International Journal of Thermophysics, Vol. 20, No. 1, 1999

Future Directions for Data Compilations[1]

R. C. Wilhoit[2,3] and K. N. Marsh[4]

As the world's supply of thermophysical property data that emerges from the laboratory increases, users of data become more dependent on evaluated compilations. However, the cost of producing and maintaining such compilations by traditional methods is becoming prohibitively expensive. The traditional compiler searches the literature, extracts, evaluates, and analyzes pertinent data and gathers it into a document or electronic database that reflects the state of knowledge of a particular subject at a particular time. Because clef the inherent time lag, it never catches up to the current state. The attempt to catch up requires that the whole procedure be repeated at intervals, with greater cost for each cycle. A more cost-effective procedure, called dynamic compilation is described. Here the user produces a compilation to-order at the time of need. It uses a suitable archive of experimental data maintained up-to-date, an automated procedure for extracting and selecting the best pertinent data, and procedures for fitting the pieces to suitable models that furnishes parameters for internally consistent data sets. With proper design of components this procedure is more economical than and superior to the traditional static compilations.

KEY WORDS: compilations; database; data selection; data uncertainty; thermodynamic properties; thermophysical properties.

1. INTRODUCTION

The world's store of thermophysical property data has doubled about every 10 years for the past century and a half. As a consequence isolated reports of investigation are becoming less accessible. Users increasingly depend on

[1] Paper presented at the Thirteenth Symposium on Thermophysical Properties, June 22–27, 1997, Boulder, Colorado, U.S.A.

[2] Thermodynamics Research Center, Texas A&M University, College Station, Texas 77843-3111, U.S.A.

[3] To whom correspondence should be addressed.

[4] Department of Chemical and Process Engineering, University of Canterbury, Private Bag 4800, Christchurch, New Zealand.

0195-928X/99/0100-0247$16.00/0 © 1999 Plenum Publishing Corporation

evaluated compilations of data based on all current information. Steps in preparing traditional compilations of numerical thermophysical property data are as follows.

1. Search the literature.
2. Read and interpret documents, identify pertinent information.
3. Extract the data into some intermediate form.
4. Evaluate the data, assign uncertainties.
5. Compare, adjust for internal consistency, select and smooth the data.
6. Prepare the final output.

This procedure generates static pre-packaged compilations which capture the state of knowledge of the subject at the time Step 1 is completed. However, the accumulation of new information requires that revisions be issued at certain intervals. Although later revisions can benefit from earlier data collections, merging new data with the old becomes increasingly more difficult. Steps 1–5 generate intermediate results. They generally disappear after each cycle. Much of this must then be repeated the next time. This is especially true when later revisions are done by different authors. Since it is difficult to predict what people may need very far in advance, much of this effort is seldom and maybe never used before the next revision cycle. Advances in data processing and communication technology help with mechanical aspects of the tasks, but for the foreseeable future, the intellectual aspects must be done by humans.

2. DYNAMIC COMPILATIONS

In order to keep the effort spent on compiling data in the future within acceptable bounds, some changes in traditional thinking are required. These changes should promote better use of modern information technology and reduce the duplication inherent in traditional procedures. However, the most radical change is to abandon most prepackaged compilations. Instead users should be able to generate compilations, to their particular specifications, as needed. They should reflect the state of knowledge at that time. We can call these compilations-on-the-fly. They are dynamic rather than static.

Dynamic compilations require two major components. The first is a suitable electronic archive of experimental thermophysical property data accessible to all users. The archive should contain all available relevant data along with the metadata to give them meaning. The second component is computer software which can accept user requests, locate the relevant data in the archive, and carry out the equivalent of Steps 4–6 in the traditional static compilation production.

The creation and maintenance of the archive requires a cooperative international effort. Once the bulk of the world' s data is collected, incorporating new information will not be overwhelming. Convenient mechanisms for entering data into the archive and extracting data from it are required. They will be made easier by the fact that more authors are sending manuscript to journals in electronic form, and electronic versions of journals are becoming more common. A method of quality control and data validation is required. Means of avoiding excessive duplication of data collection are needed. Financial procedures for giving credit for contributions and making charges for extractions are to be developed. User costs should be modest.

3. THERMOPHYSICAL PROPERTY ARCHIVE

An archive capable of supporting dynamic compilations should have the following characteristics.

