Data Quality Assurance for Thermophysical Property Databases—Applications to the TRC SOURCE Data System

Qian Dong,* Xinjian Yan, Randolph C. Wilhoit, Xiangrong Hong, Robert D. Chirico,
Vladimir V. Diky, and Michael Frenkel

Thermodynamics Research Center (TRC), National Institute of Standards and Technology (NIST),
Boulder, Colorado 80305-3328

Received November 5, 2001

To a significant degree processes of database development are based upon human activities, which are susceptible to various errors. Propagation of errors in the processing leads to a decrease in the value of original data as well as that of any database products. Data quality is a critical issue that every database producer must handle as an inseparable part of the database management. Within the Thermodynamics Research Center (TRC), a systematic approach to implement database integrity rules was established through the use of modern database technology, statistical methods, and thermodynamic principles. The four major functions of the system—error prevention, database integrity enforcement, scientific data integrity protection, and database traceability—are detailed in this paper.

I. INTRODUCTION

Experimental thermophysical and thermochemical property databases play a vital role in scientific research and industrial design. They assist engineers and technicians in designing and operating manufacturing facilities, researchers in finding and evaluating new products, environmental scientists in predicting the fate of hazardous substances in the environment, and scientists in developing new theories and models.

Data quality is a critical issue that every database producer must address in fulfilling the above functions.[1−3] Generally, data quality applied to an experimental property database can be considered as having two distinct attributes. The first quality attribute relates to “uncertainties” assigned to numerical property values by experimentalists or database professionals. Uncertainties assigned by experimentalists quantify the quality of a measurement result, including any possible source of bias.[4,5] Evaluation of these uncertainties by data professionals is often a complex task, as “uncertainties” reported by experimentalists—when provided—often represent simple experimental precisions or fail to incorporate key criteria such as sample purity or uncertainties in calibration methods. The uncertainty is an essential piece of quality information expected by anyone who uses the data. The second quality attribute refers to “data integrity”. It means that in the database, data steadfastly adhere to the original source and conform to various database rules. This attribute can be assessed by the completeness and correctness of data records stored in the database. A variety of possible errors in data entry can have a direct effect on the completeness and correctness of information stored in the database. Both quality attributes (“uncertainties” and “data integrity”) are of equal importance for overall data quality. Though a brief discussion of uncertainties will be given in the next section, the attribute of “data integrity” is the primary subject of this paper.

In a database, experimental information can be deposited, processed, and interpreted at various levels of processing complexity. Each level can add value to the original “raw” data by organizing the data in a systematic and convenient way, normalizing and standardizing the data, and assigning reliable uncertainties. Conversely, the inclusion and propagation of errors in data processing can enormously decrease the value of the original data as well as that of any product derived from the database.

Examples of errors occurring in the process of information collection from literature and input to such databases are as follows: (a) typographical errors; (b) unit-conversion errors (e.g., joules vs calories); (c) report interpretation errors (i.e., misreading the original documents); (d) metadata compilation errors (i.e., Supporting Information is incorrectly entered or interpreted); and (e) original report errors (i.e., the results reported in the original data source are wrong or misleading). It can be safely estimated that a large-scale numeric database without critical evaluation may have an error rate of 2−5%. Addressing and resolving these quality deficits are the first priority for any database management process.

The TRC SOURCE data system was designed and built as an extensive repository system of experimental thermophysical and thermochemical properties that have been reported in the world’s scientific literature.[6,7] It has grown extensively in size and functionality during the past 15 years.[8−10] TRC SOURCE now consists of over 980 000 numerical values on 17 000 pure compounds, 10 000 binary and ternary mixtures, and 3900 reaction systems. There are a total of approximately 170 000 combinations of system/property sets reported from independent measurements. Stored also in TRC SOURCE are approximately 110 000

* Corresponding author phone: (303)497-3224; fax: (303)497-5044; e-mail: qdong@boulder.nist.gov.

10.1021/ci010118e
Published on Web 05/02/2002

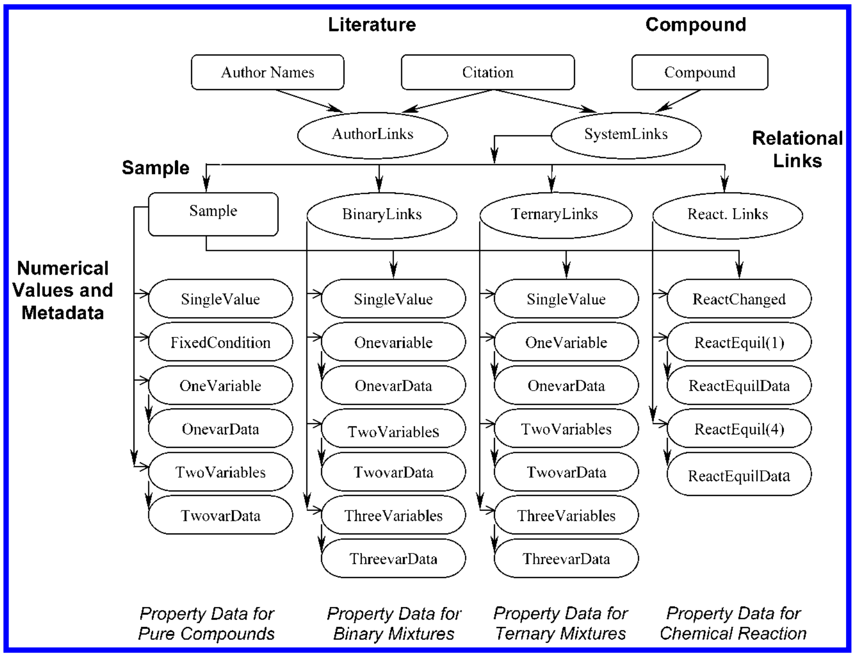

Figure 1. The TRC SOURCE schema.

records of compound identification; 83 000 records of bibliographic information; and over 70 000 records containing information pertaining to the identity of the authors of the original sources. The total number of distinct records currently exceeds 2 000 000.

