Pure Appl. Chem., Vol. 77, No. 8, pp. 1349–1367, 2005.
DOI: 10.1351/pac200577081349

Global communications and expert systems in thermodynamics: Connecting property measurement and chemical process design*,†

M. Frenkel‡

Thermodynamics Research Center (TRC), National Institute of Standards and Technology (NIST), 325 Broadway, Boulder, CO 80305-3328, USA

Abstract: Unprecedented growth in the number of custom-designed software tools for engineering applications has created an interoperability problem between the formats and structures of thermodynamic data files and required input/output structures designed for application software products. Various approaches for standardization of thermophysical and thermochemical property data storage and exchange are analyzed in this paper. Emphasis is made on the development of the XML-based IUPAC standard for thermodynamic data communications: ThermoML. A new process for global data submission and dissemination in the field of thermodynamics based on ThermoML and Guided Data Capture software is described.

Establishment of the global submission and dissemination process for thermodynamic data lays the foundation for implementation of the new concept of dynamic data evaluation formulated at NIST/TRC, which requires the development of large electronic databases capable of storing essentially all "raw" experimental data known to date with detailed descriptions of relevant metadata and uncertainties. The combination of these databases with expert software designed primarily to generate recommended data based on available "raw" experimental data and their uncertainties leads to the possibility of producing data compilations automatically "to order", forming a dynamic data infrastructure. Implementation of the dynamic data evaluation concept for pure compounds in the new NIST/TRC ThermoData Engine software is discussed.

Keywords: Thermodynamics; expert systems; communications; data; evaluation; databases.

INTRODUCTION

Thermodynamic property data represent a key foundation for development and improvement of all chemical process technologies. Design and implementation of any chemical process consists typically of six major steps: "raw" data collection, critical data evaluation, process simulation, equipment sizing, pilot-scale implementation, and full-scale implementation. The overall process quality can be characterized by three major factors: (1) the yield of the targeted chemical product, (2) the nature and amount of waste produced (environmental impact), and (3) the amount of energy consumed. Each of these three process-quality "components" greatly depend on the quality of critically evaluated thermodynamic data (if available) used in the process-simulation step. Incompleteness or poor data quality often lead to er-

*Paper based on a presentation at the 18th IUPAC International Conference on Chemical Thermodynamics (ICCT-2004), 17–21 August 2004, Beijing, China. Other presentations are published in this issue, pp. 1297–1444.

‡E-mail: frenkel@boulder.nist.gov

roneous equipment selections (pumps, reactors, heat exchangers) which preclude further process improvements at the pilot- and mass-scale stages, resulting in undesirable and possibly enormous economic losses. Lack of thermodynamic information often makes it impossible to simulate new chemical processes at all, necessitating numerous empirical and expensive trial-and-error iterations for process optimization. This commonly results in significant increases in time and resources used without any assurance of finding the true optimal conditions for the process. This situation is quite typical, especially within rapidly developing industries such as pharmaceutics, specialty chemicals, and biotechnology. In addition to the highly practical field of process development, high-quality thermodynamic property data are frequently essential prerequisites in the search for new relationships between properties of chemical systems, and constitute the basis of the scientific discovery process.

In this paper, major developments are discussed in two areas crucial for efficient thermodynamic data management and critical evaluation (critically evaluated data are defined in [1]): thermodynamic data communications and thermodynamic data expert systems.

THERMODYNAMIC DATA COMMUNICATIONS

Review of standardization efforts

Establishment of efficient means for thermodynamic data communications is absolutely critical for provision of solutions to such technological challenges as elimination of data processing redundancies and data collection process duplication, creation of comprehensive data storage facilities, and rapid data propagation from the measurement to data management system and from the data management system to engineering applications. Taking into account the diversity of thermodynamic data and numerous methods of their reporting and presentation, standardization of thermodynamic data communications is very complex.

Efforts to develop a standard for thermophysical and thermochemical property data exchange [2] were initiated in the early 1980s, reflecting a new trend in data collection through design of electronic databases, which became possible due to the rapid development of computer technology. In the time period 1985 to 1987, the Thermodynamics Research Center (TRC, then with Texas A&M University) developed the first prototype of such a standard called COSTAT (Codata STAndard Thermodynamics) [3]. This prototype was discussed extensively among numerous institutions worldwide through the auspices of CODATA. This effort played an important role in establishing the necessity of a standard and in formulating the basic principles that must be incorporated. Practical implementation of COSTAT was hindered significantly by limitations of software tools available at the time.

At the beginning of the 1990s, Global CAPE Open (initially Cape Computer-Aided Process Engineering Open) technology was initiated [4]. The Global CAPE Open project was established to develop standards for interfaces of software components of a process simulator. The main objective of the project was to enable native components of a simulator to be replaced by those from another source with minimal effort in as seamless a manner as possible. This approach was proven successful; however, the Global CAPE Open approach is not naturally modular, and therefore, implementation of any modifications of the thermodynamic data representation requires significant programming effort.

In 1998, TRC was selected as one of four data centers worldwide to be a part of a similar project funded by CODATA (IUCOSPED Task Group). A number of experts from NIST actively participated in this project, which ended in 2002. This project led to the development of the SELF [5] files closely associated with the ELDATA electronic journal formats. Though the project played a positive role in attracting the attention of the international scientific community to core issues related to thermophysical data standardization, the final outcome has profound limitations related to its noncomprehensive and nonsystematic nature.

In 1999, the Design Institute for Physical Property Data (DIPPR) under the auspices of the American Institute of Chemical Engineers (AIChE) initiated Project 991 to develop a thermophysical

property data exchange standard focusing primarily on the industrial application of the extended version of the CAPE, Physical Property Data eXchange neutral file format (PPDX) [6], and later developed its XML version PPDXML.

ThermoML

In 2003, NIST/TRC in cooperation with DIPPR developed ThermoML, an XML (Extensible Markup Language)-based approach for storage and exchange of thermophysical and thermochemical property data [1,2,7]. XML technology [8], fully developed within the last five years, provides significant advantages for the development of standards for data exchange such as its "native" interoperability based on ASCII code, modular nature, and transparent readability by both humans and computers. From a practical standpoint, it is also very important that this technology is currently supported by both the software and hardware industries. Among other X-markup languages, CML (XML for chemistry) [9] and MatML (XML for primarily mechanical properties of the materials) [10] are most closely related to ThermoML.

The development of ThermoML at TRC is a result of further improvement of the basic principles defined in COSTAT, as well as more than 50 years of experience by TRC and data groups at NIST in thermophysical property data collection and dissemination. This experience includes maintenance of the largest relational archival thermophysical property experimental data system (SOURCE [11]), which currently includes more than 120 properties for pure compounds, mixtures, and chemical reactions. In this section, the major conceptual features and structural elements of ThermoML are summarized. Details of ThermoML were previously described [1,2,7].

The ThermoML structure represents a balanced combination of hierarchical and relational elements. The ThermoML schema structure explicitly incorporates structural elements related to basic principles of phenomenological thermodynamics: thermochemical and thermophysical (equilibrium and transport) properties, state variables, system constraints, phases, and units. Meta- and numerical data records are grouped into "nested blocks" of information corresponding to data sets. The structural features of the ThermoML metadata records ensure unambiguous interpretation of numerical data and allow data quality control based on the Gibbs Phase Rule. Implementation of the Gibbs Phase Rule is a reflection of long-standing traditions and practices at NIST for assuring the highest quality in data, and would provide users with an indication of inconsistencies in thermodynamic data before the data are deposited into a data storage facility [12]. Moreover, some detailed information included in the metadata records could serve as a background for independent assessment of uncertainties, which could be propagated into uncertainties of physical parameters for reaction streams, and consequently, provide an opportunity for quantitative characterization of the quality of a chemical process design [13]. Commonly accepted IUPAC-based terminology is used as the foundation for metadata and numerical data tagging. ThermoML capitalizes on the fact that XML files are essentially textual files and can, in principle, be interpreted without customized software. This is particularly important in generating files corresponding to data directly submitted to peer-reviewed journals by scientists and engineers, who require simple verification that their data have been represented accurately. In addition, the self-explanatory approach and very limited use of abbreviations minimizes the time necessary for users to understand the schema and to convert the ThermoML formatted data with customized software or commercial XML parsers.