1. It should use modern database technology, including indexing of records, remote access, data validation, and backup and recovery. It should run on various platforms.
2. It should be designed for multiple uses. It should accommodate a wide variety of systems, properties and states. It should accommodate most kinds of published data within the defined scope of the collection.
3. The principal purpose of the archive is to store direct experimental values of properties of defined systems. These should be linked to the original published reports. They should be recorded in a manner similar to the way they were presented in the report. If other kinds of data are available, such as those based on selections of previous literature or calculated partially or completely from correlations or theory, they should be so identified.
4. The numerical values should be accompanied by essential metadata. These metadata are required to give meaning to the numbers. They include identification of the chemical components of the system, of the phase or phases present, and of the property represented. Values of the appropriate state variables and their identity are to be given explicitly or specified by appropriate constraints. The number of state variables for equilibrium thermodynamic properties is given by the Gibbs phase rule. Units should

either be identified or values should be converted to a common set of units The essential metadata should be recorded in a formal manner so that they can be recognized and interpreted by computer codes and used as a basis for search and retrieval.

5. Sufficient additional metadata should be given as a basis for the automated evaluation and selection of data values. These include method of presentation in the original document, description of samples used in the measurements, measurement techniques, and estimates of data reliability.
6. Information on data validation and pedigree should be included. These identify the person or organization contributing the data to the archive, the date of incorporation, and the source of data if it is not the original report. An indication of whether the value has been checked for accuracy should be shown. A flag should be included to show whether errors in the original report were corrected.

Each of these characteristics leads to a series of policy decisions. Item 3 emphasizes that the chief purpose of the archive is to capture original experimental data. Thus, directly measured values should be retained whenever possible. Sometimes smoothed tabulated data values are reported, and sometimes analytic functions or graphs are given. If these are also included in the archive they should be appropriately identified. Sometimes only the smoothed or fitted values are given in the original report. Values of properties derived from observed data may also be given and included in the archive as well. If so, they should also be appropriately identified. Certain kinds of properties, such as virial coefficients, activity coefficients, thermal properties of ideal gases, and excess properties of mixtures are nearly always derived from other measurements. If these are included, the original measurements should also be included whenever possible. Sometimes properties are derived from a combination of measurements by the investigator and other data either from the existing literature or from correlation or theory. If these data are included in the archive, the method of generation should be identified.

Item 4 requires that the state of the system whose property is given be well defined. Some properties, such as heat capacity, density, or conditions for phase equilibrium refer to a particular state of a system. Some properties represent a difference between two states. Internal energy and related properties such as enthalpy and Gibbs energy always represent a difference. In such cases both the initial and final states should be defined by the metadata.

4. MEASURES OF DATA RELIABILITY

Item 5 includes an estimate of the reliability of the data values. This is a complicated and controversial topic. Some data collections bypass the question by ignoring it. However, if the archive is to serve as the basis of automatic production of dynamic compilations, some measure of data reliability is required. This may take several forms.

1. A quality rating symbols, such as A, B, C,..., may be assigned to indicate an overall quality of the data values. It is a simple compact method of indicating quality. However, it is practical only for a very limited set of properties, systems, and states. Rules for assigning universally meaningful quality codes to a wide range of conditions would be very complicated. Furthermore, quality rating codes cannot be reliably converted to numerical measures of uncertainties.
2. A numerical estimate of uncertainty is often expressed as a fraction (or percent) of the property value. This is a convenient and effective method for many kinds of data. It is not suitable for properties whose values range from negative, through zero, to positive values. Examples are temperatures on the Celsius or Fahrenheit scales, virial coefficients, Joule–Thompson coefficients, thermochemical properties, and some other properties representing differences between states.
3. An uncertainty may be expressed as an additive bias to the property values. It is then expressed in the same units as the associated property. The addition of this bias to and subtraction from the property produce range of values which includes the "true" values within some probability. This method has universal application. It can easily be converted to a fractional form or to a quality rating if defining rules have been formulated. Estimates of uncertainty are used in two ways. First they are used to select data sets to be used in an evaluation from among larger sets of redundant data. Redundant data exist when more than one value of a property of a system at the same, or nearly the same, state is available from different sources. It also exists when functional relations exist among sets of data from different sources. Secondly, they are used to generate weighting factors for averaging and smoothing of data. The selection of the initial data set may also be viewed as an assignment of weighting factors. In this view, rejected data are given a zero weight. For these purposes only the relative values of the uncertainties are significant. Measures of data uncertainty may also be used to derive a tolerance or reliability estimate

in the final selected and smoothed values. The individual magnitudes of the uncertainties are significant for this purpose.

