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

Table I. Publicly Accessible Thermodynamic Data Files

name	substances	properties	organization
Physical Property Data Service	400 compounds and mixtures–mostly hydrocarbons	thermodynamic and transport	Institute of Chemical Engineering, England
UHDE Thermophysical Properties Program Package	350 compounds and mixtures	thermodynamic and transport	Friederich Uhde GMBH, Germany
DECHEMA Data Service	400 organic compounds and mixtures	thermodynamic and transport, phase diagrams	Deutsche Gesellschaft für Apparatewesen, Germany
Jose-Aesopp	2400 organic compounds and mixtures	thermodynamic and transport	Institute of the Union of Japanese Scientist and Engineers, Japan
Halcon Physical Properties System	500 compounds	transport	Halcon International, New York
DATABANK	50 metals and alloys	phase diagrams	Manlabs, Inc., Cambridge, MA
MTDATA	2000 inorganic compounds, metals, and alloys	heat capacity, enthalpy, thermochemical data for reactions, phase diagrams	Manlabs, Inc., Cambridge, MA; NPL, Teddington, England
THERMODATA	2000 inorganic compounds, metals and alloys	heat capacity, enthalpy, thermochemical data for reactions, phase diagrams	Centre Inter-universitaire de Calcul, France
CHEMTRAN	700 compounds and mixtures	physical and thermodynamic	Chemshare Corp., Houston, TX
TRC DATABANK	250 inorganic and organic compounds	100 thermodynamic properties	Thermodynamics Research Center, Texas A&M University, College Station, TX

Sciences Corporation. The system now has 282 subscribers.

My present intent is to describe considerations which arise in the design and use of a large, general purpose file of physical property, and more specifically thermodynamic, data. The effect of various choices on the nature of the file and the requirements of the associated software for retrieval of information will also be noted.

Table I gives some examples of thermodynamic data bases now operating. A more complete survey of existing data bases of this type has been made.[6] The primary use of a physical property data base is for the retrieval of physical property values of systems of specified composition. Computer programs used for the design and operation of chemical manufacturing and fuel technology are prolific consumers of physical property data. A compilation of such programs has recently been published.[7]

FILE SPECIFICATION

A data base which is used only to store data for direct retrieval is the electronic analog of a file cabinet. The primary considerations for the design of a file for this purpose are the general file system organization, the physical and logical attributes of the records, and the algorithms for locating data items and for file expansion and revision. The meaning of the stored data has little bearing on these questions. They have been thoroughly discussed in a number of textbooks on data-base management systems.

However, in constructing files of scientific data, it is tempting to devise means of using the logical power of computers at a higher level. Questions could be interpreted, searches made for relevant data, and actions taken on the basis of the information found to provide specific answers. We can visualize master programs of this kind, which can access files of many kinds of data. Some steps along this line have been taken within the files of the Chemical Information System. For such purposes, the information content of the records influences the design and structure of the file.

The design of a thermodynamic property data file requires decisions be made on the kinds of materials and properties to be stored, the way these data are to be represented, and the general organization of this information. These questions are closely related to the choices of algorithms for searching the file, interpreting and processing the data located, and procedures for quality control and for file expansion and revision.

No specific answers to such questions can be made until the purpose and scope of the file is known. These will be given in a statement of the design specifications. These will include a statement of which of the following kinds of data are to be retrieved from the file.

- Raw Data (results of direct experimental measurements)
- Normalized Data (raw data converted to a uniform set of units and reference states)
- Selected Data (a single best value selected for each property of each substance)
- Smoothed Data (selected data represented as smoothed functions of independent variables and usually requires interpolation or extrapolation)
- Correlation Parameters (experimental data represented by parameters characteristic of molecular structure and dynamics)

It is difficult to organize these different kinds of data into a single file. If more than one kind is needed, they probably would be stored in separate files. Selected and smoothed data are preferable for most applications. The use of molecular parameters is potentially a very compact and elegant means of generating physical property data.[17] However, this approach has severe limitations as a practical matter. In a file of raw data, we may often find more than one value listed for a particular substance–property combination. Furthermore, the accuracy and conditions of measurement may differ considerably among data from different sources. Raw and normalized data are used primarily as the starting point for producing selected values. There may be additional reasons for storing raw data, however. These may form part of the documentation behind the selected values of properties. The process of evaluation and selection of data is a slow and expensive one. Thus raw or normalized data may serve as a matter of expediency, until the selections have been made. We will proceed with the design considerations of a file containing basically selected and smoothed data only.