By design, there is only one unit selected for each property covered by ThermoML. These units are SI-based. For a number of properties, the selected units are multiples of SI units to ease interpretation of numerical values. Unit tagging is explicitly propagated to every numerical data point in a ThermoML file as a part of each property name, thus minimizing the possibility of unit misinterpretation. Various methods of numerical data representation commonly used in the publication of experimental property data (e.g., direct, difference from values at a reference state, ratio of the value to that at a reference state, etc.) are incorporated into ThermoML.

ThermoML covers essentially all experimentally determined thermodynamic and transport property data (more than 120 properties) for pure compounds, multicomponent mixtures, and chemical reactions (including change-of-state and equilibrium). The primary focus at present is molecular compounds. Generally, all three major types of thermodynamic data—experimental, predicted, and critically evaluated—are within the scope of ThermoML.

ThermoML consists of four major blocks, as shown in Fig. 1 [2]: *Citation*, describing the source of the data; *Compound*, characterizing the chemical system; *PureOrMixtureData*, providing information for meta- and numerical data for a pure compound or multicomponent mixture; and *ReactionData*, providing information for meta- and numerical data for a chemical reaction with thermodynamic state change or in a state of chemical equilibrium.

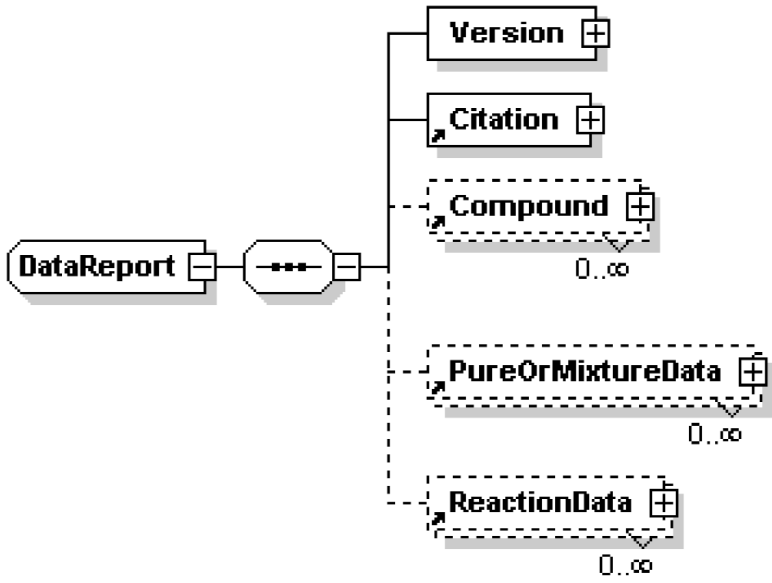

Fig. 1 Major components of ThermoML [2].

The *Citation* block provides exhaustive options for descriptions of data source types such as articles, books, conference proceedings, patents, Web sites, internal institutional communications, etc., and conforms with ISO 690:1987 [14] and ISO 690-2 [15] for the identification and description of information resources, including those in electronic format.

The *Compound* block contains provisions for all commonly used chemical identifiers, such as elemental formula, various types of chemical name, Chemical Abstract Service Registry Number (CASRN), and SMILES notation. ThermoML will also include the IUPAC–NIST Chemical Identifier (INChI) [16], once development is completed. The description for every compound is linked to a description of the sample used in the measurements with indication of its initial purity, purification methods used, final purity, and the methods used to determine it [2].

Upper-level major subelements of the metadata in the *PureOrMixtureData* block (Property, PhaseID, Constraint, and Variable [1]) reflect the elements of terminology related to the Gibbs Phase Rule. Property, in turn, can be characterized with the primary phase, type of numerical data presentation, reference phase and state, standard state, and identification of the measures of uncertainty used.

Definitions and descriptions of all quantities related to the expression of uncertainty in ThermoML [7] conform to the *Guide to the Expression of Uncertainty in Measurement* ISO [17]. These ISO recommendations were adopted with minor editorial changes as the *U.S. Guide to the Expression of Uncertainty in Measurement* [18]. Reference [18] is commonly referred to by its abbreviation—the

GUM. Reference [19] is assumed equivalent to reference [18] and is commonly referred to as the Guide. The historical development of these recommendations beginning in 1977 is described in the Guide. The recommendations have been summarized in *Guidelines for the Evaluation and Expression of Uncertainty in NIST Measurement Results*, which is available via free download from the Internet [20]. All properties, variables, and constraints in ThermoML can be characterized with values of the standard uncertainty, as well as with various measures of precision such as repeatability, measuring device specification, and deviations from fitted curves [7]. In addition, the properties can be characterized with the most comprehensive measure of the uncertainty, the combined uncertainty, which includes the impact of all sources of uncertainty including those propagated from uncertainties for variables and constraints.

All properties in *PureOrMixtureData* block are divided into 10 groups. Within each group, every property is characterized with the property name, as well as with either the experimental method used (selected from a predefined list or identified independently) or with structural elements providing information on details related to property prediction or critical evaluation [1]. The **Prediction** subelement provides information on the type of predictive method, its name, a brief description, and original sources describing it. The types of predictive methods cover a great variety of predictive techniques from the most simple, such as group contributions, to the most complex, such as ab initio [1]. The **CriticalEvaluation** subelement provides coverage for three major types of thermodynamic data critical evaluation: critical evaluation of single-property data, simultaneous critical evaluation of multiple related property data, and critical data evaluation with the use of an equation of state (simultaneously evaluating all property data).

ThermoML structure contains the elements necessary to store and exchange information related to fitting equations [1]. Storage of associated covariance matrixes, which provide the measure of uncertainty for parameters of the equations, is also accommodated. The power of XML technology and its modular nature is illustrated here in communication between two different XML languages, ThermoML and MathML [21]. The ThermoML data file includes the identities of all variables, fitted parameters, and constants that are required for a particular equation representation, but does not contain any mathematical expressions. One element of the ThermoML data file is a URL used to specify the Internet location (URL) of the full equation definition. The ThermoMLEquation schema is designed for storage and exchange of equation definitions with full mathematical content included through importation of the MathML schema. ThermoMLEquation is a general schema for the definition of any type of equation for representation of thermophysical and thermochemical properties. An equation definition file is created for definition of a particular equation. Care must be taken by the ThermoML file creator to ensure that the identities of the property, variables, constraints, and equation parameters and constants are correctly matched in the ThermoML data file and the ThermoMLEquation file.

At present, MathML is used in conjunction with ThermoMLEquation strictly for communication of mathematical content (i.e., to communicate mathematical meaning). However, MathML can also be used for transfer of information concerning presentation, making it possible to include thermodynamic property symbols, which are in full accord with IUPAC recommendations provided in the Green Book [22].

The structure of the *ReactionData* block is similar to that of the *PureOrMixtureData* block with the distinct difference of the use of the reaction participant information element in the *ReactionData* block instead of the mixture component information element in the *PureOrMixtureData* block. In addition, the *ReactionData* block includes information related to the stoichiometric coefficients of the reaction. The current ThermoML schema [23] has been extensively validated with more than 9000 data sets of experimental, predicted, and critically evaluated thermodynamic data from more than 7500 original sources.

Role of IUPAC

In 2002, IUPAC approved project 2002-055-3-024, XML-based IUPAC Standard for Experimental and Critically Evaluated Thermodynamic Property Data Storage and Capture [24,25], and established a Task Group to conduct it as one of the activities of the Committee on Printed and Electronic Publications [26]. The objective of the Task Group is to create an XML-based dictionary for storage and exchange of thermophysical and thermochemical data based on fundamental principles of phenomenological thermodynamics covering a wide variety of systems, including pure chemical compounds, multicomponent mixtures, and chemical reactions. Upon completion of the project, the developed dictionary and corresponding XML schema could become internationally accepted as a standard for thermodynamic data storage and exchange. At its meeting in January 2004 [27], the Task Group accepted ThermoML as the framework of the emerging IUPAC standard and approved the establishment of the "ThermoML" namespace for it [28].

