Reviews

ThermoML—An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 1. Experimental Data

Michael Frenkel,* Robert D. Chirico, Vladimir V. Diky, and Qian Dong

Thermodynamics Research Center (TRC), Physical and Chemical Properties Division, National Institute of Standards and Technology, 325 Broadway, Boulder, Colorado 80305-3328

Svetlana Frenkel and Paul R. Franchois

Information Technology Laboratory, National Institute of Standards and Technology, 325 Broadway, Boulder, Colorado 80305-3337

Dale L. Embry

ConocoPhillips, 850-16 Street, P.O. Box 1267, Ponca City, Oklahoma 74602-1267

Thomas L. Teague

ePlantData, Inc., 9955 South Post Oak Road, Suite 300, Houston, Texas 77096

Kenneth N. Marsh

Department of Chemical and Process Engineering, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

Randolph C. Wilhoit

Texas Experimental Engineering Station, Texas A&M University System, College Station, Texas 77843

ThermoML is an XML-based approach for storage and exchange of experimental and critically evaluated thermophysical and thermochemical property data. The basic principles, scope, and description of all structural elements of ThermoML are discussed. 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. Although the focus of ThermoML is properties determined by direct experimental measurement, ThermoML does cover key derived property data such as azeotropic properties, Henry's Law constants, virial coefficients (for pure compounds and mixtures), activities and activity coefficients, fugacities and fugacity coefficients, and standard properties derived from high-precision adiabatic heat-capacity calorimetry. The role of ThermoML in global data submission and dissemination is discussed with particular emphasis on the new cooperation in data processing between the *Journal of Chemical and Engineering Data* and the Thermodynamics Research Center (TRC) at the National Institute of Standards and Technology. The text of several data files illustrating the ThermoML format for pure compounds, mixtures, and chemical reactions, as well as the complete ThermoML schema text, is provided as Supporting Information. Some important issues related to characterization of thermodynamic data are beyond the scope of this paper (uncertainty specification) or are considered in generic terms only (critically evaluated data). These issues will be considered in subsequent papers in this series.

Background

Efforts to develop a standard for thermophysical and thermochemical property data exchange 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

* Corresponding author. Phone: (303)-497-3952. Fax: (303)-497-5044. E-mail: frenkel@boulder.nist.gov.

10.1021/je025645o CCC: $25.00
Published on Web 12/17/2002

COSTAT (COdata STAndard Thermodynamics).[1] This prototype was discussed extensively among numerous institutions worldwide through the auspices of CODATA. This effort played a very 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.

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[2] 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.

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 experimental data system (SOURCE[3]), which currently includes more than 120 properties for pure compounds, mixtures, and chemical reactions. SOURCE has been developed at TRC as a major element of a computerized expert system to implement the concept of the "dynamic data evaluation" necessary to produce critically evaluated data reports automatically "to order".[4−6] ThermoML is an application of the XML (Extensible Markup Language) technology.[7]

The principles and approach used in development of ThermoML were discussed extensively with the members of the DIPPR (Design Institute for Physical Properties of the American Institute of Chemical Engineers) 991 Project. The DIPPR 991 project team has been defining a broadly scoped physical properties data XML schema (ppdXML), to be completed in early 2003. ppdXML, which will be described in a subsequent joint paper, additionally includes support for property calculation methods and parameters, tabulated and calculated properties, and simulation stream properties. As a result of this cooperation, ThermoML terminology has been integrated into ppdXML for describing experimental data and bibliographic information. Software that provides direct translation of ThermoML data files into ppdXML data files is currently being planned for development by DIPPR.

Basic Principles

Schema Structure. 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. Metadata and numerical data records are grouped into "nested blocks" of information corresponding to data sets. The metadata records precede numerical data information, providing a robust foundation for generating "header" records for any relational database where ThermoML-formatted files could be incorporated. The structural features of the ThermoML metadata records ensure unambiguous interpretation of numerical data as well as 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 ensuring the highest quality in data, and it would provide users with an indication of inconsistencies in thermodynamic data before the data are deposited into a data-storage facility.[8] 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.[9] Definitions of uncertainties and their descriptions are complex and will be discussed in the next paper in this series.

Tagging. Commonly accepted IUPAC-based terminology is used as a 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.

Modularity. ThermoML is designed to take advantage of the modular nature of XML schemas. Structurally, it can be expanded easily into areas that are currently beyond its scope.

Units. By design, there is only one unit selected for each property covered by ThermoML. These units are SI-based; however, 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.

Data Representation. Various methods of numerical data representation commonly used in 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.

Scope

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. Polymers and ionic systems including salt and acid solutions are not fully supported by ThermoML at present. The focus of ThermoML is on properties determined by direct measurements; therefore, most derived properties are not covered. Currently, data obtained from estimations and correlations (group contributions, corresponding states, etc.), as well as theoretical calculations, are beyond the scope of ThermoML. ThermoML does cover key derived properties such as azeotropic properties, Henry's Law constants, virial coefficients, activities and activity coefficients, fugacities and fugacity coefficients, and standard properties derived from high-precision adiabatic heat-capacity calorimetry $\{S(T), H(T) - H(0)$, etc.$\}$. Common properties that do not have unambiguous thermodynamic definitions, such as decomposition temperature, flammability limits, and octane numbers, are not included.

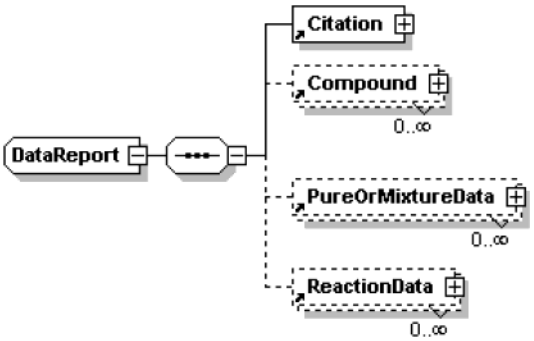

Figure 1. Major components of ThermoML.

The list of all properties within the scope of ThermoML is provided with the complete schema description. Although the current focus of ThermoML is experimentally determined property data, provisions will be added for storage of critically evaluated and calculated values as well.

Description of the ThermoML Schema

ThermoML consists of four major blocks, as shown in Figure 1. All schema figures and text of ThermoML were created with the software package XML SPY.[10] (We use trade names to specify the procedure adequately and do not imply endorsement by the National Institute of Standards and Technology. Similar products by other manufacturers may work as well or better.)

(1) *Citation* (description of the source of the data).

(2) *Compound* (characterization of the chemical system). The description for every compound is linked to a description of the sample used in the measurements with indication of its initial purity, purification method used, final purity, and the method used to determine it.

(3) *PureOrMixtureData* (metadata and numerical data for a pure compound or multicomponent mixture).

(4) *ReactionData* (metadata and numerical data for a chemical reaction with a thermodynamic state change or in a state of chemical equilibrium).