The performance specifications also determine the kind of information that is to be supplied by the system. As a minimum, the numerical values of thermodynamic properties of given pure compounds or mixtures of given composition are required. A provision for conversion to various sets of units may also be included. We may wish to have the system calculate thermochemical data or equilibrium constants for specified chemical reactions or, even more, to calculate the compositions of systems under specified conditions of chemical or physical equilibrium. In combination with the graphic capabilities of CRT terminals or hard-copy plotters, complete phase diagrams could be produced. The search of an inverted file could be made to identify systems which satisfy a given set of property values. This kind of information, either alone or in combination with search through other kinds of files,

could be used to identify unknown substances or to locate substances for some special application.

The following kinds of specifications define the range and scope of information to be retrieved.

- Substances—pure compounds, mixtures, or both; classes of compounds (organic, inorganic, metals, polymers, nonelectrolytes, electrolytes, etc.)
- Phases—solid, liquid, gas, metastable phases
- Kinds of properties and units
- Kinds of independent variables and range limitations (if any)
- Measures of accuracy and reliability of data

Adequate documentation of the data in the file is important. At the least, the immediate source of data should be recorded and retrieved. If a property value has been estimated, it is helpful to indentify the estimation method. If it was a measured value, references to the original publications should be given. Other possible documentation records are the date of entry into the file, the name of the compiler, and notes describing the selection or smoothing procedure.

The choice of properties to be recovered is the principal factor that will determine the uses to be made of the file. These will probably include phase equilibrium data for pure compounds, such as vapor pressure, melting and boiling points, and enthalpies of phase transitions. Properties of single phases will include density and volume, enthalpy and entropy, and heat capacity. Thermochemical properties are enthalpy and Gibbs energy of formation for various phases, and heats of combustion. Many of these properties could also be given for mixtures. Additional mixture properties are solubility and composition of phases at equilibrium, enthalpy and volume changes on mixing, Henry's law constants, and activity coefficients. Very often transport data, such as viscosity, thermal conductivity, and diffusion coefficents, are included along with the thermodynamic data. About 100 pure compound properties having significant practical applications can be listed. Mixtures of nonelectrolytes would increase the list by 30 to 40, and electrolytes would add another 10 to 20.

CHARACTERISTICS OF THERMODYNAMIC DATA

Certain characteristics of thermodynamic data require special attention in the design of a data file. Most properties are functions of one or more independent variables. The number of independent variables for the properties of a particular system is given by the famous Gibbs phase rule: $F = C - P + 2$. Temperature and pressure are the most common independent variables. Composition variables are required for mixtures. Alternative sets of independent variables such as density, volume, or entropy are convenient for some uses.

Thermodynamic properties are not independent of each other but are connected by a network of mathematical formulas, ultimately based on the three laws of thermodynamics. Many such relationships exist among the thermodynamic properties of a particular compound or mixture and among the thermochemical properties of certain groups of compounds. It is important to maintain the correct internal consistency among any collection of property values. Any discrepancies may be magnified considerably in subsequent calculations. The continuing appearance of new measurements means that frequent revisions and extensions of the file will be required. The requirements of internal consistency may cause the revision of a single value to propagate through many other values. Changes in fundamental physical constants, atomic weights, definitions of the practical temperature scale, or certain key values may have an extensive effect. The maintenance of internal consistency in printed tables through a series of revisions is very troublesome. It is conceivable that computer programs could be written to adjust automatically the contents

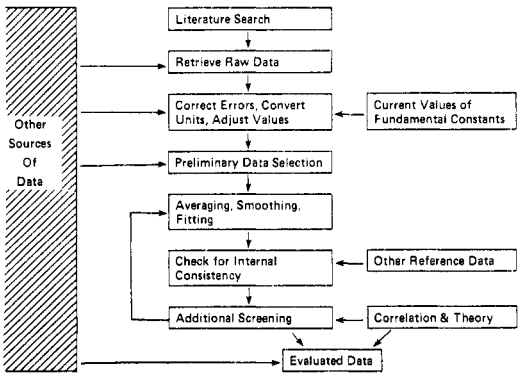

Figure 1. Steps for the generation of evaluated and smoothed data.

of a file to maintain internal consistency while accommodating new data. Some procedures of this kind have been developed for enthalpies of formation of a set of interacting compounds,[8] but algorithms for the entire set of thermodynamic properties would be very complicated.