A general discussion of uncertainties cannot be undertaken here. A useful review has been published by IUPAC Commission 1.2 on Thermodynamics [1]. A few comments are in order. Uncertainties are not the same as errors in the data. In most cases errors are unknown. The uncertainties furnish a basis for the selection and smoothing of data. Thus, they should reflect all sources of error in the reported value. They are not necessarily the same as measures of precision such as the standard deviation for a series of repeat measurements. Precision measures are useful as a lower limit to the assigned uncertainties. A major problem in creating a universal archive is to reach an agreement on the meaning of the assigned uncertainties.

5. DATA SELECTION ALGORITHM

The weighted mean of a series of values, x_i, is expressed as

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \tag{1}$$

in which weight factors, w_i, may be proportional to $1/u_i^2$, and u_i is the uncertainty. Setting w_i to zero for values of u_i above some cutoff limit has the effect of selecting the data with uncertainties below this limit.

The selection of a data set from an initial set whose values are functions of one or more independent variables depends on both the estimated uncertainties and on the distribution of the data values with respect to the independent variables. The goal is to be more selective in highly populated regions than in sparsely populated ones. Thus, the uncertainty of each data point is to be compared to uncertainties of neighboring points. The selection criterion for data point i associated with independent variable, v_i consists of the following steps. First calculate a weighting factor, z_{ij}, for each j other than i.

$$z_{ij} = \exp(-q\,|v_j - v_i|) \tag{2}$$

Then,

$$x_1 = \sum_{j \neq i} z_{ij} \tag{3}$$

$$x_2 = \sum_{j \neq i} u_j z_{ij} \tag{4}$$

$$y = u_i x_1 / x_2 \tag{5}$$

Accept point *i* if $y \leqslant d/\sqrt{x_1}$, reject it otherwise. Parameter *d* sets the level of rejection. The denominator causes the criterion to be more selective in groups of closely spaced points. Parameter *q* determines how the weighting of points other than *i* depends on their distance from point *i*. They are adjusted to optimize the selection for particular kinds of data. Generally, it is helpful to have the value of *q* depend on the range of *v* included in the set. Let Δv be this range. Then, for example,

$$q = a[1 + (\Delta v/b)^2]^{1/2}/\Delta v \tag{6}$$

where *a* and *b* are constant parameters. In order to be more selective for larger data sets, *d* can be calculated from the function,

$$d = c/\log(1 + n) \tag{7}$$

where *n* is the number of data points in the set. This procedure has been used to automatically select sets of density data for aliphatic hydrocarbons published in recent Landolt–Börnstein Tables [2, 3].

6. TRC SOURCE DATABASE

The Thermodynamics Research Center has maintained a database of experimentally measured numerical values of thermophysical properties for the past 12 years. It covers all equilibrium thermodynamic and thermochemical properties of all phases, and transport properties of fluids. It includes pure compounds, binary and ternary mixtures of organic compounds, and non-metallic inorganic compounds. It has most of the characteristics listed in Section 3.

The database exists in two forms. The working version runs under UNIX and may be accessed in hierarchical or relational modes. In consists of 50 record types (or tables) suitably indexed. Components are identified by Chemical Abstracts Services Registry Numbers. Software for maintenance, search and retrieval, and addition and revision of data has been developed. New information is added continuously. It can be searched for various combinations of component, property, year of publication, author name, and reference key. An earlier implementation of the database has been described by Wilhoit and Marsh [4], and a more detailed description of the current version is given in a TRC report [5].

A read-only version that runs under DOS or Windows on IBM PCs is also available. This version is updated every three months by downloading information from the working version. It supports several output formats in addition to the screen display. Queries can be entered from the keyboard or from batch input files. At present this version runs in nine locations. As of April 1997 it contains 1.8 million records organized into the following major sections.