Global data delivery process

As discussed above, there is a great demand for the establishment of efficient global data delivery processes. Until recently, such a process did not exist in the field of thermodynamics. In fact, there are only two well-known processes of this nature outside the field of thermodynamics: submission and retrieval of protein structures from the Protein Data Bank (PTB) [29] and submission and retrieval of crystal structures for smaller molecules from the Cambridge Structural Database (CSD) [30].

It is clear that establishing a global data delivery process is a very challenging task in comparison with the PTB and CSD processes because of the necessity to communicate information related to the hundreds of thermophysical, thermochemical, and transport properties commonly reported. Moreover, communicating these property data is further complicated by the extensive system of thermodynamic metadata (variables, constraints, phases, methods, uncertainties) required. This complexity necessitated development of a software infrastructure to support global delivery process for thermodynamic data.

In order to address this need, Guided Data Capture (GDC) software was developed at TRC [31,32]. GDC serves as a data capture expert by guiding extraction of information from the literature, assuring the completeness of the information extracted, validating the information through data definition, range checks, etc., and guiding uncertainty assessment to assure consistency between compilers with diverse levels of experience. A key feature of the GDC software is the capture of information in close accord with customary original document formats. The GDC architecture is designed to detect inconsistencies and errors in reported data (erroneous compound identifications, typographical errors, etc.), resulting in improved integrity of the captured data over that given in the original sources.

The compiler's main interactions with the GDC involve a navigation tree, which provides a visual representation in accord with the hierarchical structure of the batch data file as it is created. Each node of the tree corresponds to a record in the batch data file structure. Management of records including deletion, addition, and editing is accomplished through interactions with the navigation tree. Numerical values are not shown explicitly in the tree, but may be accessed through the property-specification nodes. Lists of established field values (journal title abbreviations, compound identifiers, properties, units, phases, experimental methods, etc.) are stored in a local database, which is a part of the GDC software. Selection of field values by the data compiler is achieved through single-value or multiple-selection lists of the predefined values, which prevent many simple errors. All predefined lists are prioritized to speed access. Keyboard input is never used for direct input of coded information, which eliminates typographical errors. Keyboard input to GDC is provided exclusively for entry of isolated numerical values, general comments, document titles, and new chemical and author names. Most numerical values are captured through electronic means (PDF files, spreadsheets, etc.) and rarely require manual input. All other input is accomplished through predefined menus, check boxes, or other controlled se-

lection processes. The GDC data processing operation encompasses both metadata and numerical data, and provides graphical representation of the numerical data (Fig. 2).

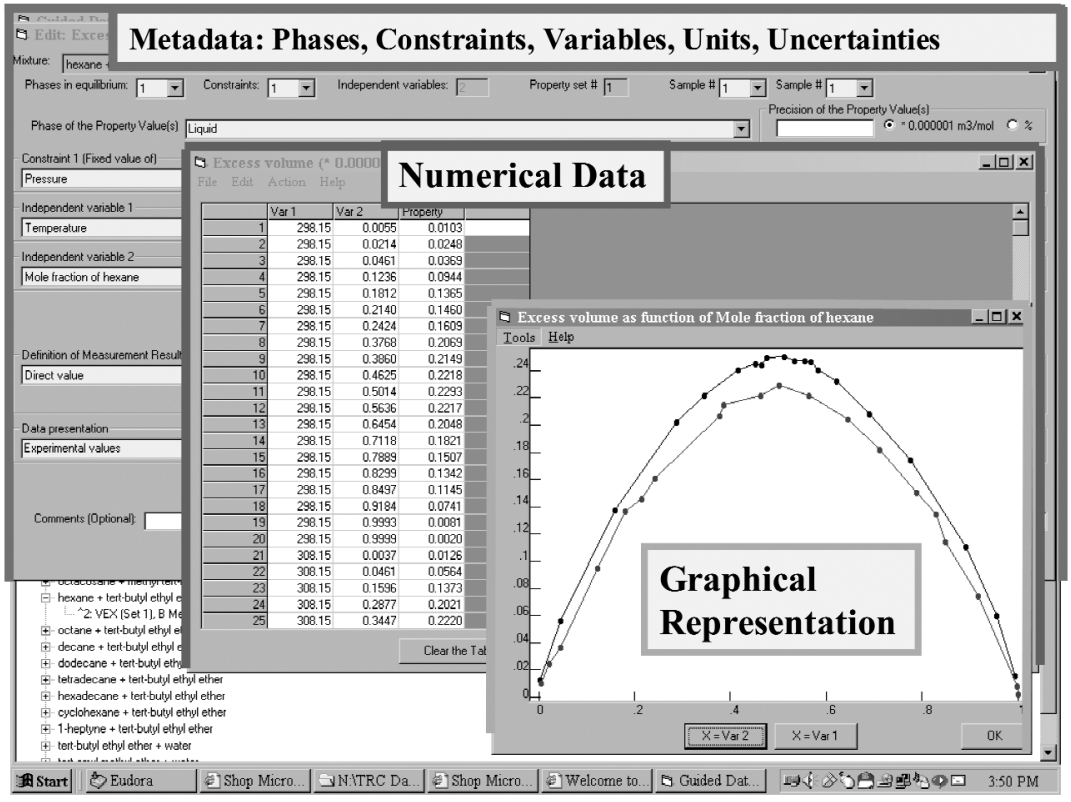

Fig. 2 Collage of the major screens of the GDC software for description of metadata, numerical data, and graphical representation of the numerical data.

The combination of the GDC software, ThermoML schema, NIST/TRC Data Entry Facility organizational and networking structure, and Web dissemination operation for the ThermoML files provides a foundation for the establishment of a global communication process for thermophysical and thermochemical property data (Fig. 3) [33]. Following the peer-review process, authors of submitted manuscripts are requested by the journal editors to download and use the GDC software to capture the experimental property data that have been accepted for publication. The output of the GDC software is an electronic data file (a plain text file), which is submitted directly to TRC. After additional consistency checks at the TRC Data Entry Facility, these electronic data files are converted into ThermoML format with software (TransThermo) developed at TRC. During this process, potential data inconsistency problems are communicated back to authors and editors for their expeditious resolution prior to publication. Upon release of the manuscript for publication, the ThermoML files are posted on the public-domain TRC Web site with unrestricted public access [34]. This procedure was first established formally by the *Journal of Chemical and Engineering Data* [35,36]. In 2004, the *Journal of Chemical Thermodynamics* joined this operation [37,38], and expansions of the operation to other journals in the field, such as *Fluid Phase Equilibria*, *Thermochimica Acta*, and the *International Journal of Thermophysics*, are expected to be implemented in 2004 and 2005 [27].

Figure 3 illustrates the data delivery process from data suppliers (thermodynamicists reporting results of measurements of thermophysical and thermochemical property data via major journals in the