There is a large variation in the quantity and quality of information available for different compounds. It ranges from the voluminous data published for compounds such as water, oxygen, methane, and carbon dioxide to nothing more than one or two property values measured for the majority of known compounds. If the properties were to be stored directly in the form of a substance–property matrix, a large multipurpose file would be a very sparse matrix.

The operation of a large property data bank requires a source of selected data. The process of evaluation and selection is a demanding and complicated one. Figure 1 outlines the steps involved. At the present time this would probably be done independently of the film maintenance,[9] although eventually certain aspects of the selection process could be automated. For the past 15 years the Office of Standard Reference Data of the National Bureau of Standards[10] has coordinated data as an outlet for the products of many data evaluation centers. The CODATA organization coordinates data evaluation activities on the international level.[11]

FORMS OF DATA REPRESENTATION

The simplest way to build a data file is to store directly the numerical values of the properties given in the performance specifications. Properties which are functions of one or more independent variables could be stored as tables. Values corresponding to intermediate values of the arguments could be obtained by interpolation. The storage of 100 properties used in engineering calcuations, along with the values of the independent variables and uncertainties, might contain 15 000 or more numbers. An extensive file at present might contain 2000 to 4000 pure compounds. Thus, an allowance for the storage of 30 million or more numbers would be made to accommodate a file of this size as a direct substance–property matrix. If the data were based on experimentally derived measurements, only a small fraction, say 5–10%, of the memory locations would be occupied with data. The occupancy rate could be increased considerably by use of estimated data. Files of this size are within the capability of large central computer installations, but they are expensive to maintain. The memory requirements could be reduced by the use of linked lists or some other such strategy, but since random access to the data is required, these would complicate the retrieval process. Considering the low rate of use of many of the numbers in such a file, this is not an attractive or economical way to store large amounts of thermodynamic data. The number of binary mixtures which can be formed from n compounds is $^1/_2 n(n-1)$. The direct storage of the properties

of more than a very small fraction of the mixtures of potential interest is totally impractical.

Most tables of smoothed values of properties would be calculated from some empirical function of the independent variables. Rather than storing the property values in the file, the values of the parameters in the smoothing equations could be stored. Then the values of the properties, or of their derivatives or integrals, could be calculated as needed at retrieval time. A set of suitable functions for representing the various properties would then be selected. The file would contain a code to call the appropriate subroutine for a particular function, for the values of the upper and lower limits of the independent variables, and for the values of the parameters in the function. A large variety of functions for any one property could be easily accommodated.

CHOICE OF PROPERTIES

The size of the file could be further reduced by taking advantage of the thermodynamic relationships among properties. The general principle that the change in any variable of state depends only on the initial and final states of the system gives rise to many such relationships. Hess's law for enthalpies of chemical reactions is an example. A few simple examples of other kinds of equations are:

$$\Delta H = \Delta U - P\Delta V \quad \text{(at constant } P\text{)}$$

$$\Delta G = \Delta H - T\Delta S \quad \text{(at constant } T\text{)}$$

$$\Delta G^\circ = -RT \ln K$$

ΔG° is the change in Gibbs energy for a chemical reaction in the standard state. Examples of some equations involving derivatives are:

$$C_p = (\partial H/\partial T)_p$$

$$dP/dT = \Delta H/T\Delta V$$

$$\left(\frac{\partial(\Delta G/T)}{\partial T}\right)_P = -\frac{\Delta H}{T^2}$$

$$(\partial H/\partial P)_T = V - (\partial V/\partial T)_P T$$

$$(\partial G^E/\partial n_i)_{T,P} = RT \ln \gamma_i$$

By the judicious use of relationships of this kind, it is possible to calculate a large number of thermodynamic properties from a small initial set. In the limit, all of the equilibrium thermodynamic and thermochemical properties of any pure compound or mixture can be calculated from an equation of state written in terms of a set of natural independent variables such as S, V, U, n_i; A, T, V, n_i; or G, T, P, n_i, combined with an appropriate choice of reference states.