1. "Citation" Block. The schema for the "*Citation*" block is shown in Figure 2. The major components are described below. The names or "tags" include special characters related to the type of information to be stored. Names beginning with "e" designate "enumeration" elements (that is, values of which are selected from a predefined list), those with "s" designate "string" elements, those with "n" specify "numerical" elements (integer or floating), those with "yr" designate elements characterizing the year, those with "date" specify date elements, and those with "url" indicate Web address elements. The elements identified by dotted boxes in the figure are optional, and those in solid-lined boxes are mandatory. Complex elements illustrated without their internal structure are identified by "+". Complex elements with internal structure displayed are identified with "−". Multiple elements of the same type are identified by lower and upper limits listed below the relevant boxes in the schema. Within ThermoML, the limits used are "0 ... ∞" for optional elements and "1 ... ∞" for mandatory elements.

eType [enumeration] indicates the type of source document (book, journal, report, patent, thesis, conference proceedings, archived document, personal correspondence, published translation, or unspecified).

eSourceType [enumeration] provides information about the nature of the source of information (original journal article, Chemical Abstracts, or other).

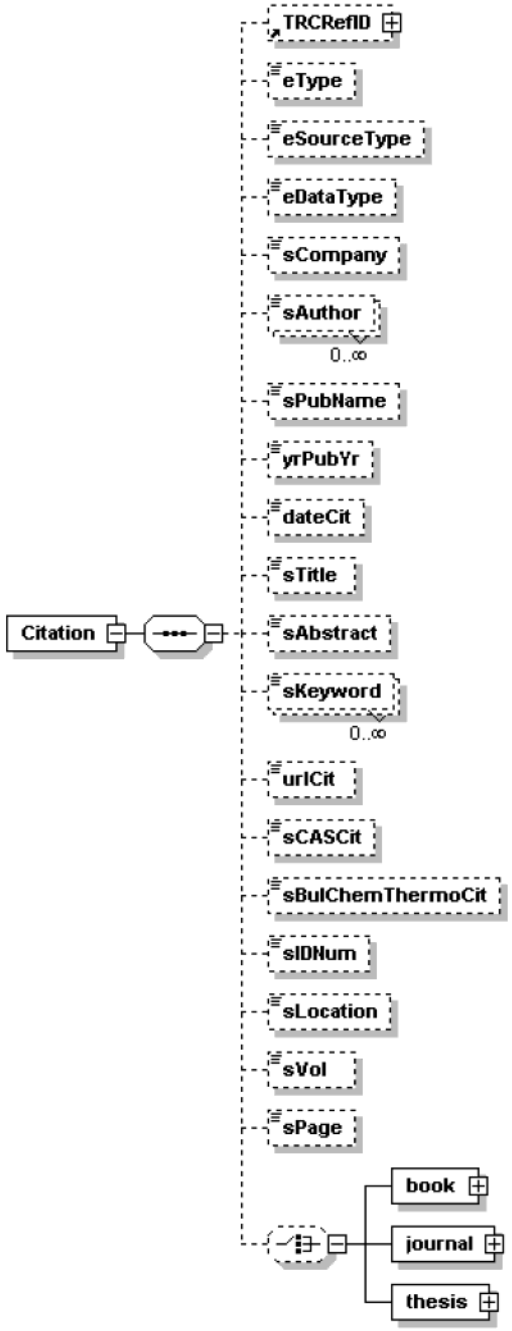

Figure 2. General schema of the "*Citation*" block.

eDataType [enumeration] describes the extent to which the information from the original document has been collected into the ThermoML file (reference only, some numerical data, data for pure compounds only, combination of data for pure compounds and mixtures, data for mixtures only, or all data).

sCompany [string] characterizes the origin of the document, such as a company name, institution, or conference.

The other elements of the "*Citation*" block are as follows: **sAuthor** [string], the author name; **sPubName** [string], the name of the publication where the citation was published; **yrPubYr** [year], the year of publication; **dateCit** [date], the date of publication; **sTitle** [string], the title of the cited document; **sAbstract** [string], the abstract for the document; **sKeyword** [string], a keyword for the document; **urlCit** [url], a url for the citation; **sCASCit** [string], the Chemical Abstract Service citation; **sBulChemThermoCit** [string], a Bulletin of Chemical Thermody-

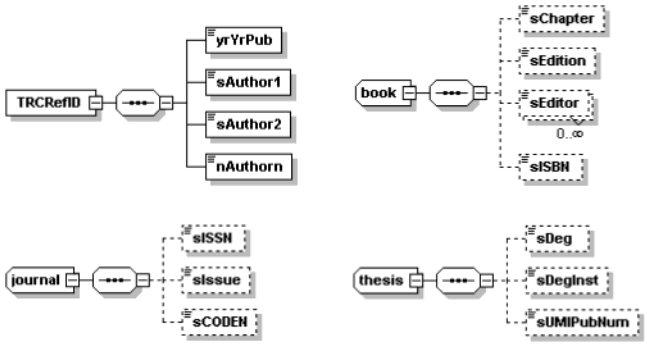

Figure 3. Structures of the **TRCRefID**, **book**, **journal**, and **thesis** elements of the "*Citation*" block.

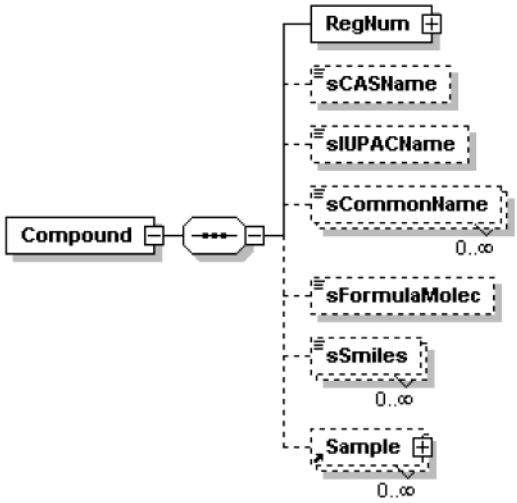

Figure 4. General schema of the "*Compound*" block.

namics citation; **sIDNum** [string], a local or global reference identifier; **sLocation** [string], the place of publication; **sVol** [string], the volume number; and **sPage** [string], the page range for the citation.

TRCRefID, **book**, **journal**, and **thesis** are complex elements within the "*Citation*" block. Their structures are illustrated in Figure 3. **TRCRefID**, the TRC reference identifier, consists of **yrYrPub** [year], the year of publication; **sAuthor1** [string], the first three characters of the first author's last name; **sAuthor2** [string], the first three characters of the second author's last name; and **nAuthorn** [numerical, integer], a numerical value to ensure the uniqueness of every distinct **TRCRefID**.