The large size of the TRC SOURCE data system, the required complex relationships among the tables and records, the many ways it can be accessed and used, and the importance of maintaining high data accuracy combine to form a major challenge to the assurance of data quality. During the last 5 years, a thorough and painstaking cleansing of the entire TRC SOURCE data system was performed through the implementation of many defined primary and foreign keys in the database tables. Consistency among the TRC SOURCE records has been enforced, and, therefore, quality has been greatly improved. However, effective quality assurance requires much more than periodic cleansing and just a few isolated and simple quality checks. It demands rather a systematic and practical approach that fully utilizes modern database technologies, statistical analysis methods, and scientific principles to automate the major tasks of data quality assurance including error prevention, anomaly detection, and data traceability.

In this paper, general features of the TRC SOURCE data system are described. Guidelines used in the development of the TRC data-quality-assurance policies are detailed, and applications of such a systematic approach to TRC SOURCE are described fully.

II. FEATURES OF THE TRC SOURCE

Measured (Experimental) Data. There are several types of data: measured data, derived data, correlated data, and recommended data, in terms of the way in which the data are generated. Correlated (or estimated) and recommended data are not stored in the TRC SOURCE data system, and examples of these are compiled in the NIST/TRC Table database and NIST/TRC Vapor Pressure database.[11,12] Measured data and some derived data based on experimental values are the primary targets of the TRC SOURCE depository, as reflected in its name, “TRC SOURCE”. It often happens that the same property on nominally the same system in the same state is measured and reported by different investigations. These are considered duplicate measurements, and all such duplicates are included in the database. Derived data are calculated based on measured data and the mathematical relationships strictly defined by thermodynamics. For example, second Virial coefficients or any property of an ideal gas or ideal solution are always derived.

Experimental data are the fundamental building blocks for generating recommended data and for developing data prediction methods. It must be noted that the mixing of estimated or correlated data with measured data into a database makes it essentially unusable for a selection of “best” values or for the development of data prediction models.

Data Organization and Structure. The structure and specifications of the TRC SOURCE data system is described in detail elsewhere.[13] To fully describe a particular measurement result, information in TRC SOURCE is organized into six types of records (Figure 1)—literature, compound, sample, metadata, numerical values, and relational links.

Literature records are organized in two different tables. The first table stores literature citations to all the documents that report experimental data as well as to other documents that contain information about measurement science, while the second table stores names of the authors. With the literature records, it is possible to generate a bibliography for particular properties, on related subjects, or for a particular author. All other records are linked to these citations.

Compound records identify chemicals by CAS registry number, empirical formula, and SMILES (Simplified Molecular Input Line Entry Specification) notation as well as chemical name (names). The first three identifier records are combined within one table, while one or more names of compounds are stored in another table. Detailed sample records, metadata, and numerical-value records are linked to these compound records.

Sample records that are contained in one database table describe over 18 900 distinct samples used in property measurements. The description includes source of sample, method of purification, and final purity as reported by the authors.

Numerical values and their metadata records are combined in 25 tables to specify property values for pure chemicals, binary and ternary mixtures, and chemical reactions. The metadata records give meaning to the numerical values, which represent the core of TRC SOURCE. Each property value applies to a system of specified chemical composition in a particular state. The state of a system is described by a set of state variables and by the phase or phases present. Properties, phases, and state variables, in the format of metadata, are identified by codes defined in the database schema. One hundred and thirty properties and state variables are recognized. All possible phases are accommodated. Also included are a variety of other metadata such as the objective and method of measurement, definition of units, identification of temperature scale, and format of reported data.

Relational link records connect citations to author identification records, citations to chemical systems (i.e., pure components, mixtures or reactions), and chemical systems to their numerical values and metadata. There are five database tables for this record.

In summary, TRC SOURCE provides not only property values but also the Supporting Information that fully describes the numerical values, including a complete description of the chemical system, with all state variables, phases, and units defined as well as a description of the measurement methods and purposes. Consequently, in addition to the extensive collection of the property values available, the metadata can be easily accessed to generate reports on the development of measurement technology.

Uncertainties. Uncertainties represent a data quality attribute. Without uncertainties, numerical property values cannot be evaluated. In TRC SOURCE, an estimate of uncertainty is assigned to each numerical property value. It is a measure of the quality of the property value, expressing the confidence in the value determined by the TRC data professionals based on reevaluation of the measurement reports and on comparisons with duplicate and mathematically related values in the database. An approximate quantitative description of the uncertainty is a range of values, which includes the “true” value of the property with a probability near 95%. In assessing uncertainty all potential sources of errors are propagated to the uncertainty of property.

The provision of uncertainties for property values in TRC SOURCE establishes the basis for determination of recommended values.[13] The uncertainties provide scientists and engineers with quantitative measures of data quality, which are essential for valid sensitivity analyses during simulation and development of chemical processes.

