, the module BLANKS would be useful. This module reads through the file and reports on all labels it found where some fields contain blanks, and tallies how many blanks in each of these fields it found.

SUMMARY reads through the entire file and, for each vector labeled as numeric, it computes the number of data items found (excluding blanks), the maximum, the minimum, the total, and the average. By examining this, the chemist can easily see if any value(s) in the file is (are) out of range. The values calculated by this module are actually stored away in the file and can be summoned for use in later operations.

Finally, the Omnidata system can prepare information for use by other processors and for computerized typesetting. The ARRAY module creates a file readable by Fortran format statements, and even informs the user of what the correct format statement should be. FIT, REGRESS, PLOT, and STATPLOTS ask simple questions, write the necessary commands, and chain to the Omnitab system for highly efficient and accurate statistical routines. The KWOC module prepares files with appropriate flags and symbols to be typeset by other programs here at NBS.

From the above brief overview of the Omnidata system, it should be evident how the system can aid the chemist in the routine handling of data. Here at NBS it has been successfully applied to chemical and physical data files and has been used to analyze data, update the file, answer inquiries, and reformat the data for dissemination and publication.

The Design of a Multipurpose File of Thermodynamic Data[†]

RANDOLPH C. WILHOIT

Thermodynamics Research Center, Texas A&M University, College Station, Texas 77843

Received February 11, 1980

Choices to be made for the design of a large, computer-readable file of thermodynamic data are identified. The effects of these choices on the file maintenance, storage efficiency, and ease of data retrieval are described. An example of a file design is given.

INTRODUCTION

Data-base design is now a well-established part of computer technology. So far, this has affected the practice of chemistry primarily through the widespread use of large files of bibliographic data. These files now furnish a familiar and indispensable bibliographic tool. The picture is changing rapidly. The increasing proliferation of computer hardware and decreasing cost of mass storage devices encourage the use of many kinds of data bases.

Although many special-purpose files of scientific data have been created and used in the past, we now appear to be on the verge of a rapid expansion in the use of large publicly accessible files of chemical data. Collections of several types of chemically related data can be seen. Information in the form of descriptive text includes biological and medical properties of chemical, synthetic and manufacturing procedures, sources of supply and economic data, and government regulatory data. Another large class of data concerns the composition and identification of materials. It includes names, formulas, and structural formula codes, as well as details of molecular and crystal structures. The most extensive file of this type is the one used in the Chemical Abstracts Registry system. Finally, there is a large and varied class of chemical information which can be conveniently represented by numbers. Examples are spectroscopic data, X-ray diffraction data, kinetic data, and the physical and engineering properties of materials.[1,2]

The most extensive collection of computer-searchable chemical data now available to the general public in the United States is the Chemical Information System.[3-5] This began several years ago as a collection of mass spectral data maintained by the National Institutes of Health. It now has twelve major components online and a number of others in various stages of development. These include all the types of chemical information listed above. The overall organization and the initial development of components is under the direction of the National Institutes of Health and the Environmental Protection Agency. Access to the file is made possible through the communications network operated by the Information

[†] 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.

0095-2338/80/1620-0138$01.00/0