Books, journals, and theses as source documents are characterized with additional tags. For **book**: **sChapter** [string] contains the chapter identifier, **sEdition** [string] the edition identifier, **sEditor** [string] the editor's name, and **sISBN** [string] the International Standard Book Number. For **journal**, the following items are specified: **sISSN** [string] specifies the International Standard Subscription Number, **sIssue** [string] the issue identifier, and **sCODEN** [string] the CODEN identification of the journal. For **thesis**: **sDeg** [string], the academic degree designation (such as M.S., Ph.D., etc.), **sDegInst** [string], the academic degree granting institution, and **sUMIPubNum** [string], the University Microfilm International Publication Number are designated.

2. "Compound" Block. The schema for the "*Compound*" block is represented in Figure 4. Compounds can be characterized with a variety of chemical names. These are **sCASName** [string], the Chemical Abstract Service name, **sIUPACName** [string], the name specified by IUPAC, and

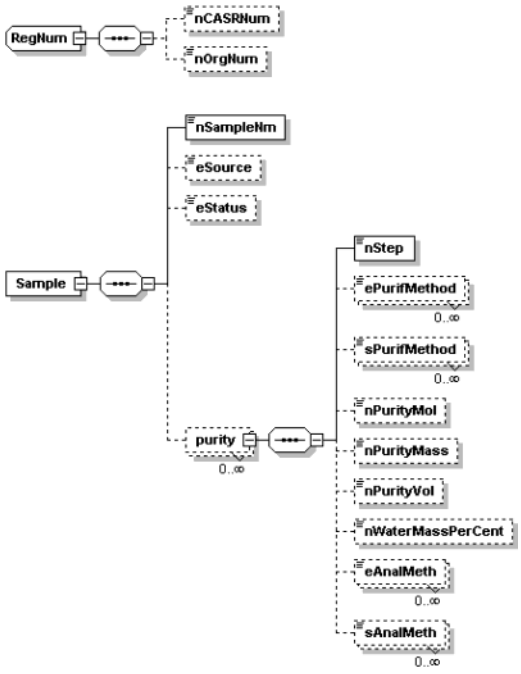

Figure 5. Structures of the **RegNum** and **Sample** elements of the "*Compound*" block.

sCommonName [string], which allows any other name. The other elements in the "*Compound*" block are **sFormulaMolec** [string], the elemental molecular formula, and **sSmiles** [string], the SMILES notation which describes the chemical formula.

RegNum (Compound Registry Number) and **Sample**, a description of the sample used for experimental measurements, are complex elements in the "*Compound*" block. Their structures are shown in Figure 5. **RegNum** is represented by the true Chemical Abstract Service Registry Number (**nCASNum**) or by an identification number assigned by a user organization (**nOrgNum**).

The **Sample** element consists of four subelements, as shown in Figure 5. These are **nSampleNm** [numerical, integer], used to distinguish different samples of the same compound, **eSource** [enumeration], used to indicate the original source of the sample before purification (commercial, synthesized by author, synthesized by someone else, isolated from a natural product, standard reference material, not stated in the citation), **eStatus** [enumeration], used to indicate the status of the sample description (unknown, not described, described in another source, no sample used), and the complex element **Purity**, used to provide information related to the purity of the sample.

The complex element **Purity** consists of **nStep** [numerical, integer], a sequential number corresponding to a purification stage, **ePurifMethod** [enumeration], the purification method applied at the specified step (impurity adsorption, vacuum degasification, chemical reagent treatment, crystallization from melt, crystallization from solution, liquid chromatography, drying with chemical reagent, drying in desiccator, drying by oven heating, drying by vacuum heating, fractional crystallization, fractional distillation, molecular sieve treatment, preparative gas chromatography, sublimation, steam distillation, solvent extraction, salting out of solution, zone refining, not specified), or **sPurifMethod** [string], the purification method, if it is not listed in the enumeration values for **ePurifMethod**. The complex element **Purity** also includes **eAnalMethod** [enumeration], the analytical method used to determine the purity after a purification stage (chemical analysis,

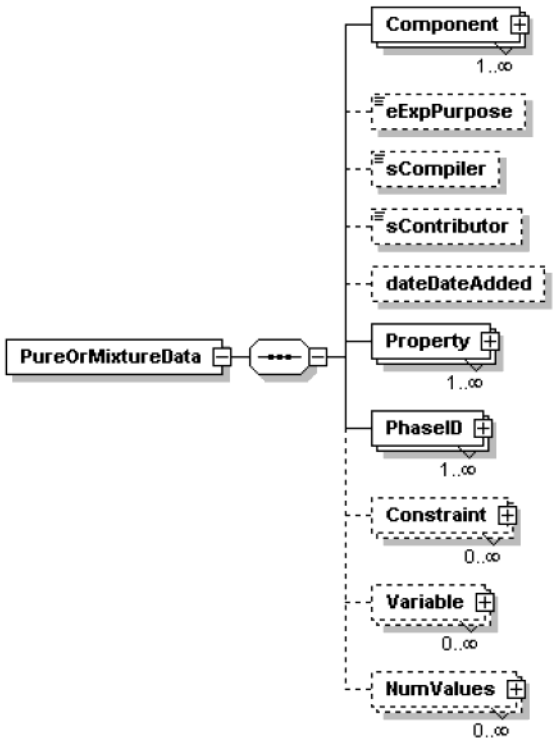

Figure 6. General schema of the "*PureOrMixtureData*" block.

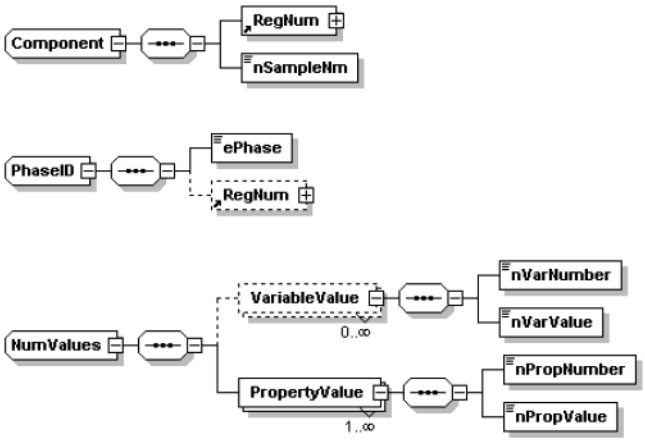

Figure 7. Structures of the **Component, PhaseID,** and **NumValues** elements of the "*PureOrMixtureData*" block.

difference of bubble point and dew point, density, differential scanning calorimetery, mass spectrometry, spectroscopy, estimation, gas chromatography, thermal analysis using a calorimeter, acid−base titration, mass loss on drying, unspecified), or **sAnalMethod** [string], the analytical method used, if it is not listed as an enumeration value for **eAnalMethod**; **nPurityMol** [numerical, floating], the mole fraction purity; **nPurityMass** [numerical, floating], the mass fraction purity; **nPurityVol** [numerical, floating], the volume fraction purity; and **nWaterMassPerCent** [numerical, floating], the mass fraction of water.