Data uncertainty is a broad and complex subject.[14,15] Correct assignment of uncertainty requires highly knowledgeable and skilled data professionals and, furthermore, includes a subjective component. The large scale of the data collection effort within TRC has required the development of computer software for “guided” uncertainty assessment. Use of the software ensures consistency in the evaluation approach and result when numerous data compilers are involved. All assessment criteria are stored in TRC SOURCE as metadata and are available for simple reevaluation, if needed.

III. A GUIDELINE FOR DATA QUALITY ASSURANCE

An essential first step in the achievement of data-quality-assurance (DQA) is to establish well-defined rules and guidelines for all aspects of the database. Initially, a statistical analysis of the TRC SOURCE records was made to obtain a preliminary status of existing data quality. The principles of both relational databases and thermodynamics were used in forming the criteria for analyzing database records and capturing anomalies. Based on this analysis, guidelines for quality management were proposed. The guidelines were compiled as an internal document[16] that clarifies the DQA rules associated with all major TRC SOURCE operations and regulates data input and output. As a foundation for quality improvement, six steps were identified: (1) literature collection, (2) information extraction, (3) data entry preparation, (4) data entry insertion, (5) anomaly detection, and (6) database rectification. Rules and criteria for achieving specified quality goals were developed for each step. The six major steps in operating the data system are summarized as follows.

Step 1. Literature Collection. Journal articles and reports that contain the following information are collected: (1) experimental property data; (2) purification of compounds and determination of purity; (3) experimental techniques and descriptions of apparatus; (4) fundamental constants, standards, symbols, and error analysis; and (5) reviews, compilations, and surveys of experimental data. The database does not include numerical data from estimations, predictions, correlations, or other calculation methods. However, the bibliographic information is collected.

Step 2. Information Extraction. The following information is mandatory in extracting data: (1) property data; (2) metadata; (3) chemicals involved; (4) sample source, purity and stability; (5) method and primary objective of measurements; and (6) measurement quality indicators (method validation, calibrations, etc.).

Step 3. Data Entry Preparation. Batch input files must conform with predefined formats as follows: (1) mandatory records; (2) record identifiers; (3) required number of items in a record; (4) defined data types, length, letter case, and allowable codes; (5) defined data ranges; and (6) scope of the database.

Step 4. Data Entry Insertion. Incoming entries to the database must adhere to the following types of database integrity criteria: (1) uniqueness of records in a table; (2) consistency of records among different database tables; (3) correct data types and lengths of columns; (4) predefined data ranges; and (5) predefined property and description codes.

Step 5. Anomaly Detection. Property values must be in good agreement with the following defined criteria: (1)

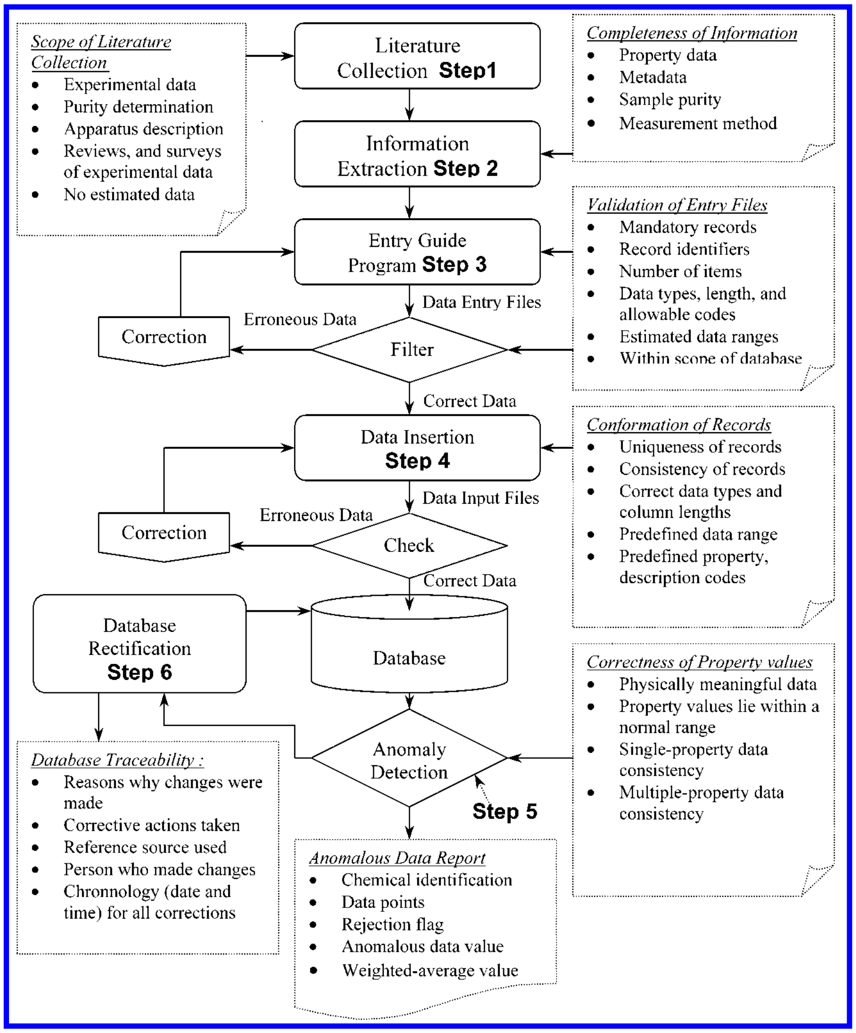
Scope of Literature Collection
• Experimental data
• Purity determination
• Apparatus description
• Reviews, and surveys of experimental data
• No estimated data
Literature Collection Step1
Completeness of Information
• Property data
• Metadata
• Sample purity
• Measurement method
Information Extraction Step 2
Validation of Entry Files
• Mandatory records
• Record identifiers
• Number of items
• Data types, length, and allowable codes
• Estimated data ranges
• Within scope of database
Entry Guide Program Step 3
Data Entry Files
Correction
Erroneous Data
Filter
Correct Data
Data Insertion Step 4
Conformation of Records
• Uniqueness of records
• Consistency of records
• Correct data types and column lengths
• Predefined data range
• Predefined property, description codes
Data Input Files
Correction
Erroneous Data
Check
Correct Data
Database Rectification Step 6
Database
Correctness of Property values
• Physically meaningful data
• Property values lie within a normal range
• Single-property data consistency
• Multiple-property data consistency
Anomaly Detection
Step 5
Database Traceability :
• Reasons why changes were made
• Corrective actions taken
• Reference source used
• Person who made changes
• Chronnology (date and time) for all corrections
Anomalous Data Report
• Chemical identification
• Data points
• Rejection flag
• Anomalous data value
• Weighted-average value