A file consisting of parameters in one of the fundamental equations of state for the various substances would be an elegant solution to the storage problem, but it is not practical now. Functions of sufficient range and accuracy have not been found for any of the fundamental equations. Even if such functions were known, sufficient experimental data to evaluate the parameters are available for only a few compounds. Finally the amount of computation required to calculate many common properties from a fundamental equation is prohibitive for routine repetitive purposes.

Other more tractable sets of nonredundant properties can be selected which can serve as the starting point for the calculation of all or part of the possible thermodynamic properties. The members of a nonredundant set are mathematically independent of each other. Storage of a nonredundant set of properties in the form of parameters in functions has several advantages over the direct storage of property values. The size of the file can be made much smaller than the equivalent one based on direct storage. Furthermore, all of the properties calculated from the nonredundant set are automatically internally consistent. Many kinds of properties could be calculated from the nonredundant set, and the kind could be changed at times, as required, without altering the file contents.

The demand for thermodynamic property data will probably always exceed what can be obtained from experimental measurements. This can be met only by calculating or estimating properties by extrathermodynamic methods. Such methods range from those based on rigorous statistical mechanics to those based entirely on empirical observations. The most successful use of statistical mechanics is the calculation of properties of ideal gases from molecular properties. For some years steady progress has been made in the use of theoretical models for deriving P–V–T-type equations of state. The chemical engineering literature is replete with examples of empirical relationships among properties.[12] Several kinds of properties can be successfully estimated as a sum of bond or group contributions.[13-17]

Estimated properties can be merged with experimental data for input to a data file without any special provision other than an appropriate tagging. In some cases the estimated value might be more accurate than the measured one. Alternatively, estimation procedures could be incorporated into the retrieval software and invoked whenever a needed item was missing from the data file. Both methods could be used, depending on the particular estimation method. For example, by the use of group contributions, properties of a very large number of compounds and mixtures can be calculated from a small set of group values. Thus it would be more efficient to calculate such properties as needed rather than storing them in the file. The group contribution values could be stored in an auxiliary file. Empirical methods are restricted to certain classes of compounds or ranges of independent variables. If empirical calculations are imbedded in the retrieval programs, some safeguards should also be included to prevent inappropriate use.

An extensive file of mixture properties must rely heavily on thermodynamic calculations and empirical procedures since directly measured data are unavailable for many mixtures of interest.

First of all, the set of properties selected for file storage must be sufficient to generate all the properties called for in the performance specification. The calculations should be reasonably simple, especially for the more commonly used properties. Finally they should be selected so as to utilize as fully as possible the available experimental or estimated data. The following kinds of properties are commonly measured: vapor pressure; temperature and enthalpy changes for phases at equilibrium; density, volume, heat capacity and enthalpy of single phases in equilibrium with a second phase; critical constants; enthalpy changes for chemical reactions; equilibrium constants; and electrical cell potentials. Additional properties often measured for mixtures are volume and enthalpy changes on mixing and composition of phases at equilibrium. Some other experimental properties which are measured less often are volume, heat capacity and enthalpy of single phases as functions of temperature, pressure, and composition; velocity of sound; Joule–Thompson coefficient; and adiabatic compressibility.

The following properties are often tabulated because of their utility in thermodynamic calculations but are usually derived from the experimental data: enthalpy, volume, entropy, fugacity, Gibbs energy and Helmholtz energy of single phases as functions of temperature and pressure; thermodynamic properties of ideal gases; enthalpy and Gibbs energy of formation of compounds and components of mixtures in the standard state. Additional derived properties of mixtures are