3. "PureOrMixtureData" Block. The schema for the "*PureOrMixtureData*" block is shown in Figure 6. This block contains information about the source of the ThermoML file, identifies the experimental purpose, specifies metadata and numerical data, and specifies the compound (or mixture) to which the data are related.

The source of the ThermoML file is recorded through the following elements: **sCompiler** [string], the name of the person who compiled the data contained in the ThermoML file; **sContributor** [string], an identifier for the project, institution, or general source of the ThermoML file; and **dateDateAdded** [date], the date that the ThermoML was created.

The compound or mixture associated with the property data is identified by the element **Component** (Figure 7), consisting of **RegNum** [complex] and a sample number, **nSampleNum** [numerical, integer]. The **RegNum** structure was described earlier (See Figure 5). The experimental-purpose element, **eExpPurpose** [enumeration], contains the following options: principal objective of the work, byproduct of some other objective, and obtained in conjunction with a synthesis to establish identity or purity.

Metadata are described by the four complex elements **Property**, **PhaseID**, **Constraint**, and **Variable. Property** (Figure 8) is characterized by **PropertyMethodID** [complex], which identifies the property and experimental method used, **PropPhaseID** [complex] indicates the phase associated with the property, **ePresentation** [enumeration] indicates the mathematical form used to report the data {i.e., absolute or some type of specified relative value(s)}, **eRefStateType** [enumeration] describes the thermodynamic reference state (if required), **nRefTemp** [numerical, floating] lists the value of a reference temperature, **nRefPressure** [numerical, floating] lists the value of a reference pressure, **RefPhaseID** [complex] indicates the reference phase for a particular data set, **Solvent** [complex] identifies the solvent used, **eStandardState** [enumeration] indicates the thermodynamic standard state (if required by the property definition), and **nPropNumber** [numerical, integer] is a sequential property number for the case in which multiple properties are listed as a function of the same variable values. The provision for **nPropNumber** is convenient for storage of tie-line phase equilibria data.

The element **PropertyMethodID** includes **PropertyGroup** [complex] and **RegNum** [complex], as shown in Figure 8. **RegNum** [complex] has the same structure as described earlier, but it should be used for mixtures only if the property definition involves a specific component, such as mole fraction of a particular compound. The **PropertyGroup** element is represented by 10 property groups: **Criticals**, **VaporPBoilingTAzeotropTandP** (an abbreviation of "*Vapor pressure, Boiling temperature, and Azeotropic temperature and pressure*"), and so forth, as listed in the upper right of Figure 8. Thermophysical properties are divided into these 10 groups to simplify the property-selection process for the ThermoML user. Each group is characterized by the **ePropertyName** [enumeration] and **eMethodName** [enumeration]. These are shown in expanded form for the **Criticals** group in Figure 8. If the option "Other" is used as a value for **eMethodName**, **sMethodName** [string] should be used to identify the method. This option could be used to identify data obtained from critical evaluations or calculations. Descriptions of critically evaluated data will be addressed explicitly in the third paper of this series.

The list of options for **ePropertyName** and **eMethodName** for each **PropertyGroup** is provided below together with the units for each property.

Criticals *[complex]*

ePropertyName [enumeration] (critical temperature, K; critical pressure, kPa; critical density, $kg \cdot m^{-3}$; critical molar volume, $m^3 \cdot mol^{-1}$; critical specific volume, $m^3 \cdot kg^{-1}$; critical compressibility; lower consolute temperature, K; upper consolute temperature, K).

eMethodName [enumeration] (visual observation in an unstirred cell, visual observation in a stirred cell, DSC/

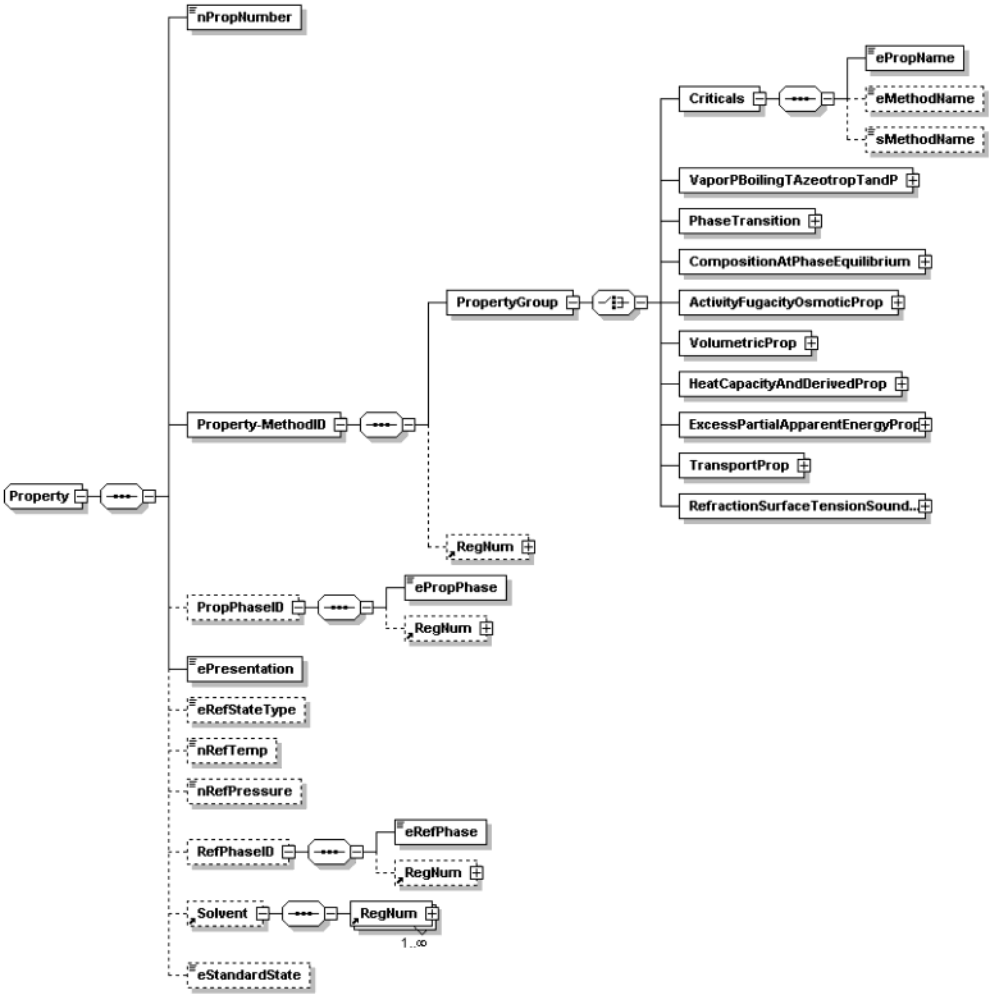

Figure 8. Structure of the **Property** element of the "*PureOrMixtureData*" block.