Figure 2. A systematic approach for DQA.

physically meaningful data; (2) property values that lie within the normal range; (3) single-property data consistency; and (4) multiple-property data consistency.

Step 6. Database Rectification. The following information must be furnished for database rectification activities: (1) reasons why significant changes were made; (2) corrective actions taken; (3) reference source used; (4) person making changes; and (5) chronology (date and time) for all corrections.

These rules have been evolving into a knowledge base for supporting quality improvement processes and have been tested, adjusted, and refined in practice in order to reach the comfortable level of quality control.

IV. A SYSTEMATIC APPROACH FOR DATA QUALITY ASSURANCE

Goals for data quality assurance cannot be met by relying solely on human reliability and responsibility and cannot be achieved passively through remedial action only. TRC has implemented a knowledge-based quality improvement system, illustrated in Figure 2, that includes the main functions of (a) error prevention, (b) database integrity enforcement, (c) scientific data integrity protection, and (d) database rectification. The objective of the system is to implement data quality assurance guidelines and to ensure that information handled in every step conforms to these rules.

Error Prevention. Error prevention aims at enforcing rules of literature collection, information extraction, and data entry preparation. These are the first three steps of database operations, shown at the top of Figure 2. Error prevention is realized in the preparation of data entry files and the following check process.

A complete set of input data for the TRC SOURCE data system contains information concerning the literature citation, compound and chemical system, sample purity, numerical values of property as well as state variables, units, measurement objectives, and methods, and other metadata. It is prepared and stored in a data entry file with a predefined data format and structure according to the rules of steps indicated as 1, 2, and 3 in Figure 2. A filter program written in PERL then checks the file, and any detected errors in data entries are corrected. These are the main tasks that are implemented in the data entry preparation, including the Entry Guide Program block and the following Filter block (Step 3) shown in Figure 2.

These initial steps (1−3, Figure 2) can be very labor intensive, representing some key components of the entire database operation. Due to the complexity of physicochemical, extensive training is required to become a data compiler, whose responsibility is to prepare data entry files. Moreover, expertise in data and measurements is needed for the assessment of uncertainties for each property value. Rejoining

NIST in 2000 with a goal of capturing more than 80% literature coverage in this field, TRC has established a data entry facility with undergraduate engineering students forming the major part of the work force for acceleration of the data entry process and expansion of the data coverage. Crucial questions involved (1) how to ensure data quality at this faster pace with broader coverage and (2) how to eliminate the most errors before entering data entries into the data system. To meet these goals, an interactive Entry Guide Program (EGP), written in Visual Basic, was developed to ensure that all relevant information is captured and that proper formatting requirements are met in the preparation of data entry files. No data without sufficient check and validation are allowed for direct entry into the database. The program serves as a data entry expert in the following ways: assisting extraction of information from literature, simplifying entry preparation procedures, ensuring the completeness of the information extracted, validating the input by setting data definition, and most importantly, automating uncertainty assignment to ensure consistency among compilers.

One of the key features of the EGP program is its error prevention mechanism in the light of rules established for steps 1, 2, and 3 (Figure 2). These are implemented by

a. Enforcing the scope of the database by rejecting invalid document types, data types (nonobserved data), and property types;

b. Requiring completeness of input information, such as mandatory records, predefined record identifiers, and all record items by order-predefined forms;

c. Minimizing manual input through use of pull-down menus, radio or button selections, check boxes, and other graphical means; meanwhile, ensuring the input validity;

d. Eliminating certain common input errors, such as errors of data types, length, letter case, and allowable codes, by performing rigorous checks of consistency with data definition during an edit process.

A compiler's job is to interact with the EGP program to produce data entry files. Since all the operations involve interactions between numerous forms and compilers, it is important to realize the error prevention mechanisms on this stage of the data entry. The compiler is prompted at various points by the program regarding possible errors, then these errors are corrected, and the following steps are continued until the corrected batch files are produced.

