International Journal of Thermophysics, Vol. 22, No. 1, 2001

TRC SOURCE Database: A Unique Tool for Automatic Production of Data Compilations[1]

M. Frenkel,[2,5] Q. Dong,[2] R. C. Wilhoit,[3] and K. R. Hall[4]

Thermochemical and thermophysical properties of chemicals are the basis of most simulations of commercial processes. The TRC SOURCE database is an extensive archive of numerical experimental values of thermophysical properties extracted from the world's scientific literature. A suite of input/output routines that utilize current database technology has been developed. The database in combination with suitable software enables the automatic retrieval, interpretation, selection, and formatting of a wide range of thermophysical properties. Their output supports the automatic production of user specified compilations and the direct interaction with user-written application programs.

KEY WORDS: automatic data productions; database applications; data compilation; data correlation; experimental values; relational database; thermophysical properties.

1. INTRODUCTION

SOURCE, created and maintained by the Thermodynamics Research Center (TRC), is a large, general-purpose archive of experimental data covering thermodynamic, thermochemical, and transport properties for pure compounds and mixtures of well-defined composition. The database contains numerical values for various kinds of thermodynamic and thermochemical properties of systems in all phases and values of transport properties of fluids. SOURCE does not include properties whose values depend

[1] Paper presented at the Fourteenth Symposium on Thermophysical Properties, June 25–30, 2000, Boulder, Colorado, U.S.A.

[2] Thermodynamics Research Center, Physical and Chemical Properties Division, National Institute of Standards and Technology, Boulder, Colorado 80303-3328, U.S.A.

[3] Texas A&M University System, College Station, Texas 77843-3122, U.S.A.

[4] Chemical Engineering Department, Texas A&M University, College Station, Texas 77843-3122, U.S.A.

[5] To whom correspondence should be addressed.

0195-928X/01/0100-0215$19.50/0

upon the history of samples. The database identifies literature sources for molecular properties that may have been used in statistical calculations of thermophysical properties, but it does not include such property values. At present, SOURCE emphasizes organic and nonmetallic inorganic compounds.

A goal of TRC is to incorporate all relevant, published data within the scope of SOURCE and to maintain coverage as new data appear. TRC intends that the database be available for input and output to the users of both academic and industrial communities throughout the world. Additionally, TRC supports the production of evaluated and selected data required for technical applications.

Wilhoit and Marsh [1] have suggested the concept of dynamic compilations that enable users of numerical data to create dynamic compilations upon demand. This concept contrasts with static compilations that are available in advance of need. Dynamic compilations reduce the effort spent anticipating future needs of users and the effort required to keep static compilations current. A recently developed software product, DataExpert [2], implements this concept to generate a set of mutually consistent numerical data for a number of physicochemical properties.

Dynamic compilations depend upon two resources: (a) a suitably designed and organized database of raw experimental data (such as SOURCE) and (b) a software package that can interact with the database, search and recognize data relevant to a user request, perform transformations, evaluate and select best values, and fit the results to suitable models. The DataFetch library [3] and a combination of directly linked retrieval subroutines with a customized interface [4] are sufficient for this purpose.

It is critically important that the SOURCE database includes the estimated uncertainties for practically all the numerical data stored. This feature allows, in principle, determination of the quality of recommended data based upon the original experimental data collected in SOURCE.

This paper provides a detailed description of SOURCE, illustrates potential means of communicating with it, and demonstrates how to use SOURCE for automatic production of critical data compilations. A variety of the formats (both hard-copy and electronic) is possible.

2. INFORMATION IN SOURCE

SOURCE contains four major types of information.

2.1. Compound Identification

Registry numbers identify pure compounds and components of mixtures throughout the database. SOURCE uses numbers assigned by Chemical

Abstracts Services when available and numbers assigned by TRC otherwise. Registry numbers link to an empirical formula, a coded representation of the structural formula, and one or more names. The database contains 113,000 registry numbers and 218,000 names. Reacting systems of one or more compounds also receive registry numbers. Among the stored compounds, approximately 15,800 pure compounds, 9000 binary and ternary mixtures, and some 2500 reaction systems have data records. Chemical reactions have a classification code and registry numbers of four species in the reaction.

2.2. Sample Descriptions

SOURCE describes over 17,900 distinct samples used in property measurements. The description includes the source of sample, method of purification, and final purity as reported by the authors of the document. Formal abbreviations exist and, by sample numbers, identify different samples of the same compound used for measurements in the same document.