DTA, derived from *PVT* data, extrapolated vapor pressure, rectilinear diameter, appearance of two phases, disappearance of two phases, other).

VaporPBoilingTAzeotropTandP *[complex]*

ePropertyName [enumeration] (vapor or sublimation pressure, kPa; normal boiling temperature, K; boiling temperature at pressure *P*, K; azeotropic pressure, kPa; azeotropic temperature, K).

eMethodName [enumeration] (manometric method, closed cell—static method, diaphragm manometer, inclined piston gauge, isochoric PVT apparatus, isoteniscope, Knudsen effusion method, distillation, ebulliometric method—recirculating still, twin ebulliometer, transpiration method, rate of evaporation, azeotropic temperature or pressure determination when X = Y, azeotropic temperature or pressure determination by temperature or pressure extreme).

PhaseTransition *[complex]*

ePropertyName [enumeration] (triple point temperature, K; triple point pressure, kPa; normal melting temperature, K; enthalpy of transition or fusion, kJ·mol^{-1}; cryoscopic constant, K^{-1}; enthalpy of vaporization or sublimation, kJ·mol^{-1}; quadruple (quintuple) point temperature, K; quadruple (quintuple) point pressure, kPa; solid—liquid equilibrium temperature, K; liquid—liquid equilibrium temperature, K; eutectic temperature, K).

eMethodName [enumeration] (visual observation, heating or cooling curves, DSC or DTA, adiabatic calorimetry, large-sample thermal analysis, drop calorimetry, drop ice or diphenyl ether calorimetry, obtained from cryoscopic constant, derived from phase diagram analysis, static calorimetry, flow calorimetry, derived by the Second Law, depression of a freezing point of a dilute solution, other).

CompositionAtPhaseEquilibrium *[complex]*

ePropertyName [enumeration] (azeotropic composition—mole fraction; azeotropic composition—mass fraction; eutectic composition—mole fraction; eutectic composition—mass fraction; eutectic composition—volume fraction; lower consolute composition—volume fraction; lower consolute composition—mole fraction; lower consolute composition—mass fraction; mass per volume of solution, kg·m^{-3}; mass ratio to solvent; molality, mol·kg^{-1}; molarity, mol·dm^{-3}; mole fraction; mole fraction in LLG critical state; mole ratio to solvent; moles per mass of solution, mol·kg^{-1}; upper consolute composition—volume fraction; upper consolute composition—mole fraction; upper consolute composition—mass fraction; volume fraction; volume ratio to solvent; mass fraction; mass fraction in LLG critical state; Henry's Law constant for mole fraction, kPa; Henry's Law constant (molality), kPa·kg·mol^{-1}; Henry's Law constant (molarity), kPa·L·mol^{-1}; Bunsen coefficient; Oswald coefficient; partial pressure, kPa).

eMethodName [enumeration] (azeotropic composition determination when X = Y, azeotropic composition determination by temperature of pressure extreme, chromatography, spectrophotometry, determined by refractive index and/or density, calculated by Gibbs—Duhem equation, titration method, static method, dynamic method, phase equilibration, derived from phase diagram analysis, ap-

pearance of two phases, disappearance of two phases, photoacoustic method, other).

ActivityFugacityOsmoticProp *[complex]*

ePropertyName [enumeration] (activity; activity coefficient; fugacity, kPa; fugacity coefficient; osmotic pressure, kPa; osmotic coefficient).

eMethodName [enumeration] (chromatography, spectroscopy, mass-spectrometry, NMR-spectrometry, static method, isopiestic method, other).

VolumetricProp *[complex]*

ePropertyName [enumeration] (specific density, $kg \cdot m^{-3}$; specific volume, $m^3 \cdot kg^{-1}$; molar density, $mol \cdot m^{-3}$; molar volume, $m^3 \cdot mol^{-1}$; second virial coefficient, $m^3 \cdot mol^{-1}$; second acoustic virial coefficient, $m^3 \cdot mol^{-1}$; third virial coefficient, $m^6 \cdot mol^{-2}$; third acoustic virial coefficient, $m^6 \cdot mol^{-2}$; third interaction virial coefficient C_{112}, $m^6 \cdot mol^{-2}$; third interaction virial coefficient C_{122}, $m^6 \cdot mol^{-2}$; excess virial coefficient, $m^3 \cdot mol^{-1}$; interaction virial coefficient, $m^3 \cdot mol^{-1}$; excess volume, $m^3 \cdot mol^{-1}$; partial molar volume, $m^3 \cdot mol^{-1}$; relative partial molar volume, $m^3 \cdot mol^{-1}$; apparent molar volume, $m^3 \cdot mol^{-1}$; adiabatic compressibility, kPa^{-1}; isothermal compressibility, kPa^{-1}; coefficient of expansion, K^{-1}; compressibility factor; thermal pressure coefficient, $kPa \cdot K^{-1}$).

eMethodName [enumeration] (pycnometric method, buoyancy method, vibrating tube method, isochoric *PVT* measurement, other *PVT* measurement, Burnett expansion technique, constant-volume piezometry, direct dilatometry, derived analytically, derived graphically, calculated with densities of this investigation, calculated with a solvent density reported elsewhere, other).

HeatCapacityAndDerivedProp *[complex]*

ePropertyName [enumeration] (heat capacity at constant pressure, $J \cdot K^{-1} \cdot mol^{-1}$; heat capacity at vapor saturation pressure, $J \cdot K^{-1} \cdot mol^{-1}$; heat capacity at constant volume, $J \cdot K^{-1} \cdot mol^{-1}$; heat capacity at constant pressure per unit mass, $J \cdot K^{-1} \cdot kg^{-1}$; heat capacity at constant pressure per unit volume, $J \cdot K^{-1} \cdot m^{-3}$; heat capacity ratio C_p/C_v, standard entropy, $S(T) - S(0)$, $J \cdot K^{-1} \cdot mol^{-1}$; standard enthalpy, $H(T) - H(0)$, $kJ \cdot mol^{-1}$; enthalpy function, $\{H(T) - H(0)\}/T$, $J \cdot K^{-1} \cdot mol^{-1}$; Gibbs energy function, $\{G(T) - H(0)\}/T$, $J \cdot K^{-1} \cdot mol^{-1}$; Gibbs energy, $G(T) - H(0)$, $kJ \cdot mol^{-1}$; Helmholtz energy, $A(T) - E(0)$, $kJ \cdot mol^{-1}$; internal energy, $E(T) - E(0)$, $kJ \cdot mol^{-1}$; Joule–Thompson coefficient, $K \cdot kPa^{-1}$; pressure coefficient of enthalpy, $J \cdot mol^{-1} \cdot kPa^{-1}$).

eMethodName [enumeration] (vacuum adiabatic calorimetry, small sample (less than 1 g) adiabatic calorimetry, flow calorimetry, large sample (1 g) DSC, small sample (50 mg) DSC, drop calorimetry, drop ice or diphenyl ether calorimetry, open cup calorimetry, closed cup calorimetry, differential flow calorimetry, extra sensitive DSC, twin closed cell calorimetry, derived from speed of sound, derived from equation of state, expansion technique, other).