The Entry Guide program eliminates most of the errors when preparing data entry files. Nevertheless, some errors might still remain to be checked by the filter program, because a balance is needed between the functional capability and software complexity of the EGP program. For example, the filter program will further reduce errors in numerical values by checking the rationality of each record on a scientific basis. Besides, it double tests the various kinds of errors, such as duplicates within a data set or erroneous codes for properties, variables, or phases; if any are found, a message is displayed. At Step 3 (Figure 2), whether by the EGP program or by the filter program, the codes are checked only for appropriate format and characters, not for specific adherence to those defined in the database. Errors found by the filter program are corrected and rerun repeatedly until no more are found.

Database Integrity Enforcement. Database integrity is enforced in Step 4 (Figure 2), Data Insertion, when data are entered into the database. During this process, the data entry file is automatically converted into several data input files, which are closely associated with individual database tables. These are the files that are loaded directly into TRC SOURCE. Modern relational database systems provide declarative constraints as the mechanism of database integrity protection and prevent invalid data entry into the database tables. If any data in the loaded file violate a predefined integrity constraint, the database returns an error message.

In TRC SOURCE, there are five different types of declarative constraints that may constrain a database column or a database table: Primary Key, Not Null, Unique, Check, and Foreign Key. A Primary Key constraint is applied to eliminate duplicate records in a table; a Not Null constraint guards against a required column value being missing; a Unique constraint is identical to a Primary Key, except that a missing value is allowed; a Check constraint defines a discrete list of values that are valid to a column; and a Foreign Key constraint implements referential integrity, which enforces the relations among columns of different tables. Correctly defining Primary Keys and Foreign Keys enforces effective and consistent relationships among data records and columns and ensures the implementation of the data organization and structure (Figure 1). Check constraints are used extensively when only a limited number of valid entries are possible for a column. Within TRC SOURCE, there are 36 Primary Keys, 92 Foreign Keys, and 294 Check constraints that are assigned to database columns and tables.

When data input files are loaded into the database, the rules identified for Step 4 (Figure 2, Data Insertion) are enforced by the database to test whether the following suspicious records exist.

Duplicate Records. Records that have the same Primary Key as that of a record in the database are rejected by the database. It is necessary to determine the reason for such rejections. Duplicates between input files and the database most likely occur when a part of the data from a document has been previously entered.

Missing Records. The rejected records indicate a violation in defined relationships among records and fields for parent-child tables, usually declared by Foreign Keys. A common cause for this violation is not usually related to a rejected record but rather to a missing record in a parent table.

Out of Range. If certain items in the input file are outside an allowed range of values, they are rejected. This generally means that there exists some kind of error in the input files or that the order of the items is incorrect.

Invalid Codes. Codes that do not conform to the database specifications are invalid.

Records from the data input file that pass these final tests are added to the database. Those that do not are rejected. Errors in the rejected records are corrected, and then the corrected data are again submitted for data loading.

Handling database constraints is one of the key database management arts. How many constraints are reasonable against a table? What is an acceptable speed for the data insertion operation? What checks should be performed by the database and what checks should be performed by an application, such as the Entry Guide Program? Answers to these questions are adjusted in light of database performance.

Table 1. Check Criteria Examples for Property Value Ranges of Pure Organic Compounds in SOURCE

property code	property definition	range of property value
TMN TP	normal freezing point, K; triple point, K	TMN or TP < TBN < Tc; 65 < TMN or TP < 800 K
TB	boiling point temperature, K	TMN or TP ≤ TB <Tc; 80 < TB < 1000 K
TBN	normal boiling point, K	TMN < TBN <Tc;
Tc	critical temperature, K	Tc > TBN > TMN; 190 < Tc < 1000 K; 1.0< Tc/TB < 1.8
Pc	critical pressure, kPa	100 < Pc < 15000 kPa
PV	vapor pressure, kPa	PV < Pc; 50 < T < 1000 K; 0 < PV < 1000 for crystal values (include triple point); 0 < PV < 15000 for liquid values
VDC	critical density, kG/m^3	160 < VDC < 1000
VDN	specific density	500 > VDN< 2500 for saturated liquid
Vc	critical molar volume, m^3/mol	0.00008 < Vc < 0.01 m^3/mol
Z	compressibility factor	Z > 0 (Z = PV/RT)
Zc	critical compressibility factor	0.19 < Zc < 0.34 (Zc = PcVc/RTc)
VPA	adiabatic compressibility, kPa^{1-}	10^{-10} < VPA < 0.01 kPa^{1-}
VTP	coefficient of expansion, K^{-1}	10^{-6} < VTP < 0.02 K^{-1}
HTR	enthalpy of transition, kJ•mol^{-1}	0 ≤ HTR < 200. Here, HTR includes all phase changes: solid to solid (generally, HTR < 100); solid to liquid (generally, HTR < 200); for larger molecules, HTR can be larger than 200.
X, W, V	mole, weight, volume fractions	0 to 1.0
RID, RIX	refractive index	1.0 to 1.8

In our experience, most checks on data completeness and correctness should be conducted through database application programs, with database declarative integrity enforcement as a backup.

Scientific Data Integrity Protection. While the first four steps in the operation of the database, noted in Figure 2, remove most of the errors before data are added to the database, scientific data integrity is further protected by additional automatic checks of numerical values and uncertainties in the database. These checks are termed "Anomaly Detection (Step 5)" in the six major database operations described in Section III and are made at specified time intervals. These checks globally analyze multiple data sets for a given property and check mathematically related data through thermodynamic equations and correlations. Suspicious values are compared against the original documents to determine whether a transcription or interpretation error has been made.