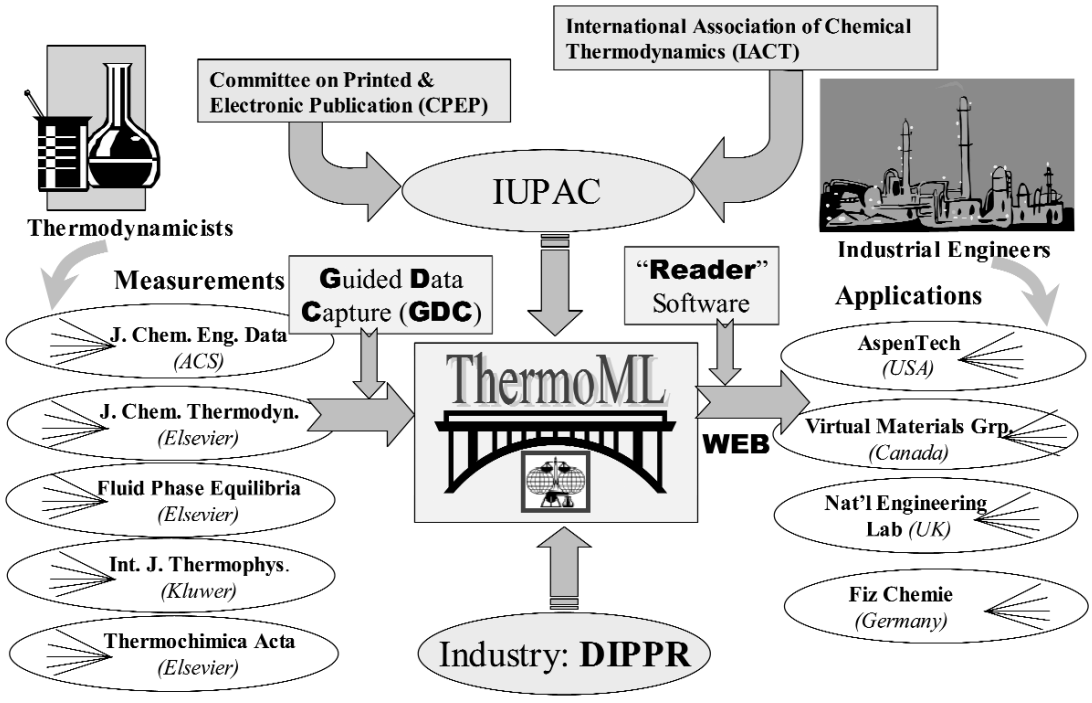

Fig. 3 Global thermodynamic data communication process [39].

field) to data users (chemical engineers via engineering software applications including chemical process design). GDC represents a key software support element for data submission, and ThermoML serves as the media to assure interoperability for propagation of the data across different platforms. ThermoML software "readers" have been developed by a number of organizations in cooperation with NIST/TRC to transfer data from the ThermoML format to customized formats suitable for application software and databases [39]. This process is supported by standardization efforts with the participation of industry (DIPPR), IUPAC, and the International Association of Chemical Thermodynamics (IACT) [40].

THERMODYNAMIC DATA EXPERT SYSTEMS

Definitions

To discuss various concepts of critical data evaluation for thermodynamic data, it is necessary to establish some definitions. There appear to be no such definitions in the literature, which might reflect a consensus of the scientific community. The definitions provided here were suggested recently [1]. These definitions are not intended in any way to serve as a rigorous guide (or "standard") for distinguishing various types of property data, but rather to provide clarification in the discussion of various aspects of the data evaluation process.

True data (hypothetical)

True data are exact property values for a system of defined chemical composition in a specified state. These data have the following characteristics. They are (1) unique and permanent, (2) independent of any experiment or sample, and (3) a hypothetical concept with no known values. The difference between the values of experimental, predicted, and critically evaluated data, on one hand, and true values, on the other, is defined as the *error*. The *error* is never known, however, it is given that it is never zero.

A measure of the quality or confidence in an experimental, predicted, or critically evaluated value is expressed in terms of the "uncertainty" [17–20], which is a range of values believed to include the *true* value with an estimated probability. All data types can and should have associated uncertainty estimates.

Experimental data

Experimental data are defined as those obtained as the result of a particular experiment on a particular sample by a particular investigator. The feature that distinguishes *experimental* data from *predicted* and *critically evaluated* data is use of a chemical sample including characterization of its origin and purity.

Predicted data

Predicted data are defined as those obtained through application of a predictive model or method. Clearly, there is no sample associated with this type of property data.

Critically evaluated data

Like predicted data, there is no chemical sample involved with *critically evaluated* data. The feature that distinguishes *critically evaluated* data from *predicted* data is the involvement of the judgment of a data evaluator or evaluation system. Critically evaluated data are recommended property values generated through consideration of available *experimental* or *predicted* data, or both.

Derived data

Derived data are defined as property values calculated by mathematical operations from other data, possibly including *experimental*, *predicted*, and *critically evaluated* data.

Critical data evaluation

Based on the definitions given above, critical data evaluation can be defined as the process of generation of critically evaluated data obtained from the analysis of available experimental and predicted data, as well as their uncertainties.

Traditional (static) critical data evaluation and problems associated with it

Traditionally, critical data evaluation is an extremely time- and resource-consuming process, which includes extensive use of manpower in data collection, data mining, analysis, fitting, etc. Because of this, it must be performed far in advance of a need within an industrial or scientific application. As a result, in spite of the enormous cost associated with the critical data evaluation process, a very significant part of the existing recommended data has never been used in any meaningful application. This is because data requirements often shift between the initiation and completion of an evaluation project. In addition, it is quite common that by the time the critical data evaluation process for a particular chemical system or property group is complete (sometimes after years of data evaluation projects involving highly skilled data experts), it must be reinitiated because significant new data have become available. This type of slow and inflexible critical data evaluation is defined here as "static". Essentially, all existing data evaluation projects fall into this category. In addition, the quality of results obtained within a static data evaluation process often deteriorates immensely owing to the fact that a mix of experimental, derived, predicted, and critically evaluated data is used to generate the targeted recommended data, commonly producing virtual (i.e., baseless) data instead. Another major problem with existing critical data evaluation is related to absent or ambiguous uncertainty estimates, which make it impossible to propagate the overall data quality into the quality of a chemical process design or model, resulting in enormous economic waste in all stages of the chemical process implementation. These shortcomings have become magnified dramatically within the last 5 to 10 years owing to the significant increase in the amount of reported experimental and predicted thermodynamic data to be analyzed during the critical data evaluation process. Moreover, the static data evaluation process for thermodynamic data has been unable to provide adequate conceptual solutions for chemical process design in such rap-

idly developing fields as biotechnology, where there is a demand for simulation of hundreds of new technologies every year.

Dynamic data evaluation concept

The new concept of dynamic data evaluation has been developed at TRC [41,42]. This concept requires the development of large electronic databases capable of storing essentially all experimental data known to date with detailed descriptions of relevant metadata and uncertainties. The combination of these electronic databases with artificial intellectual (expert-system) software, designed to automatically generate recommended data based on available experimental data, leads to the ability to produce critically evaluated data dynamically or "to order" (Fig. 4). This concept contrasts sharply with static critical data evaluation, which must be initiated far in advance of need. The dynamic data evaluation process dramatically reduces the effort and costs associated with anticipating future needs and keeping static evaluations current. The critically evaluated data produced by the deployment of the dynamic data evaluation concept can rigorously be characterized with their quality assessments providing the ability to propagate reliable data quality limits to all aspects of chemical process design.