ExcessPartialApparentEnergyProp *[complex]*

ePropertyName [enumeration] (apparent enthalpy, $kJ \cdot mol^{-1}$; apparent entropy, $J \cdot mol^{-1} \cdot K^{-1}$; apparent Gibbs energy, $kJ \cdot mol^{-1}$; apparent molar heat capacity, $J \cdot mol^{-1} \cdot K^{-1}$; enthalpy of mixing with binary solvent, $kJ \cdot mol^{-1}$; excess enthalpy [enthalpy of mixing], $kJ \cdot mol^{-1}$; excess entropy, $J \cdot mol^{-1} \cdot K^{-1}$; excess Gibbs energy, $kJ \cdot mol^{-1}$; excess heat capacity, $J \cdot mol^{-1} \cdot K^{-1}$; partial molar enthalpy, $J \cdot mol^{-1}$; partial molar entropy, $J \cdot mol^{-1} \cdot K^{-1}$; partial molar Gibbs energy, $kJ \cdot mol^{-1}$; partial molar heat capacity, $J \cdot mol^{-1} \cdot K^{-1}$; relative partial molar enthalpy, $kJ \cdot mol^{-1}$; relative partial entropy, $J \cdot mol^{-1} \cdot K^{-1}$; relative partial molar Gibbs energy, $kJ \cdot mol^{-1}$; relative partial molar heat capacity, $J \cdot mol^{-1} \cdot K^{-1}$; standard state enthalpy, $kJ \cdot mol^{-1}$; standard state entropy, $J \cdot mol^{-1} \cdot K^{-1}$; standard state Gibbs energy, $kJ \cdot mol^{-1}$; standard state heat capacity, $J \cdot mol^{-1} \cdot K^{-1}$).

eMethodName [enumeration] (flow calorimetry, Calvet calorimetry, other).

TransportProp *[complex]*

ePropertyName [enumeration] (viscosity, $Pa \cdot s$; kinematic viscosity, $m^2 \cdot s^{-1}$; fluidity, $Pa^{-1} \cdot s^{-1}$; thermal conductivity, $W \cdot m^{-1} \cdot K^{-1}$; thermal diffusivity, $m^2 \cdot s^{-1}$; binary diffusion coefficient, 10^{-9} $m^2 \cdot s^{-1}$; self-diffusion coefficient, 10^{-9} $m^2 \cdot s^{-1}$; tracer diffusion coefficient, 10^{-9} $m^2 \cdot s^{-1}$).

eMethodName [enumeration] (capillary tube [Oswald, Ubbelohde] method, cone and plate viscometry, concentric cylinders viscometry, falling or rolling sphere viscometry, oscillating disk viscometry, vibrating wire viscometry, parallel plate method, coaxial cylinder method, hot wire method, optical interferometry, dispersion, diaphragm cell, open capillary, closed capillary, Taylor dispersion method, NMR spin–echo technique, other).

RefractionSurfaceTensionSoundSpeed *[complex]*

ePropertyName [enumeration] (refractive index [Na D-line]; refractive index [other wavelength]; surface tension liquid–gas, $N \cdot m^{-1}$; interfacial tension, $N \cdot m^{-1}$; speed of sound, $m \cdot s^{-1}$).

eMethodName [enumeration] (standard Abbe refractometry, precision Abbe refractometry, dipping refractometry [monochromatic], interferometer, capillary rise, drop weight, drop volume, maximal bubble pressure, pendant drop shape, ring tensimeter, linear variable-path acoustic interferometer, sing-around technique in a fixed-path interferometer, annular interferometer, pulse-echo method, spherical resonator, light diffraction method, other).

The element **ePresentation** [enumeration] specifies the mathematical form of the presentation of the numerical values. The numerical values for a property can be reported directly or in terms of some specified reference conditions. The mathematical relationship involving the reference value(s) is defined here. Definition of the reference state is described in subsequent paragraphs. The **ePresentation** options are as follows: (direct value, difference between upper and lower temperatures $\{X(T_2) - X(T_1)\}$, difference between upper and lower pressures $\{X(P_2) - X(P_1)\}$, mean between upper and lower temperature $\{X(T_2) + X(T_1)\}/2$, difference from the reference state $\{X - X(\text{ref})\}$, ratio with the reference state $\{X/X(\text{ref})\}$, ratio of difference with the reference state to the reference state $\{X - X(\text{ref})\}/X(\text{ref})$).

The type of reference state is specified within the element **eRefStateType** [enumeration]. The enumerations for the reference state are as follows: (reference state with the same composition at fixed temperature and pressure; reference state with the same composition, temperature, and pressure; reference state at fixed temperature and the same pressure; reference state at the same temperature and fixed pressure; mixture at equilibrium with primary phase at the same temperature and pressure; pure components in the same proportion at the same temperature and pressure; pure solvent at the temperature of the same phase equilibrium; pure solvent at the same temperature and pressure; and pure solute at the same temperature and pressure.

nRefTemp [numerical, floating] and **nRefPressure** [numerical, floating] represent the values of the reference temperature and reference pressure, if required. **RefPhaseID** consists of **RegNum**, which is necessary in cases where the reference phase is a pure compound phase and is used in the representation of mixture data, and **eRefPhase** [enumeration] specifies the reference phase