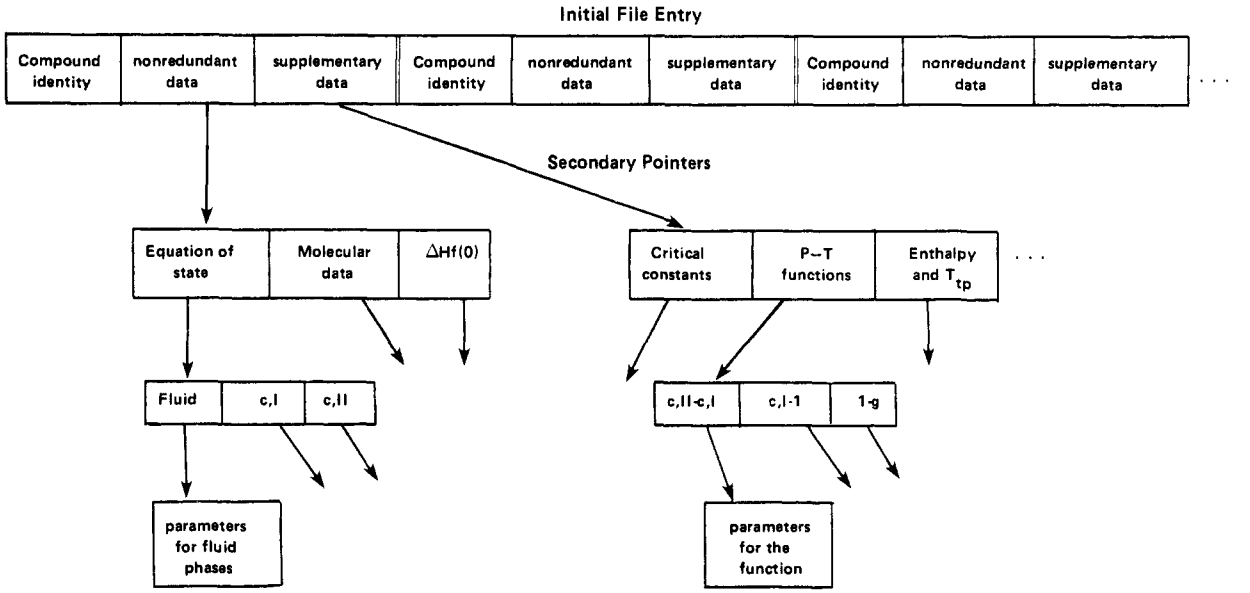

Figure 2. Schematic diagram of data organization in a thermodynamic data file.

excess thermodynamic functions, activity coefficients, and osmotic coefficients. One solution is to create several files for different classes or types of compounds with different selections of stored properties. However, the number of combinations of available data is so large that a large number of such files would be needed to be very effective. This would then complicate the file maintenance and retrieval procedures to an intolerable extent.

EXAMPLE OF FILE DESIGN

The following approach to a pure compound file selects a reasonable fundamental, but practical, nonredundant set of properties. These are supplemented with some additional properties which are often measured and used in calculations.

Nonredundant Set—An equation of state of the type $f(P,V,T) = 0$ for the fluid phases, and for each thermodynamically stable solid phase; molecular parameters for calculating ideal gas partition functions; enthalpy of formation of the crystal at 0 K.

Supplementary Set—Critical constants; functions between pressure and temperature along phase boundaries and values of the triple points; enthalpies of phase changes at the triple points; an equation for the ideal gas heat capacity as a function of temperature; volumes and heat capacities of condensed phases along the phase boundaries; values of enthalpy of formation, entropy and enthalpy relative to 0 K for the condensed phase and the ideal gas at 25 °C.

In principle, all of the properties in the supplementary set can be calculated from those in the nonredundant set, although such calculations may be quite lengthy. Whenever sufficient data exist to calculate the complete set of nonredundant properties, the supplementary properties are calculated from them and stored in the file. Most of the properties requested for retrieval from the file would be calculated from the supplementary set since such calculations for the common properties are fairly simple. When necessary, recourse could be had to the nonredundant set. For this case, the calculated properties would all be internally consistent.

More often the available data are not sufficient to determine a complete set of nonredundant properties. In this case, as many of the nonredundant properties as possible would be stored, and as many of the supplementary properties as could be calculated from them would also be stored. Any additional supplementary properties which could be obtained by measurement or estimation would also be stored. Unless special tests are made, these may not be internally consistent.