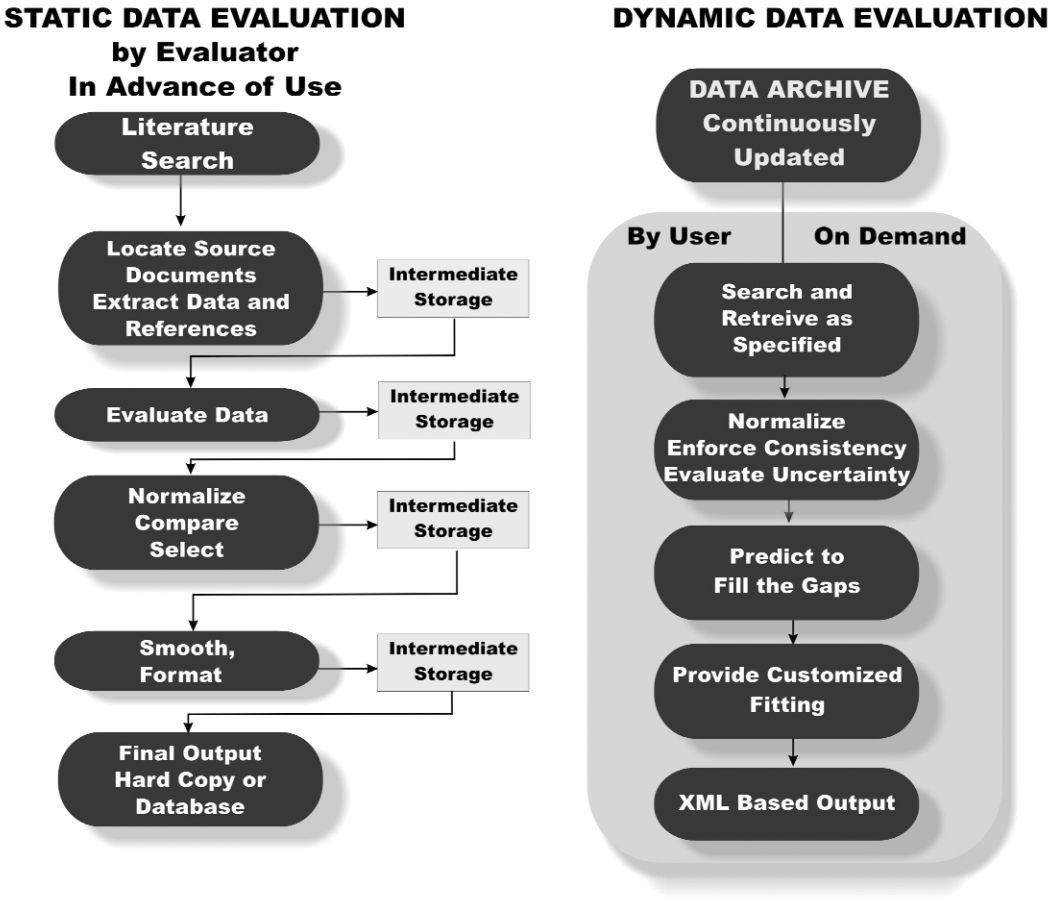

Fig. 4 Functional comparison of static and dynamic data evaluation concepts.

Realization of the dynamic data evaluation concept, based on available experimental data and their uncertainties, provides an opportunity to avoid essentially all principle problems related to the static data evaluation methods currently employed. However, the recommended data generated through implementation of the dynamic data evaluation concept might still have significant "gaps" for particular chemical systems and properties, which have never been studied experimentally. Presently, numerous varied correlation and prediction methods (group contributions, molecular mechanics, semi-empirical quantum, molecular dynamics, ab initio) are available to estimate thermophysical and thermochemical properties. Nevertheless, the "applicability regions" for most of these methods, with regard to the nature of the chemical systems or the properties involved, are not well defined. Moreover, in most cases there are no definitive procedures to assess uncertainties of the predicted property values.

Implementation of the dynamic data evaluation concept together with knowledge-based algorithms to apply prediction and correlation methods which optimize the {recommended data quality/computational time} ratio leads, in principle, to the possibility of generation of the complete set of thermodynamic property data (with the estimated uncertainties) as automatically generated output for any particular chemical entity of interest without regard to its ever having been studied, or for that matter, even synthesized.

Implementation of the dynamic data evaluation concept consists of the solution of a number of major tasks: (1) design and development of a comprehensive database system structure based on the principles of physical chemistry and capable of supporting a large-scale data entry operation for the complete set of thermophysical, thermochemical, and transport properties for chemical systems including pure compounds, binary mixtures, ternary mixtures, and chemical reactions; (2) development of software tools for automation of the data entry process with robust and internally consistent mechanisms for automatic assessments of data uncertainty; (3) design and development of algorithms and software tools to assure quality control at all stages of data entry and analysis; (4) development of algorithms and computer codes to implement the stages of the dynamic data evaluation concept; (5) development of algorithms to implement, target, and apply prediction methods depending on the nature of the chemical system and property, including automatic chemical structure recognition mechanisms; and (6) development of procedures allowing generation of output in a format suitable for application in major commercial simulation engines for chemical-process design.

Comprehensive data archival system

The first three of the six requirements outlined in the previous section for the implementation of the dynamic data evaluation concept are related to the design, maintenance, and population of a comprehensive data storage facility.

Among existing thermodynamic property databases, DIPPR 801 [43], PPDS [44], the Dortmund Data Bank [45], and DETHERM [46] are well established and commonly used in a variety of engineering applications [39]. The DIPPR 801 database contains critically evaluated data for pure compounds, PPDS stores critically evaluated property data for pure compounds and binary interaction coefficients, the Dortmund Data Bank and DETHERM are primarily focused on experimental properties for mixtures although they contain a very significant collection of pure compound properties as well. Even though the databases mentioned above are high-quality data storage facilities, none of them contain information related to the uncertainties [17–20] of the experimental data. Furthermore, these databases do not provide information for thermodynamic property data of chemical reactions.

The TRC SOURCE [11,47] was designed and built as an extensive relational data archival system for experimental thermophysical, thermochemical, and transport properties, which have been reported in the world's scientific literature. It has grown extensively in size and functionality during the past 15 years. The SOURCE now consists of over 1 500 000 numerical property values on more than 17 200 pure compounds, 17 000 binary and ternary mixtures, and 4000 reaction systems. Stored also in the SOURCE are approximately 110 000 records of compound identification; 85 000 records of biblio-

graphic information; and over 70 000 records containing information pertaining to the identity of authors of the original sources. The total number of distinct records currently exceeds 2 600 000. This large data depository system covers data for more than 120 distinct properties. The SOURCE data system contains estimated uncertainties for practically all the numerical data stored, which makes the SOURCE database uniquely positioned to serve as the foundation for implementation of the dynamic data evaluation concept. The design of the SOURCE is based strictly on the principles of chemical thermodynamics—in particular, the Gibbs phase rule.

The SOURCE is managed by the Oracle database management system [39]. 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 serve for the interpretation and display of data. At TRC, the database server, Oracle RDBMS Enterprise Edition, resides on a SUN-280R [39] computer running the Unix operating system, while development tools and other client tools reside on the NIST local network. In this configuration, client software programs run on PCs, and the associated server processes run on the SUN machine using the NIST network and the Oracle network software. Several ways exist to access SOURCE for input and output. These ways include a primary tool for daily data entry and maintenance, batch input and output programs on the server machine, and data reports generated from both the server and the client.

TRC has established a Data Quality Assurance (DQA) program [12] related to both uncertainties characterization and data integrity. As a foundation of the quality control, six steps have been identified: (1) literature collection, (2) information extraction, (3) data entry preparation, (4) data entry insertion, (5) anomaly detection, and (6) database rectification. The GDC software discussed above [31,32] is extensively used for DQA stages 2 through 6.

The TRC Data Entry Facility was established in 2001 to support a mass-scale data entry operation for the SOURCE data system. The operational schema of the facility is illustrated in Fig. 5. TRC operates a large in-house data capture effort staffed chiefly by undergraduate students of chemistry and chemical engineering. The operation is supported by two networks, the SUN Oracle network and the Windows NIST (Boulder) network (Fig. 5). A high-quality scanner is included in the networking system supporting distribution of information obtained by scanning hard-copy documents to each workstation. The TRC Data Entry Facility is a unique facility of its kind worldwide, and operates with a data entry rate of nearly 300 000 data points a year under strict data quality assurance guidelines.