2.3. Literature References

SOURCE contains citations of original documents and associated information (such as titles, document types, classification of information, and comments) and links to data values. Names of authors appear in a dedicated table linked to the citations. Thus, it is possible to retrieve literature references by year of publication, author, compound identity, property, or combinations of them. The database contains 82,000 citations, of which over 22,000 citations contribute numerical values to SOURCE.

A reference key identifies a document. A reference key consists of the year of publication, the first three letters of the last names of the first two authors (or three letters and three blanks for only one author), and a number to provide a unique key.

2.4. Numerical Values

Each numerical property value appears in a data record. A data record also contains values of state variables and an estimate of the uncertainty of the property. Each data record links to the three kinds of information listed above to codes that identify the property, the primary phase, other phases in equilibrium with the primary phase, other information about the way the data appear in the original document, and a trail of its entry into the database. Property values are converted to SI units on entry while retaining

sufficient information to regenerate the original numbers. The database now contains 850,000 data records.

Data records are grouped into data sets. A data set contains values for a particular property of a system with components from particular samples consisting of particular phases reported in the same document. Other considerations enter into defining data sets. The objective is to mirror the way the data have been measured and reported in the document.

IMPLEMENTATION OF THE ORACLE VERSION OF SOURCE

TRC has used the Oracle Relational database management system and its development environment since 1997 after embarking upon a large project to rebuild SOURCE and eliminate the SIR database. During the past 2 years, TRC has updated the computer system, installed the Oracle software, and implemented and modernized SOURCE.

3.1. Oracle Architecture

The Oracle server/client environment allows splitting processes between the database server and client application programs. The computer running a database server handles the database transactions while PCs running database applications concentrate upon the interpretation and display of data. At TRC, the database server, Oracle RDBMS Enterprise Edition 8.1.5, resides on a DEC alpha personal workstation running Digital Unix V4.0D, while development tools, Developer 2000 suite, and other client tools reside on a TRC local network. In this configuration, client software programs run on PCs and the associated server processes run on the DEC machine using the university network and the Oracle network software, Oracle Net8. Oracle software products on the server involve Oracle RDBMS 8I, SQL* Plus command line, SQL Loader utility, and Import/Export utility. The client software system includes a development tool suite—Developer 2000, a GUI interface server manager—Enterprise Manager, Oracle Net8 Assistant, and SQL* Plus.

3.2. Organization of the Database

SOURCE is a relational database. It consists of 35 tables and relations. Each table contains a sequence of rows. Each row contains one or more items, called columns or attributes. Columns contain character strings, integers, floating-point numbers, or certain special structures such as dates. All columns in a table contain the same number and type of columns. The data definitions are compatible with the third normal form except in a

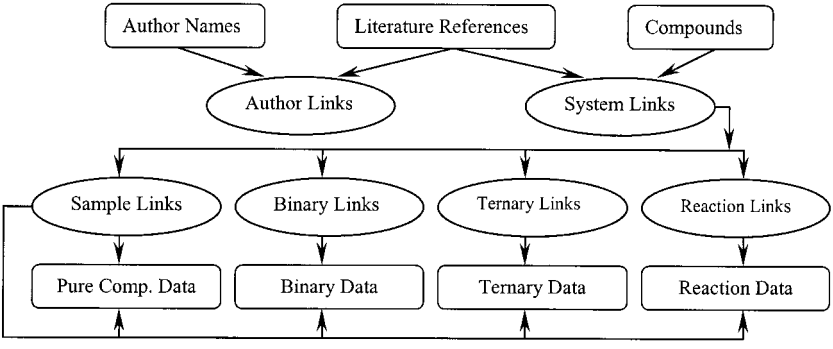

Fig. 1. The SOURCE data model.

few special cases. Combination of one or more columns in a table forms the primary key that uniquely identifies each row. Figure 1 presents the relational data model for SOURCE, a further subdivision of the four main types of data listed previously. Each box in this figure represents one or more database tables. Tables in some boxes store the indicated information, while others serve as links to relate one kind of information to another.

SOURCE groups tables that contain numerical data according to the number of components—pure compounds, binary mixtures, ternary mixtures, and chemical reactions. Within each of these groups, tables express the "effective degrees of freedom" for the data sets. The effective degrees of freedom determine the number of state variables in a data record. The Gibbs Phase Rule establishes the total degrees of freedom for a system according to the number of independent components and number of phases. The effective degrees of freedom are less than the total degrees of freedom if the system contains additional constraints for a particular data set. Examples of additional constraints are a state variable held constant, a property code that includes a constraint in the definition, and a special state such as a critical or an azeotropic state.