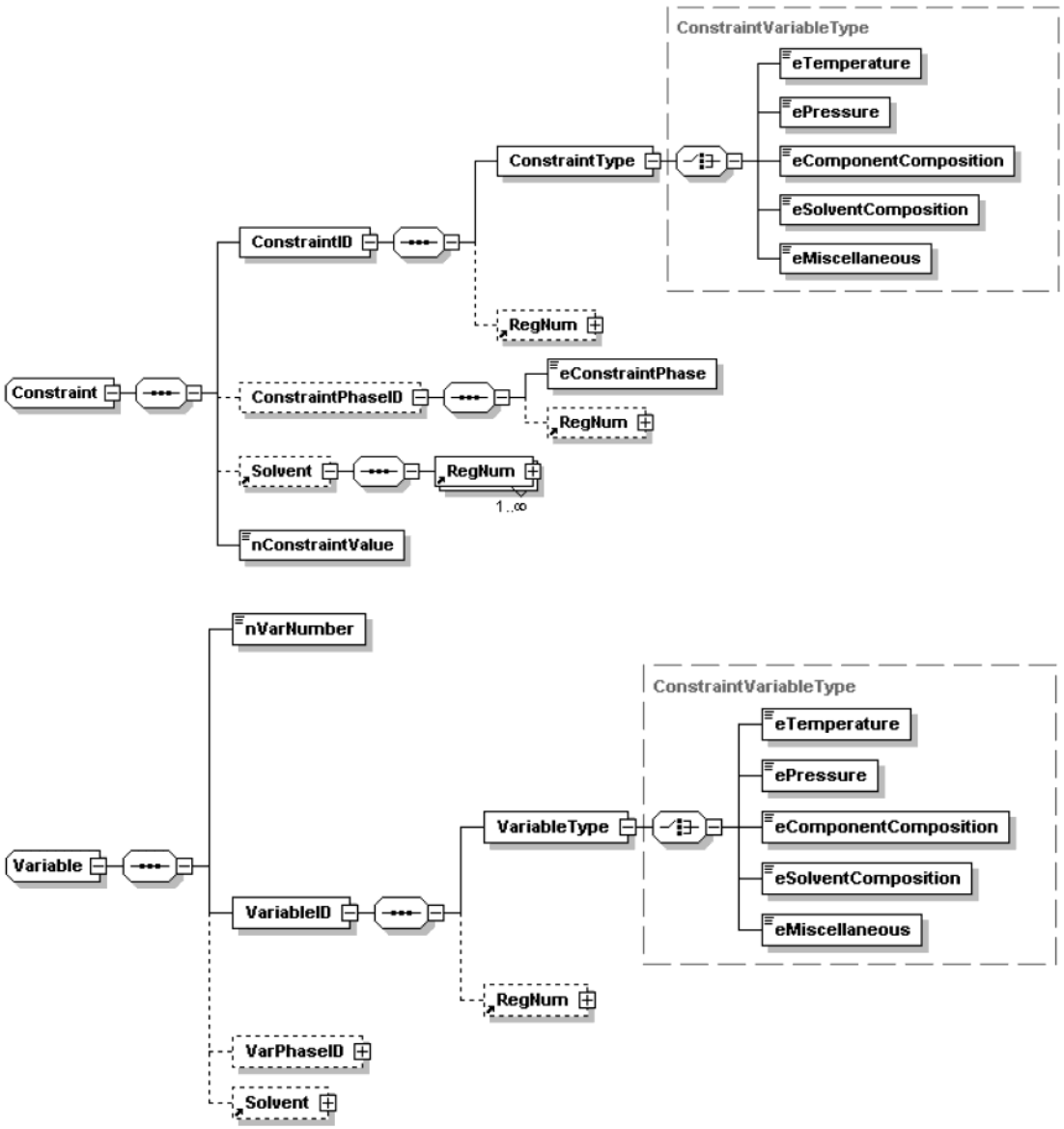

Figure 9. Structures of the **Constraint** and **Variable** elements of the "*PureOrMixtureData*" block.

from the following list: (crystal 4, crystal 3, crystal 2, crystal 1, crystal, crystal of unknown type, crystal of intercomponent compound 1, crystal of intercomponent compound 2, crystal of intercomponent compound 3, metastable crystal, glass, smectic liquid crystal, nematic liquid crystal, cholesteric liquid crystal, liquid, liquid mixture 1, liquid mixture 2, fluid [supercritical or subcritical phases], ideal gas, gas, air at 1 atm).

Solvent [complex] is identified by its components using **RegNum** (described earlier)**,** as shown in Figure 8. **eStandardState** [enumeration] specifies the thermodynamic standard state, which is required for complete specification of certain properties. The listed standard states are as follows: (pure compound, hypothetical pure solute, hypothetical unit molality solute, hypothetical unit molarity solute, and infinite dilution solute).

The element **PhaseID** [complex] (Figure 7) lists all phases in equilibrium for the property (for example crystal, liquid, and gas for the triple-point temperature, or liquid and gas for VLE data). **PhaseID** is described by **RegNum** and **ePhase** [enumeration]. Structure of **RegNum** was described earlier. Enumeration for **ePhase** is identical to that for **eRefPhase**, which was also described earlier.

The metadata elements **Constraint** and **Variable** have very similar structures, as shown in Figure 9. **Constraint** [complex] has four subelements: **ConstraintID** [complex], **ConstraintPhaseID** [complex], **Solvent** [complex], and **nConstraintValue** [numerical, floating]. **ConstraintID** consists of **RegNum** [complex, see Figure 5] and **ConstraintType** [complex]. **ConstraintType** [complex] enumerates five types of constraints: **eTemperature** [enumeration], **ePressure** [enumeration], **eComponentComposition** [enumeration], **eSolventComposition** [enumeration], and **eMiscellaneous** [enumeration]. **RegNum** should be used for mixtures only if the constraint is a composition expressed in terms of the concentration of a particular compound.

The values of the enumerated elements for **eTemperature** [enumeration] and **ePressure** [enumeration] are (temperature, K; upper temperature, K; lower temperature, K) and (pressure, kPa; upper pressure, kPa; lower pressure, kPa), respectively. The "upper" and "lower" values are required in the specification of enthalpy-increment metadata. The element **eComponentComposition** [enumeration, always defined with **RegNum**] lists (mole fraction; mass fraction; molality, mol·kg^{-1}; molarity, mol·dm^{-3};

volume fraction; moles per mass of solution, mol·kg^{-1}; mass per volume of solution, kg·m^{-3}; mole ratio to solvent; mass ratio to solvent; volume ratio to solvent; activity; and activity coefficient). The element **eSolventComposition** [enumeration, always defined with **RegNum**] lists (solvent—mole fraction; solvent—mass fraction; solvent—molality, mol·kg^{-1}; solvent—molarity, mol·dm^{-3}; solvent—volume fraction; solvent—mole ratio to other component of a binary solvent; solvent—mass ratio to other component of a binary solvent; solvent—volume ratio to other component of a binary solvent). The element **eMiscellaneous** [enumeration] identifies various other types of constraints and includes the following enumerations: (wavelength, nm; molar volume, m^3·mol^{-1}; specific volume, m^3·kg^{-1}; density, kg·m^{-3}; molar density, mol·m^{-3}; entropy, J·K^{-1}·mol^{-1}).

The structure of the element **ConstraintPhaseID** [complex] (Figure 9) is identical to the structures of **PhaseID** (Figure 7) and **RefPhaseID** (Figure 8). **Solvent** is identified through specification of **RegNum** for each solvent component. **nConstraintValue** [numerical, floating] represents the numerical value of the constraint.

The structure of the element **Variable** [complex] (Figure 9) is nearly identical to that of **Constraint** [complex]; however, **Variable** does not include the element **nConstraintValue** and includes the additional element **nVarNumber**. **nVarNumber** [numerical, integer] designates the sequential variable number for the list of variables. This reinforces correct association of numerical values with variables.