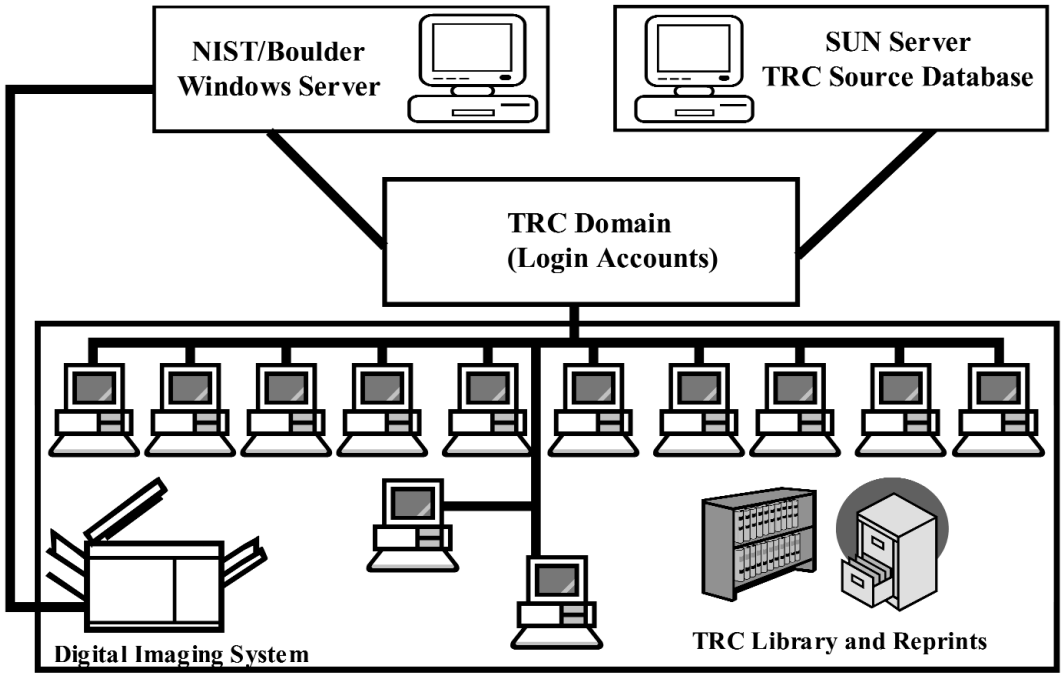

Fig. 5 Organizational and networking structure of the TRC Data Entry Facility.

NEL/NIST(TRC) DataExpert system

DataExpert [39,48] was the first software product developed jointly by TRC and the National Engineering Laboratory (NEL, UK) in the mid-1990s to implement the concept of dynamic data evaluation. The DataExpert software contains three major utilities—VARIABLE, CONSTANT, and SCREEN as well as LOADER2 for the Windows prediction package [49] developed at NEL. The utilities VARIABLE and CONSTANT serve to retrieve data from the NIST/TRC SOURCE data system (temperature-dependent properties) and NIST/TRC Table database [50] (critically evaluated data for critical constants, boiling points, melting points, and thermodynamic data in the ideal-gas state). The retrieved data are combined and preprocessed by the SCREEN utility. The output of the SCREEN utility feeds LOADER2 in order to provide predicted property data, if needed.

Development of DataExpert was a very important first step in implementation of the dynamic data evaluation concept. However, it cannot be operated in a fully autonomous mode, and it requires extensive communications between the software and the user, and has limited enforcement of consistency between properties related mathematically.

New generation of thermodynamic data expert systems (ThermoData Engine)

Recently, a new approach (ThermoData Engine software) for data expert systems implementing dynamic data evaluation concept has been discussed [51–54]. This approach encompasses the development of algorithms and computer codes to implement the stages of the dynamic data evaluation concept as well as incorporation of the prediction methods depending on the nature of the chemical system and property, including automatic chemical structure recognition mechanisms (requirements 4 and 5 for implementation of the dynamic data evaluation concept). The ThermoData Engine software incorporates all major stages of the concept implementation (Fig. 6), including data retrieval, grouping, normalization, sorting, consistency enforcement, fitting, and prediction. The SOURCE data system [11,47] is

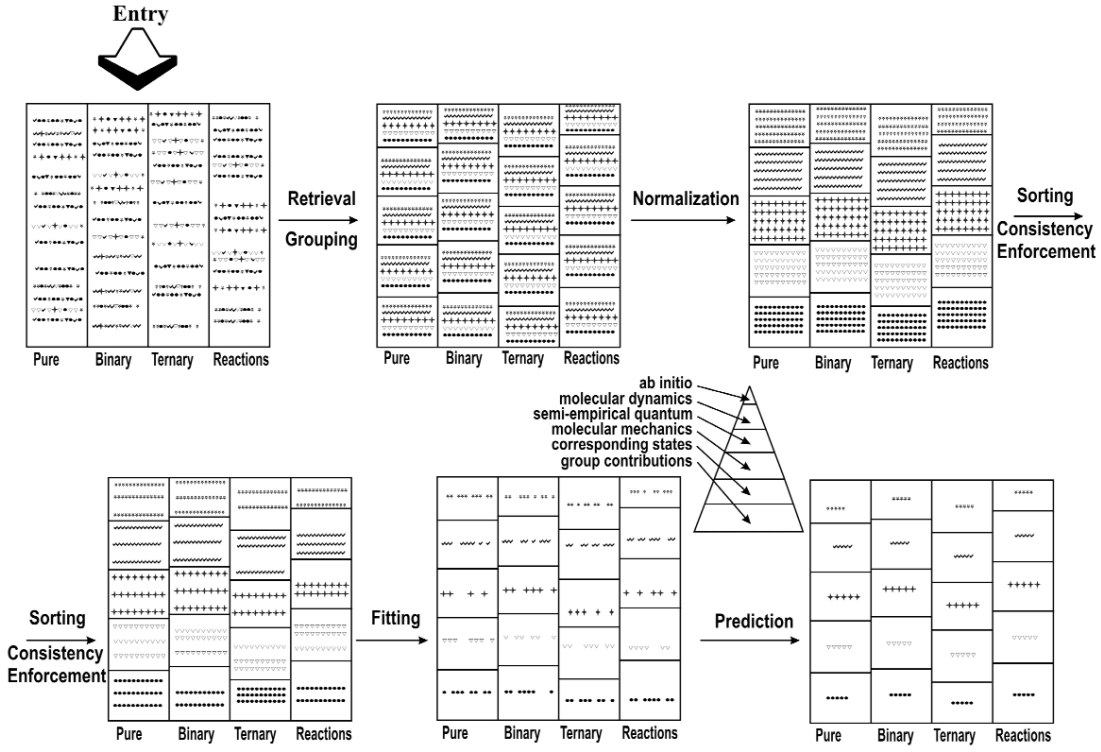

Fig. 6 Major steps of the dynamic data evaluation process.

used in conjunction with ThermoData Engine as a comprehensive storage facility for experimental thermophysical and thermochemical property data. In addition, the NIST/TRC Ideal Gas Database [55,56] is used as a source of thermodynamic property data in the ideal-gas state. The ThermoData Engine software architecture emphasizes enforcement of consistency between related properties (including those obtained from predictions), assumes an imperfect source of original data, provides for flexibility in selection of default data models depending on the particular data scenario, incorporates a large variety of models for secondary fitting, and allows saving of critically evaluated data in the ThermoML format. The latter assures compatibility of the ThermoData Engine software with any engineering application equipped with a ThermoML software "reader". The principle block-schema of the ThermoData Engine "expert" software is shown in Fig. 7. The properties are subdivided into four blocks: phase diagram properties (triple point, critical point, saturated vapor pressure), volumetric properties (critical density, equilibrium density along the saturation line, single-phase density, volumetric coefficients), energy-related properties (enthalpies, heat capacities, speeds of sound), and other properties (transport properties, surface tensions, and refractive indices). ThermoData Engine is tasked with making fully automated and transparent decisions in the process of dynamic data evaluation. Some of those decisions are illustrated in Fig. 8. Thermodynamic consistency conditions enforced by ThermoData Engine include equality of vapor pressures over the solid and liquid phases at triple point, convergence of condensed phase boundary lines at a triple point, convergence of gas and liquid saturation density lines at the critical temperature, infinite first derivatives of saturated density against temperature at the critical temperature, convergence of single-phase densities to saturated densities at equilibrium phase boundaries, etc.