Although the Gibbs Phase Rule does not apply to transport properties, the concept of effective degrees of freedom applies to them as well. Thus, for example, the boiling point of a pure compound has one degree of freedom with pressure usually being the state variable. However, the normal boiling point has 0 degrees of freedom because the pressure is 1 atm by definition. References 3 and 4 discuss the choice of properties and state variables and the concept of effective degree of freedom.

Properties for systems with 0 degrees of freedom appear in tables corresponding to the number of components in the system. The properties are constants. Data sets for these properties contain only one data record

and each row in these tables contains one data set. In these cases the data records include all the descriptive information and metadata.

Properties of systems with $\geqslant 1$ effective degrees of freedom appear in a pair of tables constructed for the number of components and effective degrees of freedom. Data sets for these properties contain one or more data records. One table contains descriptive information, while another table contains the associated data sets. For example, a pair of tables for systems of one component and 1 effective degree of freedom contains values for the vapor pressures of a pure compound. PVT data for a pure compound could appear in tables for 2 effective degrees of freedom. Vapor–liquid equilibrium properties for binary systems are in the binary system block in tables with 1 degree of freedom for isothermal or isobaric data sets or in tables with 2 degrees of freedom if both the temperature and pressure change.

Often a data set could fit into the database in various places. These locations reflect the way the original document presents the data. Retrieval software should locate and extract data relevant to a user request.

A database document [4] reveals definitions of terms, contents of tables, database policies, and examples in considerable detail. This document lists 130 codes that identify properties plus codes for various composition variables and phases. Some of these codes apply to a particular table, while other codes appear in several tables. Some codes describe both properties and state variables.

Symbols exist for metastable phases, glass phases, various liquid crystal phases, and more complicated phases in mixtures.

3.3. Data Integrity and Protection

Data quality is one of the most important issues for databases, especially for SOURCE, which contains a large variety of properties on a large variety of systems with sophisticated relations among different groups of data. SOURCE employs five mechanisms for data integrity: data type definitions, primary key constraints, foreign key constraints, check constraints, and database triggers. Each table or each column defines its mechanisms to prevent invalid data entry. The mechanisms also enforce the predefined data rules associated with information in SOURCE. If a data transaction violates an integrity constraint, it is rejected and an error message is displayed.

Primary key constraints ensure the uniqueness of each row in the table. In SOURCE, each table contains one primary key constraint. A primary key is a concatenated sequence of one or more columns. For tables in the reference group, the primary key is the reference key. In the compound group, the registry number is the primary key. For the numerical data

group, a concatenated key of up to 20 columns identifies records, a combination of one or more registry numbers, the reference key, sample numbers, property code, and data set numbers.

As with any relational database, SOURCE relates the contents of different tables through common columns. FOREIGN KEY constraints maintain these relationships, thus ensuring that certain columns in one table match the primary key of some other table. The foreign keys also prevent creating rows in tables that have no relationship to anything else.

Check constraints promote data integrity by enforcing specific or complicated integrity rules restricting the contents of certain columns to one of a valid list or restricting numbers within a certain range. These constraints help screen out invalid data entries.

3.4. Database Input and Output

Several ways exist to access SOURCE for input and output. These include a primary tool for daily data entry and maintenance—Data Entry Form on a client machine, batch input and output programs on the server machine, and SQL* Plus data reports generated from both the server and the client.

A special program displays registry numbers of compounds by searching the database for names, partial names, or formulas, or combinations. This program runs on the UNIX OS and interacts with a special database containing these data. It is a vital adjunct to all the I/O processes.

3.4.1. Interactive Data Entry Form

The Data Entry Form (built using Oracle Developer 2000) operates through the server/client model. Client machines are the PCs connected to either the local TRC network or the university network. The local TRC file server holds a single copy of Oracle Developer 2000 and the runtime version of the Data Entry Form. From a PC workstation, users can execute the Data Entry Form and interactively manipulate the database on the back-end machine, a DEC alpha workstation. The Form permits the user full access privileges to the SOURCE tables, including operations of SELECT, INSERT, DELETE, and UPDATE. The Form contains the relations among database tables with a master-detail mechanism that allows the user to take advantage of the foreign key constraints predefined in the database. The users can query and modify the database using the Form display without detailed knowledge of the database schema and constraints. The principal use of the Form is to browse the table rows and to enter or correct small amounts of data.