Numerical property values and uncertainties become a focus of anomaly detection at this point for two reasons. One is that the magnitude of the error rate of property values and uncertainties has a significant impact on overall data quality of the database. The other reason is that being able to directly handle database records makes it possible to do various comparisons of multiple data points and sets and to make consistency checks on relationships among different types of properties throughout the entire database. In TRC SOURCE, multiple data points for the same compound/property as well as various types of thermophysical and thermochemical properties are two of the advantages for such further comparisons and anomaly detection.

Basic thermodynamic principles and estimation models are two bases that are used to form scientific data integrity criteria. They prove to be a powerful approach to further identify anomalous values. Three principles deduced from basic thermodynamic theory and empirical observations are used to generate criteria for checking properties in TRC SOURCE. They are as follows: Principle I, each property value must lie in a certain numerical range; Principle II, each property value for a given compound must be in accord with other data for that compound; and Principle III, similar compounds should have similar properties.

Sets of check criteria to identify anomalous values are devised according to the three principles. Examples of simple check criteria for property-value ranges of pure compounds are listed in Table 1. More sophisticated check criteria for properties with multiple data points involve comparison of deviations from the weighted-average value (WAV), which mean that the value is considered as an anomalous value if its deviation from WAV is greater than two times the average deviation of the data set. Similar check criteria are applied to properties as a function of temperature.

Principle II constitutes a single-property consistency check and a multiple-property consistency check. Here the single-property consistency means that property values must lie in a special range for a given property and compound. Three cases may exist. (a) There are multiexperimental data points or sets. For example, a critical temperature (T_c) value of one compound should not deviate largely from WAV of other values of the same compound. (b) There is only one data point. The value should not differ greatly from that of similar molecules or from the estimated value of the best models. (c) Properties are functions of variable(s). For example, all vapor pressure data of a pure compound should be on a smooth curve and cannot be larger than the critical pressure of the compound or less than zero. The multiple-property consistency is used to check mathematical-correlated relationships among properties for a given system. For example, the ratio of critical temperature/normal boiling point of organic compounds should not be less than 1 or larger than 1.8. Another example is that a vapor pressure equation fitted from a set of experimental data for a given compound should entail suitable values for the heat of vaporization.

Let us take a single-property consistency check as an example. As we indicated in Section II, data collection of all duplicate measurements is one of the features of TRC SOURCE; as a result, each property for a given compound in the database may contain one, or two, or several, or up to hundreds of values. In the case where the total number of data points is less than or equal to 5, the value is examined by comparing it with other reliable data sources or with the calculated value from recommended models such as group contribution methods. On the other hand, in the case of multiple data points for a given system/property, all check

Table 2. Partial Report on Anomalous Values for Critical Temperature (T_c)[a]

NO	CASRN	DP	EO	REFKEY	TRCTB	WAV	VALUE	UNC	DEV
1	60-29-7	28	B	1879-saj-0	466.7	466.67	463.15	3.0	3.5
2	60-29-7	28	B	1891-hei-0	466.7	466.67	472.45		5.8
3	60-29-7	28	B	1892-bat-0	466.7	466.67	470.15	3.0	3.5
4	64-17-5	35	B	1822-del-0	513.92	514.49	531.9	0.2	17.4
5	64-17-5	35	B	1879-saj-0	513.92	514.49	507.5		7.0
6	64-17-5	35	B	1880-han-hog-0	513.92	514.49	507.8		6.7
7	64-17-5	35	B	1880-han-0	513.92	514.49	507.5		7.0
8	64-17-5	35	B	1881-sch-6	513.92	514.49	506.85	8.0	7.6
9	64-17-5	35	B	1882-han-0	513.92	514.49	508.6		5.9
10	64-17-5	35	B	1891-sch-1	513.92	514.49	507.5	7.0	7.0
11	64-17-5	35	B	1904-cri-0	513.92	514.49	518.55	5.0	4.1
12	64-17-5	35	B	1910-pri-1	513.92	514.49	523		8.5
13	101-81-5	6	B	1957-gla-rul-0	770	773.29	845.65	7.0	72.3
14	101-81-5	6	B	1980-wie-kob-0	770	773.29	760	2.0	13.4
15	101-81-5	6	B	1985-smi-0	770	773.29	767	10	6.4
16	101-81-5	6	B	1992-ste-0	770	773.29	760	8.0	13.4
17	110-54-3	41	B	1942-ipa-mon-0	507.5	507.56	500.65	2.0	6.9
18	110-54-3	41	C	1993-nik-pav-0	507.5	507.56	504.5	4.0	3.1
19	111-87-5	8	B	1906-bro-0		652.34	658.6		6.3
20	111-87-5	8	B	1943-fis-rei-0		652.34	657.8		5.5
21	111-87-5	8	B	1966-efr-0		652.34	668	2.0	15.7
22	811-97-2	16	B	1988-wil-bas-0	374.26	374.15	673.65	0.25	300
23	2551-62-4	5	A	1951-ata-sch-0	318.72		591.86	0.05	151
24	2551-62-4	5	B	1985-mat-sch-0	318.72		568.8	0.7	128