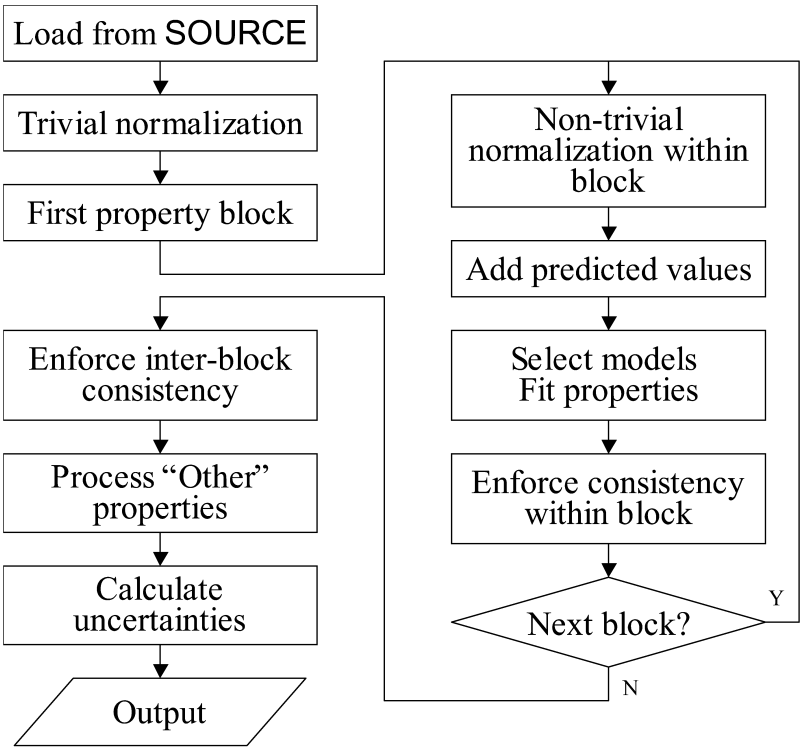

Fig. 7 General processing algorithm in the ThermoData Engine software.

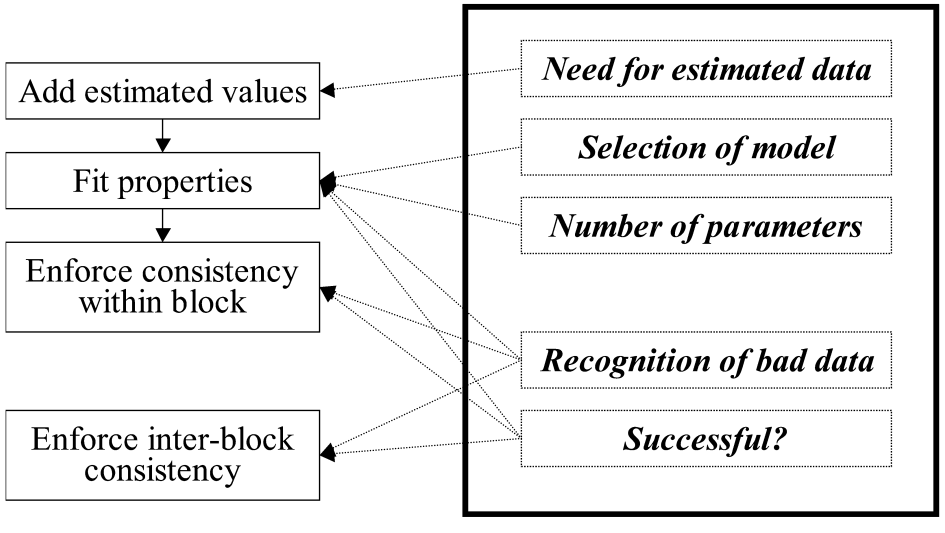

Fig. 8 Examples of data processing decisions made within the ThermoData Engine software.

ThermoData Engine provides comprehensive assessment of the uncertainties for critically evaluated data based on the uncertainties of experimental and predicted data, data "density", propagation of the uncertainty values between related properties, and covariance matrix analyses for coefficients of the fitting equations. To extend the scope of its coverage, ThermoData Engine includes a structure-drawing facility.

Full implementation of the dynamic data evaluation concept requires continuous update of the data storage facility that can be delivered to the computer of a local user through a multi-tier Web-dissemination architecture [57]. ThermoData Engine can communicate with chemical process simulation engines via ThermoML to provide critically evaluated data on demand for analyses of feasibility for conceptual processes and the improvement of existing chemical and biochemical industrial processes (Fig. 9).

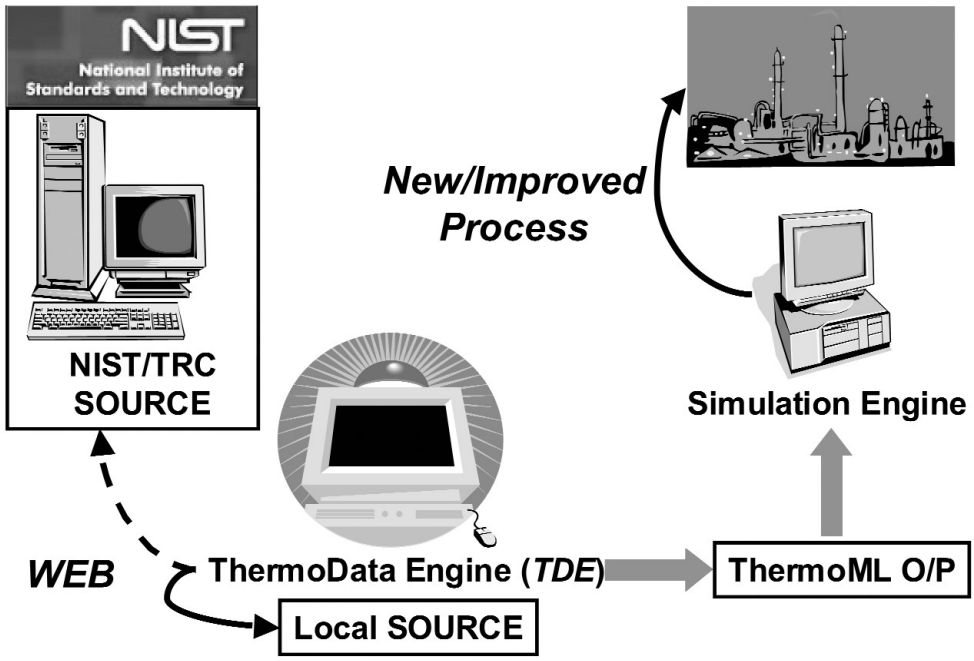

Fig. 9 Global data delivery and communication process linking ThernoData Engine and chemical process simulation applications.

CONCLUSIONS

- Dramatic progress has been made within the last five years in the development of new technologies for thermodynamic data communications including XML-based standards, software support infrastructure, and global data delivery processes.
- An emerging new generation of expert systems has been designed to fully implement all elements of the dynamic data evaluation concept generating critically evaluated thermodynamic data on-demand. These systems include comprehensive data storage facilities and highly efficient "intellectual" software, and are able to characterize evaluated data with reliable measures of their quality (uncertainties), and through application of new XML-based communication technologies, can communicate readily with a great variety of engineering applications.
- New technologies for global communications and expert systems in thermodynamics can make enormous economic and "knowledge discovery" impact.

ACKNOWLEDGMENTS

The author expresses appreciation to Drs. R. D. Chirico, V. V. Diky, X. Yan, and Ms. Q. Dong of NIST/TRC whose efforts led to many of the interesting results discussed in the present paper. Special thanks are given to colleagues from DIPPR; Drs. T. L. Teague (ePlantData), D. L. Embry

(ConocoPhillips), A. K. Dewan (Shell); and IUPAC Task Group members Dr. J. H. Dymond (University of Glasgow, UK), Prof. E. Königsberger (Murdoch University, Australia), Prof. K. N. Marsh (University of Canterbury, New Zealand), Dr. S. E. Stein (NIST/Gaithersburg, Maryland, USA), and Prof. W. A. Wakeham (University of Southampton, UK). In addition, the author thanks Drs. J. W. Magee (NIST/Boulder, Colorado, USA), A. R. H. Goodwin (Schlumberger, Sugar Land, Texas, USA), M. Thijssen (Elsevier, Amsterdam, Netherlands), S. Watanasiri (AspenTech, Cambridge, Massachusetts, USA), M. Satyro (Virtual Materials Group, Calgary, Canada), A. I. Johns (National Engineering Laboratory, Glasgow, UK), and M. Schmidt (Fiz Chemie, Berlin, Germany) for their valuable advice and practical suggestions for establishment of global communication processes for thermodynamic data. The author wishes to acknowledge the late Dr. Randolph Wilhoit of Texas A & M University, a long-time friend, colleague, and associate, who was an inspiration for implementation of the dynamic data evaluation concept.