This scheme is very flexible and would be compatible with a wide range of type and amount of experimental or estimated data. Notice that there may be more than one way of calculating a particular property from the data stored in the file. In such cases, the software should include algorithms for deciding which path to use. When the nonredundant data are complete, any valid procedure would give the same result. The simplest one would then be selected. When the nonredundant set is not complete, inconsistencies may cause different calculation paths to give different results. Ideally the choice would then be based on an analysis of the propagation of errors through the different procedures.

In order to accommodate varying amounts of data without wasting storage, the contents can be organized in a tree-like structure. This is illustrated in Figure 2. The initial entry to the file consists of a series of compound records. Each record contains an identification code followed by two pointers: one for the nonredundant set of properties and one for the supplementary set. An appropriate code will signal the absence of either or both of these sets for a particular compound. The primary pointers will locate a series of secondary pointers which will either locate the appropriate parameters and auxiliary data or, as in the case of phase equilibria, will locate additional pointers for the various phase combinations.

MEASURES OF UNCERTAINTY

Few, if any, of the existing thermodynamic data banks contain quantitative information on the accuracy and reliability of the data they contain. This is a serious deficiency since tolerances in the retrieved data are very critical for some applications. Furthermore, the method of calculating certain properties from those in the file may be dictated by the effect of data errors on the calculation procedure. The calculation of the uncertainties in the retrieved data requires three steps. First, a value of the uncertainties in the original observed or estimated data must be obtained. Elaborate mathematical procedures have been devised for calculating uncertainty estimates for the effect of random errors subject to certain conditions. However, uncertainties in experimental data are dependent on many subtle factors not suitable for statistical analysis, and, in the end, the estimate must rest on the subjective judgment of an experienced and impartial evaluator. In the more general context, it is even difficult to satisfactorily define what is meant by uncertainty in data. The second step

consists of calculating the uncertainties in the smoothed and selected data stored in the file. In general, the property values and their uncertainties as well are functions of the independent variables of the system. Although this is a mathematical problem, no general solution considering the functional dependence among thermodynamic variables has been found. The final step is to calculate the uncertainty in the values of properties retrieved from the file. This is a trivial problem if the retrieved property is identical with a stored property, but if the retrieved property is calculated by combining several stored properties, especially if derivatives or integrals are involved, it can be a very difficult problem. This whole subject demands much additional work and thought.