[a] 1. Check criteria: a. For all compounds, the value $<$ 80 K or the value $>$ 1000 K. b. For less than five data points, use the recommended models and other reliable data source. c. For more than five data points, the deviation $>$ 2 × average deviation. 2.Title description: NO, sequence number of anomalous records; CASRN, Chemical Abstract Service Registry Number; DP, total number of data point for a particular compound; EO, experimental objective—A, state of the art at the time of publication; B, the measurement as a principal objective; and C, the measurement as some secondary purpose. REFKEY, TRC SOURCE reference key used to identify measurement reports; TRCTB, evaluated data from NIST/TRC TABLE (K); WAV, weighted-average value (K); VALUE, property value of the data point being examined (K); UNC, property uncertainty of data point being examined (K); DEV, deviation of VALUE from WAV (K). 3. Compound names: 60-29-7, diethyl ether; 64-17-5, ethanol; 101-81-5, diphenylmethane; 110-54-3, hexane; 111-87-5, 1-octanol; 811-97-2, 1,1,1,2-tetrafluoroethane; 2551-62-4, sulfur hexafluoride

criteria are implemented through mathematical models devised in procedural SQL programs, called Anomaly Detection Procedures (ADPs). The models employ a weighted-average method that is based on several weight factors, among which, besides a primary weight factor (numerical uncertainty), some additional weight factors (year of publication, sample purity, and experimental objective) are also contributing factors in calculating WAV. The factors are adopted in consideration of possible major problems involving greater deviations, missing uncertainty values, and incorrectly assigned uncertainties, presented in a noncritically evaluated data set. To calculate a relatively precise WAV from such a data set, we devised an algorithm, which consists of four steps, utilized in ADPs. In Step 0, simple rules, such as property value range check partially listed in Table 1, are used to reject obviously erratic data. Step 1 calculates a WAV using assigned uncertainties and predefined uncertainties for data, which have not been assigned an uncertainty. The difference between each property value and WAV obtained at Step 1 is taken as a new uncertainty in second weighted-average calculation, conducted in Step 2. If some new uncertainty value is significantly greater, for example, uncertainty (T_c) $>$ 30 K or uncertainty (P_c) $>$ 600 kPa, this property value will be eliminated from weighted-average calculation in order to exclude absurd values that obviously deviate from most other data values. In doing so, a new average value is obtained, based upon the remaining selected data and calculated uncertainties. In Step 3, the process of Step 2 is repeated, except that a stricter criterion, such as uncertainty (T_c) $>$ 3 K or uncertainty (P_c) $>$ 300 kPa, is chosen to eliminate more data points from calculations. This is called third weighted-average calculation that generates our final WAV. Any property values with a discrepancy from WAV larger than two times the average deviation are considered as anomalous values.

The ADP programs run every month, producing reports in which anomalous records are identified and summarized. Summary information includes the chemical compound identification, the number of data points analyzed, the experimental objective, the TRC SOURCE reference key, the WAV, the property value being identified as an anomalous value, the uncertainty, and the discrepancy between WAV and the property value. WAV data and uncertainty ranges are provided to aid in the assessment of property values and, therefore, to determine whether further investigation is needed.

Anomaly detection in TRC SOURCE can be illustrated by our effort in analyzing data quality for critical temperature (T_c) data. TRC SOURCE includes 95% of the experimental values for critical temperature published since 1822. There are some 1900 experimental property values of critical temperature reported for 630 pure organic compounds. Running the ADP against all T_c records generated a report that included 366 anomalous records, for which a partial report is included in Table 2. In this report, Column VALUE contains existing database values identified as anomalous values because all the values in the DEV column are two times greater than the average deviation, and Column UNC lists their uncertainty values assigned by TRC (As shown in the table, for some historical reason, not every data value has an assigned uncertainty.). Values in both columns are to be examined and adjusted. For a larger data set, WAV can be generated in accord with the value from reliable data sources such as NIST/TRC TABLE or IUPAC Project 2000-026-1-100 and Project 121/10/87[17] (Column TRCTB includes values from the NIST/TRC Table database as an additional data source.), while for some compounds with less than five data points, the reliable data sources and the recommended models are used to determine anomalous values.

In the following labor-intensive process, most original articles and reports involved in the anomalous records were located and reviewed. One of the difficulties in the process was to solve semantic ambiguities in scientific terminology and presentation across a large time span and across many languages. The main goal was to identify a cause for each anomalous record. Of the 366 anomalous records, about 70 were found to be due to data-entry errors, such as typographical error, interpretation error, unit conversion error, or an invalid type of data, i.e., the data were estimated or correlated and should not have been included in the database (Rows 2, 10, 13, 14, 17, 21, 22, 23, and 24 of Table 2 belong to this category.). For some 216 anomalous records, no errors were found in T_c values. However, because the values were further dispersed from the WAV values, adjustments were made to uncertainties to reflect the T_c accuracy realistically (Rows 4, 5, 6, 7, 9, 12, 19, and 20 of Table 2 fit into this category.). Some 80 records in the list were recognized to be error free (Rows 1, 3, 8, 11, 15, 16, and 18 of Table 2 represent this category.). In summary, an error rate of about 3.6% error rate was found for T_c records in TRC SOURCE.

Database Traceability. Database Rectification (Step 6, Figure 2) is the last stage of the quality-improvement process. As a long-term strategy of quality assurance for TRC SOURCE, data-quality schemas are needed for tracking modification activities for each numeric table. To increase the traceability of TRC SOURCE values, rectifications and deletions of data within the database are recorded in special tables, called “traceability tables”. For example, two traceability tables, “correction” and “deletion”, were created for each numerical table in the database to record a history of significant modifications. The “correction” table describes modifications of numerical property values and uncertainties, including the changes made, reasons for the changes, the referenced source used in data checking, the person making the changes, and the time of the modification. Similar information is stored in the “deletion” table except that in this case the information is related to the data records deleted from numerical tables. The two tables contain important information about the reasons why certain data were deleted or corrected in order to avoid repetitions of the same errors and to guide subsequent data processing.