REFERENCES

1. M. Frenkel, R. D. Chirico, V. V. Diky, K. N. Marsh, J. H. Dymond, W. A. Wakeham. *J. Chem. Eng. Data* **49**, 381 (2004).
2. M. Frenkel, R. D. Chirico, V. V. Diky, Q. Dong, S. Frenkel, P. R. Franchois, D. L. Embry, T. L. Teague, K. N. Marsh, R. C. Wilhoit. *J. Chem. Eng. Data* **48**, 2 (2003).
3. R. C. Wilhoit and K. N. Marsh. *COdataSTAndardThermodynamics. Rules for Preparing COSTAT Message for Transmitting Thermodynamic Data*, Report to CODATA Task Group on Geothermodynamic Data and Chemical Thermodynamic Tables, Paris (1987).
4. <www-i5.informatik.rwth-aachen.de/lehrstuhl/projects/gco/>.
5. <www.fiz-karlsruhe.de/dataexplorer/test/iucosped/dataexplorer.html>.
6. A. K. Dewan, D. L. Embry, T. J. Willman. *Book of Abstracts of the 14th Symposium on Thermophysical Properties*, p. 169, Boulder, CO (2000).
7. R. D. Chirico, M. Frenkel, V. V. Diky, K. N. Marsh, R. C. Wilhoit. *J. Chem. Eng. Data* **48**, 1344 (2003).
8. C. Finkelstein and P. Aiken. *Building Corporate Portals with XML*, McGraw-Hill, New York (1999).
9. P. Murray-Rust and H. S. Rzepa. *J. Chem. Inform. Comp. Sci.* **39**, 938 (1999).
10. <www.matml.org/>.
11. M. Frenkel, Q. Dong, R. C. Wilhoit, K. R. Hall. *Int. J. Thermophys*. **22**, 215 (2001).
12. Q. Dong, X. Yan, R. C. Wilhoit, X. Hong, R. D. Chirico, V. V. Diky, M. Frenkel. *J. Chem. Inform. Comp. Sci*. **42**, 473 (2002).
13. W. B. Whiting. *J. Chem. Eng. Data* **41**, 935 (1996).
14. <www.collectionscanada.ca/iso/tc46sc9/standard/690-1e.htm#1>.
15. <www.collectionscanada.ca/iso/tc46sc9/standard/690-2e.htm>.
16. <www.iupac.org/projects/2000/2000-025-1-800.html>.
17. *Guide to the Expression of Uncertainty in Measurement* (International Organization for Standardization, Geneva, Switzerland, 1993). This *Guide* was prepared by ISO Technical Advisory Group 4 (TAG 4), Working Group 3 (WG 3). ISO/TAG 4 has as its sponsors the BIPM, IEC, IFCC (International Federation of Clinical Chemistry), ISO, IUPAC (International Union of Pure and Applied Chemistry), IUPAP (International Union of Pure and Applied Physics), and OIML. Although the individual members of WG 3 were nominated by the BIPM, IEC, ISO, or OIML, the *Guide* is published by ISO in the name of all seven organizations.
18. *U.S. Guide to the Expression of Uncertainty in Measurement*, ANSI/NCSL Z540-2-1997, NCSL International, Boulder, CO (1997).
19. B. N. Taylor and C. E. Kuyatt. *Guidelines for the Evaluation and Expression of Uncertainty in NIST Measurement Results*, NIST Technical Note 1297, NIST, Gaithersburg, MD (1994).

20. <http://physics.nist.gov/cuu/>.
21. P. Sandhu. *The MathML Handbook*, Charles River Media, Hingham, MA (2003). See also: <www.w3.org/Math/>.
22. I. Mills, T. Cvitas, K. Homann, N. Kallay, K. Kuchitsu. *Quantities, Units and Symbols in Physical Chemistry* (The Green Book), Blackwell Science, Oxford (1993).
23. <www.trc.nist.gov/ThermoML.xsd>.
24. <http://www.iupac.org/projects/2002/2002-055-3-024.html>.
25. *Chem. Int*. **26** (1), 17 (2004).
26. <http://www.iupac.org/standing/cpep.html>.
27. *Chem. Int*. **26** (4), 26 (2004).
28. <www.iupac.org/namespaces/ThermoML/>.
29. <www.rcsb.org/pdb/>.
30. <www.ccdc.cam.ac.uk/products/csd/>.
31. <www.trc.nist.gov/GDC.html>.
32. V. V. Diky, R. D. Chirico, R. C. Wilhoit, Q. Dong, M. Frenkel. *J. Chem. Inform. Comp. Sci*. **43**, 15 (2003).
33. M. Frenkel, R. D. Chirico, V. V. Diky, Q. Dong, S. Frenkel, P. R. Franchois, D. L. Embry, T. L. Teague, K. N. Marsh, R. C. Wilhoit. *Book of Abstracts of the 15th Symposium on Thermophysical Properties*, p. 124, Boulder, CO (2003).
34. <www.trc.nist.gov/ThermoML.html>.
35. <www.boulder.nist.gov/div838/trc/journals/jced/2004v49/i02/jced2004v49i02.html>.
36. K. N. Marsh. *J. Chem. Eng. Data* **48**, 1 (2003).
37. <www.boulder.nist.gov/div838/trc/journals/jct/2004v36/i07/jct2004v36i07.html>.
38. *J. Chem. Thermodyn.* **36**, iv (Editorial) (2004).
39. Names of commercial products and/or commercial entities are provided for complete scientific description and as a service to the reader of this publication. Such identification neither constitutes nor implies endorsement of such products or companies by NIST or by the U.S. Government. Other products or services may be found to be just as good.
40. <www.iactweb.org/Projects/projects.htm>.
41. R. C. Wilhoit and K. N. Marsh. *J. Chem. Inform. Comp. Sci.* **29**, 17 (1989).
42. M. Frenkel. In: *Report on Forum 2000: Fluid Properties for New Technologies Connecting Virtual Design with Physical Reality*, J. C. Rainwater, D. G. Friend, H. J. M. Hanley, A. H. Harvey, C. D. Holcomb, A. Laesecke, J. W. Magee, C. Muzny (Eds.), NIST Special Publication 975, p. 83, Gaithersburg, MD (2001).
43. <http://dippr.byu.edu/>.
44. <www.ppds.co.uk/products/ppds.asp>.
45. <www.ddbst.de/new/Default.htm>.
46. <www.dechema.de/detherm-lang-en.html>.
47. X. Yan, Q. Dong, M. Frenkel, K. R. Hall. *Int. J. Thermophys*. **22**, 227 (2001).
48. <www.ppds.co.uk/products/dataexpert.asp>.
49. <www.ppds.co.uk/products/loader.asp>.
50. <www.nist.gov/srd/nist85.htm>.
51. V. V. Diky, R. D. Chirico, X. Yan, R. C. Wilhoit, M. Frenkel. *Book of Abstracts of the 225th National Meeting of the American Chemical Society*, Vol. 1, abstract CINF15, New Orleans (2003).
52. V. V. Diky, R. D. Chirico, X. Yan, R. C. Wilhoit, M. Frenkel. *Book of Abstracts of the 15th Symposium on Thermophysical Properties*, p. 91, Boulder, CO (2003).
53. M. Frenkel. *Book of Abstracts of the 10th International Conference on Properties and Phase Equilibria for Product and Process Design*, p. 63, Snowbird, Utah (2004).

54. M. Frenkel, R. D. Chirico, V. V. Diky, X. Yan, Q. Dong. *Book of Abstracts of the 59th Calorimetry Conference*, p. 96, Santa Fe, NM (2004).
55. <www.nist.gov/srd/nist88.htm>.
56. M. Frenkel, G. J. Kabo, K. N. Marsh, G. N. Roganov, R. C. Wilhoit. *Thermodynamics of Organic Compounds in the Gas State*, Vols. 1, 2, TRC, College Station, TX (1994).
57. J. Garmany and D. K. Burleson. *Oracle Application Server 10g. Administration Handbook*, McGraw-Hill, New York (2004).