3.4.2. Batch Input Mode

The utility, Oracle Loader, is a convenient tool for loading large amounts of data into SOURCE. The input data files required for the Loader are non-Oracle sources such as ASCII files with flexible data formats. The data files are specific for each table in SOURCE, and items in these files correspond to columns in the database table. Because the Loader input file has a database table orientation, it may contain information for one table from a number of articles. In contrast, information extracted from an article occupies one input file even though it might be for several database tables. For example, property values of density measurements for pure compounds and binary mixtures go to the reference table, the compound table, the sample table, pure data tables, and binary data tables.

Programs exist (written in PERL) to convert files of user-generated information into ones acceptable to the Loader facility. The user-generated files are in free format form and are much more compact and human-friendly than those formatted for Loader. The conversion programs check the input files for integrity constraints and other sources of errors before attempting to load them into the database. Programs are available for registry numbers and compound names, references and author names, and numerical property data.

Any direct text editor can create user-generated files. Such editors also furnish the primary means for experts outside TRC to submit data. These procedures are the only ones used to load the data collected through the TRC International Data Exchange Program, which encompasses a number of research institutions throughout the world (Institute of Chemical Technology, Prague, Czech Republic; Belarusssian State University, Minsk; Thermodynamics Center of Oil and Gas Industry of Ukraine, Kiev; Laboratory of Computational Chemistry of the Chinese Academy of Sciences, Beijing; etc.).

3.4.3. Batch Output Mode

The batch output utility available on the server includes three major programs that search the database with user specification of information desired. The return query results in ASCII data files. The Reference Extract Program retrieves complete literature information for reference keys supplied by the user. The Compound Extract Program extracts names, formulas, and SMILES line notations for compounds identified by registry numbers. The Data Extract Program locates various property values of pure compounds and binary and ternary mixtures. Query results contain direct measurement values along with associated information that specifies the components, properties, phases, literature references, sample descriptions,

and estimated uncertainties. The results appear in a tabular format in the output files with detailed explanations on each column value.

3.4.4. SQL PLUS Reports*

SQL* PLUS is the Oracle front-end command line interface to an Oracle database. Both the server and the client can invoke it. In addition to the tasks of data definition, data manipulation, and database maintenance, generation of a query report is one of the most frequently performed tasks of SQL* PLUS. Using the SQL SELECT statement and some special formatting commands permits extracting any data directly from the database tables and reformatting column data in the query results.

4. AUTOMATIC UTILIZATION OF DATA

Critical evaluation of thermophysical, thermochemical, and transport property data is usually very time and resource consuming. Often new data necessitate redoing an evaluation report for a particular compound/property combination that is nearly complete. The concept of dynamic compilation makes this process much more efficient. TRC has developed a set of procedures to automate production of reference books providing recommended density and phase transition property data for thousands of organic compounds [6–12]. This unique process, illustrated in Fig. 2, combines data entry into SOURCE, uniform extraction of data, data processing leading to generation of mutually consistent recommended data sets, automatic typesetting using a variety of specially designed Macros for Word processor, compilation of the list of references and Chemical Abstract Registry Number Index and Chemical Name Index as direct output from the database, and generation of a camera-ready copy at the end of the process.

Another automatic utilization of the data is the design of electronic databases for critically evaluated data. For example, the TABLE database, an electronic version of TRC Tables—Hydrocarbons and TRC Tables—Non-Hydrocarbons, receives updates and revision automatically using this process providing convenient user-friendly access to the data, the means to make multiple searches, and interpretation and plotting of the data. This process also can produce a series of selected compound/property databases such as Density, Vapor Pressure, Ideal Gas, and Virial Coefficients as illustrated in Fig. 2.

The third alternative to utilize the results of automatic production of the thermophysical data compilation is to employ the output file of the recommended data as an input file for a chemical process simulation. TRC currently collaborates with a number of simulation software design companies to implement this schema in a variety of commercial products.