- 107,000 compound registry numbers, each with the chemical formula, and one or more names.
- Property data for 20,000 pure compounds and 15,000 mixtures.
- 77,000 literature references and 68,000 author names.
- 420,000 property values of pure compounds, with appropriate independent variables, phase identification, and sample descriptions; most are accompanied by estimates of uncertainties and other descriptive information.
- 280,000 property values of mixtures, also with independent variables, phase identification, sample descriptions and uncertainty estimates.
- 2000 thermochemical property values (calorimetric heats of reaction and equilibrium constants).

The working version occupies 360 MB of disk space and the PC version 35 MB. We estimate that at present the SOURCE database contains around 30% of the world's supply of thermophysical data on pure organic compounds and 10–20% of the supply of mixture data.

The following organizations have contributed substantial amounts of data: IUPAC Commission of Thermodynamics (critical constants); Institute of Physical Chemical Technology, Prague; the Institute of Physical Chemistry and Coal Chemistry and the Polish Academy of Science in Poland; the Thermodynamics Center, Kiev, Ukraine; Laboratory of Thermodynamics of Organic Compounds, Byelorussian State University, Minsk, Belarus; ESD International, London;, Nanjing Institute of Chemical Technology, Nanjing; Laboratory of Computer Chemistry, Institute of Metallurgy, Chinese Academy of Science, Beijing; and Kyoto Institute of Technology, Kyoto. It is the intention of the TRC to make this information readily available to investigators and compilers of data.

The Thermodynamics Research Center also maintains the TABLE Database, which contains all of the numerical property data in the two serial publications *Thermodynamic Tables—Hydrocarbons and Thermodynamic Tables—Nonhydrocarbons* [6].

One of the output options of the SOURCE database on the PC is one that can be read by the LOADER2 program developed by the National Engineering Laboratory, Glasgow [7]. This program accepts experimental thermophysical property data for pure compounds, supplements them with correlated data, and fills in missing gaps. It then analyzes these data, adjusts them for internal consistency, and fits them to a series of empirical smoothing functions. It uses estimated uncertainties to generate weighting factors. The result is a set of fitted parameters and tables of smoothed internally consistent property values. Properties processed are vapor pressure, saturated liquid density, ideal gals heat capacity, saturated vapor enthalpy, enthalpy of vaporization, saturated liquid heat capacity, surface tension, and liquid and gas viscosity and thermal conductivity. The user can select from a series of correlating procedures and smoothing functions or accept the defaults based on the characteristics of the compound. A screening routine for selecting the best data from the SOURCE database output based on the principles of Section 5, as well as a preliminary test of internal consistency, has been developed for the LOADER2 formatted files.

The combination of the TRC SOURCE database with the LOADER2 output, the screening routine, and the LOADER2 program furnishes an existing practical method for the production of dynamic compilations for pure compounds. This procedure is being extended to mixtures.

REFERENCES

1. IUPAC Commission 1.2 on Thermodynamics, *J. Chem. Thermodyn.* **13**:603(1981).
2. R. C. Wilhoit, K. N. Marsh, X. Hong, N. Gadalla, and M. Frenkel, *Landolt-Börnstein, Group IV: Physical Chemistry, Vol. 8, Thermodynamic Properties of Organic Compounds and Their Mixtures, Subvolume B. Densities of Aliphatic Hydrocarbons: Alkanes* (Springer-Verlag, Berlin, 1996).
3. R. C. Wilhoit, K. N. Marsh, X. Hong, N. Gadalla, and M. Frenkel, *Landolt-Börnstein, Group IV: Physical Chemistry, Vol. 8, Thermodynamic Properties of Organic Compounds and Their Mixtures, Subvolume C. Densities of Aliphatic Hydrocarbons: Alkenes, Alkadienes, Alkynes, and Miscellaneous Compounds* (Springer-Verlag, Berlin, 1996).
4. R. C. Wilhoit and K. N. Marsh, *J. Chem. Inform. Comput. Sci.* **29**:17 (1989).
5. R. C. Wilhoit and K. N. Marsh, *Documentation for the TRC Source Database* (Thermodynamics Research Center, College Station, TX, 1996).
6. *TRC Thermodynamic Tables—Hydrocarbons and—Non-Hydrocarbons* (Thermodynamics Research Center, Texas A&M University System, College Station, 1997).
7. *LOADER2 for Windows. User Manual and Reference Guide* (Physical Property Data Service, NEL, East Kilbride, Glasgow, UK, 1996).