V. CONCLUSION

The successful development of large-scale scientific databases requires in-depth consideration of many complex issues, such as data storage and depository, security, database design, and data processing, etc. At the heart of database management, however, is the issue of data quality assurance. In the specific case of the TRC SOURCE data system, we have shown that through application of guided data entry, database integrity constraints, exploitation of thermodynamic principles employed in anomaly detection programs, and strict traceability protocols, an effective and robust data-quality assurance system is developed.

REFERENCES AND NOTES

(1) Bhat, T. N.; Bourne P.; Feng, Z.; Gilliland, G.; Jain S.; Ravichandran V.; Schneider, B.; Schneider K.; Thanki N.; Weissig H.; Westbrook J.; Berman H. M. The PDB Data Uniformity Project. *Nucleic Acids Res.* **2001**, *29*(1), 214−218.

(2) Kuhn, P.; Deplanque, R.; Fluck, E. Criteria of Quality Assessment for Scientific Databases. *J. Chem. Inf. Comput. Sci.* **1994**, *34*(3), 517−519.

(3) Kaufman, J. G. Standards for Computerized Material Property Data − ASTM Committee E-49. *Computerization and Networking of Materials Data Base, ASTM STP 1017;* Glazman, J. S., Rumble J. R., Jr., Eds.; American Society for Testing and Materials: Philadelphia, 1989; pp 7−22.

(4) Taylor, B. N.; Kuyatt, C. E. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*; Technical Note 1297; National Institute of Standards and Technology; Washington, DC, 1994; 24 p.

(5) Croarkin, M. C. Statistics and Measurements. *J. Res. Natl. Inst. Stand. Technol.* **2001**, *106*(1), 279−292.

(6) Frenkel, M.; Dong, Q.; Wilhoit, R. C.; Hall K. R. TRC SOURCE Database: A Unique Tool for Automatic Production of Data Compilations. *Int. J. Thermophys.* **2001**, *22*(1), 215−226.

(7) Yan, X.; Dong, Q.; Frenkel, M.; Hall, K. R. Window-Based Applications of TRC Databases: Structure and Internet Distribution. *Int. J. Thermophys.* **2001**, *22*(1), 227−241.

(8) Wilhoit, R. C.; Marsh, K. N. Future Directions for Data Compilation. *Int. J. Thermophys.* **1999**, *20*(1), 247−255.

(9) Marsh, K. N.; Wilhoit, R. C. Thermodynamics Research Center Databases. In *Transport Properties of Fluids. Their Correlation, Prediction and Estimation*; Millat, J., Dymond, J. H., Nieto de Castro, C. A., Eds.; Cambridge University Press: Cambridge, 1996.

(10) Wilhoit, R. C.; Marsh, K. N. Automation of Numerical Data Compilations. *J. Chem. Inf. Comput. Sci.* **1989**, *29*(1), 17−22.

(11) (a) Yan, X.; Dong, Q.; Hong, X.; Frenkel, M. NIST/TRC Table Database: NIST/TRC Standard Reference Database 85, WinTable 1.5, 2001; http://www.nist.gov/srd/nist85.htm. (b) Yan, X.; Dong, Q.; Hong, X.; Frenkel, M. NIST/TRC Table Database: NIST/TRC Standard Reference Database 85, WinTable 1.5, 2001; http://www.nist.gov/srd/webguide/nist85/85_1.htm. (c) Yan, X.; Dong, Q.; Hong, X.; Frenkel, M. NIST/TRC Table Database: NIST/TRC Standard Reference Database 85, WinTable 1.5, 2001; http://trc.nist.gov/database/Table/wintable.htm.

(12) (a) Frenkel, M.; Wilhoit, R. C.; Marsh, K. N.; Dong, Q.; Hardin, G. R. NIST/TRC Vapor Pressure Database: NIST/TRC Standard Reference Database 87, 2001; http://www.nist.gov/srd/nist87.htm. (b) Frenkel, M.; Wilhoit, R. C.; Marsh, K. N.; Dong, Q.; Hardin, G. R. NIST/TRC Vapor Pressure Database: NIST/TRC Standard Reference Database 87, 2001; http://www.nist.gov/srd/webguide/nist87/87_1.htm.

(13) Documentation for the TRC SOURCE Database; Thermodynamics Research Center, NIST: Boulder, CO, 2001.

(14) Whiting, W. B. Effects of Uncertainties in Thermodynamic Data and Models on Process Calculations. *J. Chem. Eng. Data* **1996**, *41*, 935−941.

(15) Vasquez, V. R.; Whiting, W. B. Uncertainty and Sensitivity Analysis of Thermodynamic Models Using Equal Probability Sampling (EPS). *Computers Chem. Eng.* **2000**, *23*(11/12), 1825−1838.

(16) Data Quality Assurance Guidelines for the SOURCE Database; Thermodynamics Research Center, NIST: Boulder, CO, 2001.

(17) IUPAC Project 2000-026-1-100, Critical Compilation of Vapor Liquid Critical Properties, http://iupac.chemsoc.org/projects/2000/2000-026-1-100.html.

CI010118E