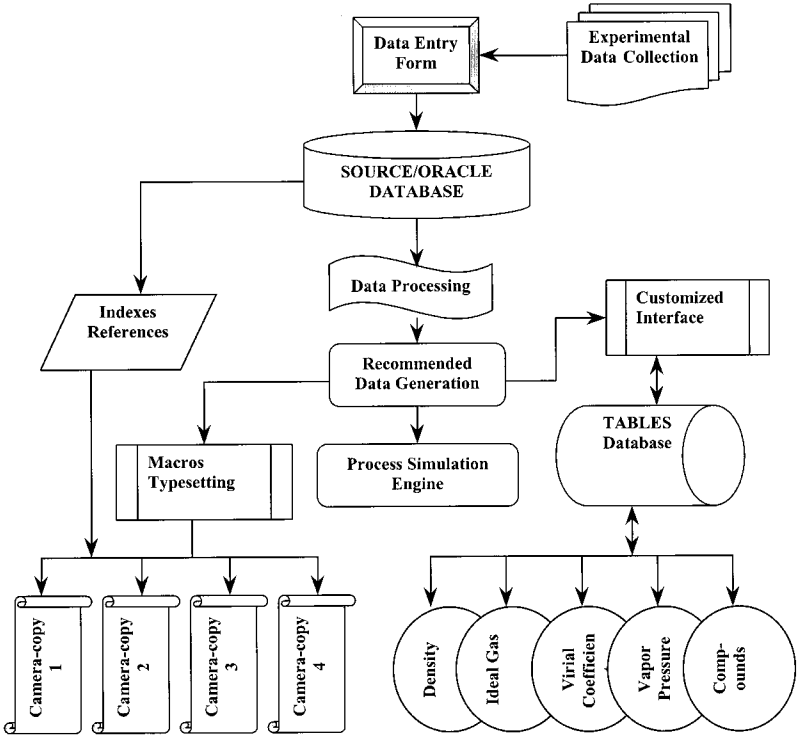

Fig. 2. Automatic data compilation production schema.

5. STAND-ALONE WINDOW VERSIONS OF THE TRC DATABASES

Several routes exist to produce versions of SOURCE that run on local servers or individual workstations. These local versions would receive periodic updates by downloading from the ORACLE version at the TRC office. The first operational version uses the Microsoft Windows operating system [5].

Replication of an MS Access version from Oracle databases on the server through ODBC is a practical approach to establish a database support system for redevelopment of a Windows version of TRC databases. ODBC (Open Data Base Connectivity) is a software layer that provides all databases with a common interface to communicate with each other. Microsoft Access is an end-user data manipulation and query tool. MS Access is familiar to most developers and users because of its simplicity and ease of use. It is important to have an MS Access version of SOURCE for distribution and redevelopment purposes.

6. INTERNET APPLICATION DEVELOPMENT

Next-generation Web applications are thin-client (as opposed to fat-client), implemented in a multitier rather than a two-tier architecture. Internet-based technology enables companies to deliver new functionality quickly, *via* a nonproprietary mechanism, to anyone having a Web browser while immediately reducing client-side management costs.

TRC has begun to investigate emerging Web technology on next-generation database applications. The decision has been made to move from the two-tier server/client architecture to thin-client computing. The new architecture adds a middle tier, Oracle Application Server, to the existing server/client system. At the middle tier, Oracle Application Server runs on the Windows NT Server platform. This architecture enables TRC to deploy a dynamic web site, WEBTRC, that is the center of TRC electronic products with full capacity for interacting with either Oracle Databases or MS Access databases.