REFERENCES AND NOTES

(1) F. V. Wetzler, "Data Banks R and D", *Res./Dev.*, **28**(6), 54-64 (1977).
(2) R. W. Counts, Y. S. Touloukian, and J. W. Phillips, "A New Dimension to Numerical Data Banks", *Res./Dev.*, **27**(1), 36-38 (1976).
(3) S. R. Heller, G. W. A. Milne, and R. J. Feldman, "A Computer-Based Chemical Information System", *Science*, **195**, 253-369 (1977).
(4) S. R. Heller and G. W. A. Milne, "The NIH/EPA Chemical Information System", a chapter in "Retrieval of Medicinal Chemical Information", M. Milne and J. Howe, Ed., ACS Symposium Series No. 84, American Chemical Society, Washington, D.C., 1978, Chapt 10.
(5) G. W. A. Milne, S. R. Heller, A. E. Fein, E. F. Free, R. G. Marquart, J. A. McGill, J. A. Miller, and D. S. Spiers, "The NIH-EPA Structure and Nomenclature Search System", *J. Chem. Inf. Comput. Sci.*, **18**, 181-186 (1978).
(6) L. M. Rose, "A Survey of Available Physical Property Data Banks", *ACHEMA Jahrb.*, I, 32 (1977/79). A description of 23 data banks collected by the Working Party for Information and Documentation of the European Federation of Chemical Engineers.
(7) J. N. Peterson, C.-C. Chen, and L. B. Evans, "Computer Programs for Chemical Engineers: 1978—Part 1", *Chem. Eng. N.Y.* 85, 145-154 (June 5, 1978); Part 2, *ibid.*, 69-82 (July 3, 1978); Part 3, *ibid.*, 79-86 (July 31, 1978); Part 4, *ibid.*, 107-115 (Aug 28, 1978).
(8) D. Garvin, V. B. Parker, D. D. Wagman, and W. H. Evans, "A Combined Least Sums and Least Squares Approach to the Evaluation of Thermodynamic Data Networks", Interim Report, Office of Standard Reference Data, NBSIR 76-1147, July 1976.
(9) W. H. Evans and D. Garvin, "The Evaluator Versus the Chemical Literature", *J. Chem. Doc.*, **10**, 147-150 (1970).
(10) "Critical Evaluation of Data in the Physical Sciences—A Status Report on the National Standard Reference Data Systems", *Natl. Bur. Stand. (U.S.), Tech. Note*, No. 947, 77-600016 (Jan 1977).
(11) N. Kurti, "Capture, Evaluation and Storage of Data—As Seen By CODATA", *CODATA Newslett.*, 8 (Sept 1977).
(12) R. C. Reed, J. M. Prausnitz, and T. K. Sherwood, "The Properties of Gases and Liquids", McGraw-Hill Book Co., New York, 1977.
(13) G. R. Somayajulu and B. J. Zwolinski, "Generalized Treatment of Alkanes—Part 2", *J. Chem. Soc., Faraday Trans.*, **68**, 1971-1987 (1972); "Part 4: Atomic Additivity Applications to Substituted Alkanes", *ibid.*, **70**, 973-993 (1974); "Part 5: Branching and Buttressing Effects", *ibid.*, **72**, 2213-2234 (1976).
(14) G. R. Somayajulu and B. J. Zwolinski, "Generalized Treatment of Aromatic Hydrocarbons. Part 1. Triatomic Additivity Applications to Parent Aromatic Hydrocarbons", *J. Chem. Soc. Faraday Trans. 2*, **70**, 1928-1941 (1974).
(15) S. W. Benson, "Thermochemical Kinetics", 2nd ed., Wiley, New York, 1976, Chapt 11.
(16) A. Fredenslund, G. Gmehling, and P. Rosmussen, "Vapor-Liquid Equilibria Using UNIFAC. A Group Contribution Method", Elsevier, New York, 1977.
(17) Y. Yoneda, "A Proposal of an Estimation and Retrieval System EROICA for Physical Properties of Organic Compounds by Chemo Inputs", in "Information Chemistry", S. Fujiwaya and H. B. Mark, Ed., University of Tokyo Press, Tokyo, 1975, pp 239-253.

Problems in Physical Property Data Retrieval†

DONALD T. HAWKINS

Bell Laboratories, Murray Hill, New Jersey 07974

Received February 11, 1980

Problems in the retrieval of numerical data on physical properties are reviewed. Data compilations are valuable, but often do not exist for the properties of interest. Those that are available are usually arranged by substance, which causes problems when the substance has a variety of names. Finding a substance which fits a given set of properties is not possible with today's handbooks; an "online handbook", containing numerical data on properties of substances and searchable online, would allow such a question to be answered. On another level, two projects which have been undertaken at Bell Laboratories to help staff members obtain needed numerical data are an index to a list of compilations held by the National Bureau of Standards and a pathfinder to numerical values of the properties of silicon.

INTRODUCTION

The scientist or engineer who wishes to obtain and use numerical data faces a monumental challenge. Numeric data are widely scattered in the literature; data compilations and Information Analysis Centers (IAC's) exist only for some fields. This paper discusses some of the problems faced by an information retrieval service in searching for and retrieving numeric data. Although the discussion here is limited to data on physical properties, the general situation is the same for many fields. Experiences related here are those gained at Bell Laboratories, a large and diversified information community; they are by no means atypical.

Bell Laboratories, the research and development unit of the Bell System, has more than 11 000 technical personnel, many of whom have a potential need for numeric physical property data. The numeric data needs of Bell Laboratories are very wide, and many technical staff members are able to satisfy their own needs. One way a scientist or engineer can satisfy this need is to collect the relevant handbooks and data compilations, which is no small task. One metallurgist at Bell Labs, who serves as a consultant in his field, has such a collection. He has acquired handbooks of data on metals and alloys and related materials. His collection comprises 135 volumes–enough to fill several bookshelves! Often, many of these volumes must be consulted when a particular numeric value is wanted. Each volume is organized differently; frequently the values are stated in different units, at varying temperatures; for a wide range of experimental conditions, and so on. The location and comparison of the data are therefore difficult and time-consuming tasks. When data on a particular substance are located, one may find that the parameters under which the value was measured are too far removed from those

† 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.

0095-2338/80/1620-0143$01.00/0