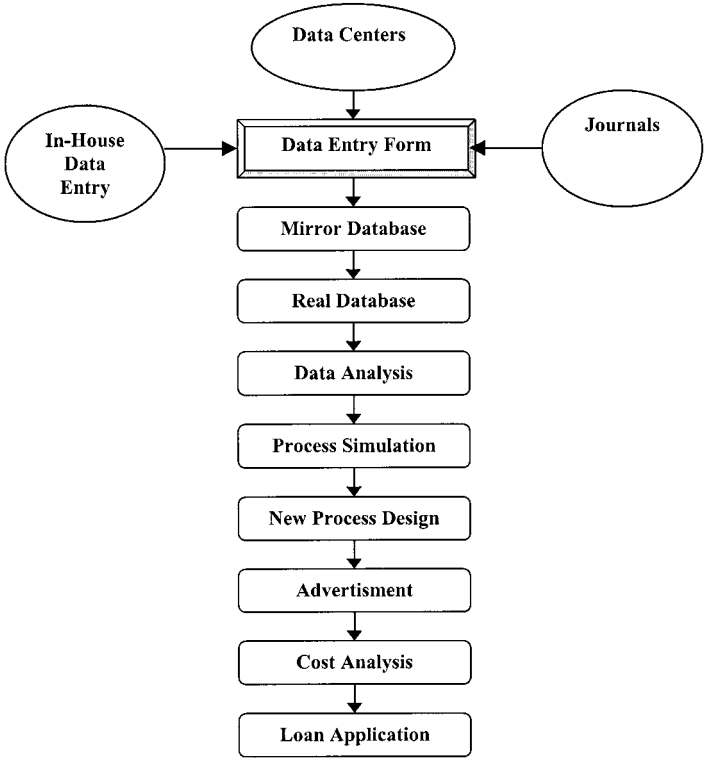

Fig. 3. Schema of the Internet communication with SOURCE.

WEBTRC is under construction at TRC. The first web application is the Source Data Entry Form. Anticipated web products are the SOURCE and TRC TABLE databases. Oracle Web-enabled Developer 2000 and Oracle Application Server allow existing server/client codes to run on the Oracle web server and access from any browser. However, certain areas exist where the codes must be adjusted to web environment. Availability of the SOURCE Data Entry Form through the Internet access would allow dramatically enhanced coverage of SOURCE, making it possible to submit data to the database from individual experts as soon as the data become available. It also can create an opportunity to collect data more efficiently using the International Data Exchange Program. On the other hand, distribution of the data over the Internet in the future and communication of the hosting web site with other web sites related to the expert analysis of the physicochemical processes might provide new opportunities for chemical engineers and small engineering companies as suggested in Fig. 3.

REFERENCES

1. R. C. Wilhoit and K. N. Marsh, *Int. J. Thermophys.* **10**:247 (1999).
2. R. C. Wilhoit, A. I. Johns, M. Frenkel, and K. R. Hall, *DataExpert System: A New Concept for Evaluating Thermophysical Properties*, 15th IUPAC Conference on Chemical Thermodynamics, Porto, Portugal, Conference Booklet, Contribution C4-7 (1998).
3. R, C. Wilhoit, *Int. J. Thermophys.* (in press).
4. *Documentation for the TRC Source Database* (Thermodynamics Research Center, College Station, TX 77843, Nov. 1999). (An abridged version can be downloaded from the TRC website: trcweb.tamu.edu.)
5. X. Yan, Q. Dong, M. Frenkel, and K. R. Hall, *Int. J. Thermophys.* (in press).
6. Z.-Y. Zhang, M. Frenkel, K. N. Marsh, and R. C. Wilhoit, *Enthalpies of Fusion and Transition of Organic Compounds: Landolt–Bornstein*, New Series, Vol. IV/8A (Springer-Verlag, Berlin, 1995).
7. R. C. Wilhoit, K. N. Marsh, X. Hong, N. Gadalla, and M. Frenkel, *Densities of Aliphatic Hydrocarbons: Alkanes*, *Landolt–Bornstein*, New Series, Vol. IV/8B, (Springer-Verlag, Berlin, 1996)
8. R. C. Wilhoit, K. N. Marsh, X. Hong, N. Gadalla, and M. Frenkel, *Densities of Aliphatic Hydrocarbons: Alkenes*, *Alkadienes*, *Alkynes and Miscellaneous Compounds*, *Landolt–Bornstein*, New Series, Vol. IV/8C (Springer-Verlag, Berlin, 1996).
9. R. C. Wilhoit, X. Hong, M. Frenkel, and K. R. Hall, *Densities of Monocyclic Hydrocarbons*, *Landolt–Bornstein*, New Series, Vol. IV/8D (Springer-Verlag, Berlin, 1997).
10. R. C. Wilhoit, X. Hong, M. Frenkel, and K. R. Hall, *Densities of Aromatic Hydrocarbons*, *Landolt–Bornstein*, New Series, Vol. IV/8E (Springer-Verlag, Berlin, 1998).
11. R. C. Wilhoit, X. Hong, M. Frenkel, and K. R. Hall, *Densities of Polycyclic Hydrocarbons*, *Landolt–Bornstein*, New Series, Vol. IV/8F (Springer-Verlag, Berlin, 1999).
12. M. Frenkel, X. Hong, R. C. Wilhoit, and K. R. Hall, *Densities of Alcohols*, *Landolt–Bornstein*, New Series, Vol. IV/8G (Springer-Verlag, Berlin, 2000).