The schema element **NumValues** [complex] (Figure 7) consists of **VariableValue** [complex], which represents numerical values of variables, and **PropertyValue** [complex], which represents numerical values of properties. Each contains a sequential identifier for the variable or property (**nVarNumber** for **VariableValue** and **nPropNumber** for **PropertyValue**) and numerical values (**nVarValue** [numerical, floating] for **VariableValue** and **nPropValue** [numerical, floating] for **PropertyValue**).

4. "ReactionData" Block. The "*ReactionData*" block is for storage of data for chemical reactions, and is shown in Figure 10. This block includes a number of elements, **sExpPurpose**, **sCompiler, sContributor**, **dateDateAdded**, and **NumValues** [complex], which are identical to those used in the "*PureOrMixtureData*" block and were described earlier. **Variable** [complex] and **Constraint** [complex] differ from those in the "*PureOrMixtureData*" block by the absence of **Solvent** and **PhaseID** elements (Figure 13).

The element **Participant** [complex] (Figure 11) stores information about a participant in a chemical reaction. This element includes **RegNum** [complex, see Figure 5], **nSampleNum** [numerical, integer, see Figure 5], **nStoichiometricCoef** [numerical, floating], to store stoichiometric coefficients (negative values for reactants and positive for products), **ePhase** [enumeration, with values the same as those quotes for **ePhase** in the **PhaseID** element of the '*PureOrMixtureData*' block], **eCompositionRepresentation** [enumeration], and **nNumericalComposition** [numerical, floating].

eCompositionRepresentation [enumeration] (Figure 11) stores the composition representation for a participant (mole ratio of solvent to participant; molality—moles of participant per kilogram of solvent, mol·kg^{-1}; moles of participant per kilogram of solution, mol·kg^{-1}; molarity—moles of participant per liter of solution; mole ratio—moles of participant per mole of solvent; mass ratio—mass of participant per mass of solvent; volume ratio—volume of participant per volume of solvent; mass of participant per volume of solution, kg·m^{-3}). **nNumericalComposition** [numerical, floating] (Figure 11) indicates the numerical value of the composition representation. **eCompositionRepresentation** and **nNumericalComposition** are used for change-of-state reactions only.

The element **eReactionType** [enumeration] (Figure 10) stores a description of the general type of chemical reaction. The complete enumeration list includes the following: (combustion with oxygen, combustion with other elements or compounds, addition of various compounds to unsaturated compounds, addition of water to a liquid or solid to produce a hydrate, atomization or formation from atoms, esterification, exchange of alkyl groups, exchange of hydrogen atoms with other groups, formation of a compound from elements in their stable state, halogenation—addition of or replacement by a halogen, hydrogenation—addition of hydrogen molecules to unsaturated compounds, hydrohalogenation, hydrolysis of ions, other reactions with water, ion exchange, neutralization, oxidation with oxidizing agents other than oxygen, oxidation with oxygen, homonuclear dimerization, polymerization—all other types, solvolysis—solvents other than water, stereoisomerization, structural isomerization, other reactions).

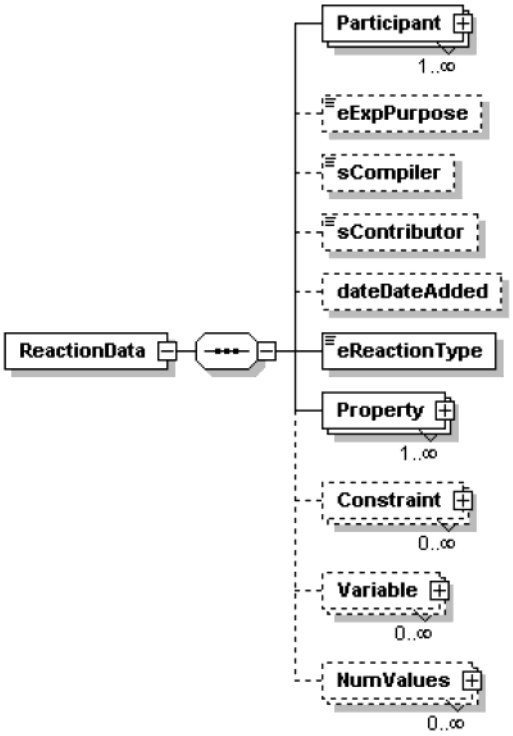

Figure 10. General schema of the "*ReactionData*" block.

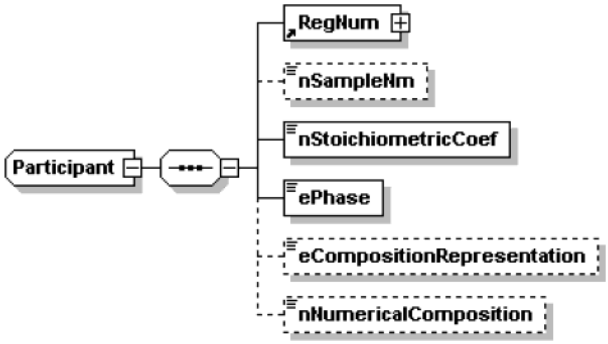

Figure 11. Structure of the **Participant** element of the "*ReactionData*" block.

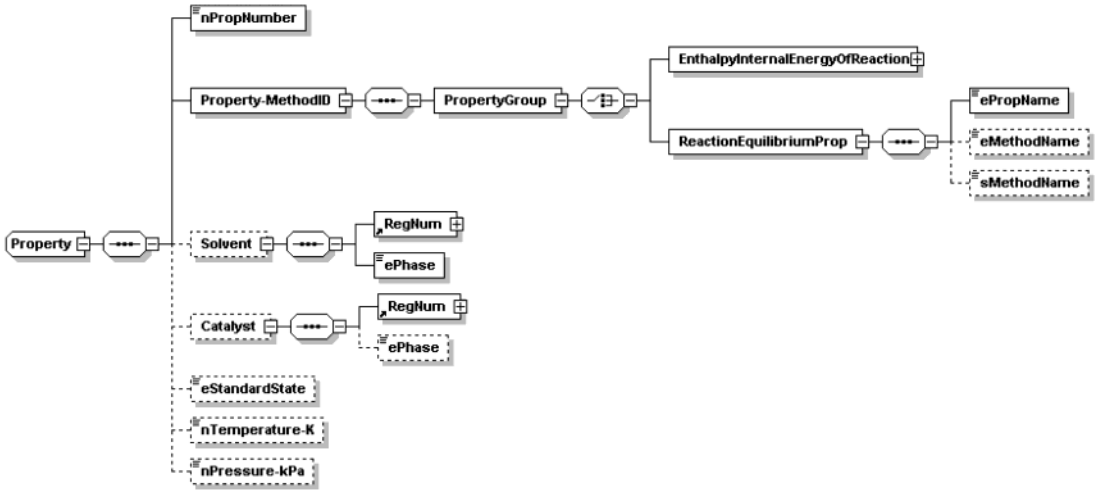

Figure 12. Structure of the **Property** element of the "*ReactionData*" block.

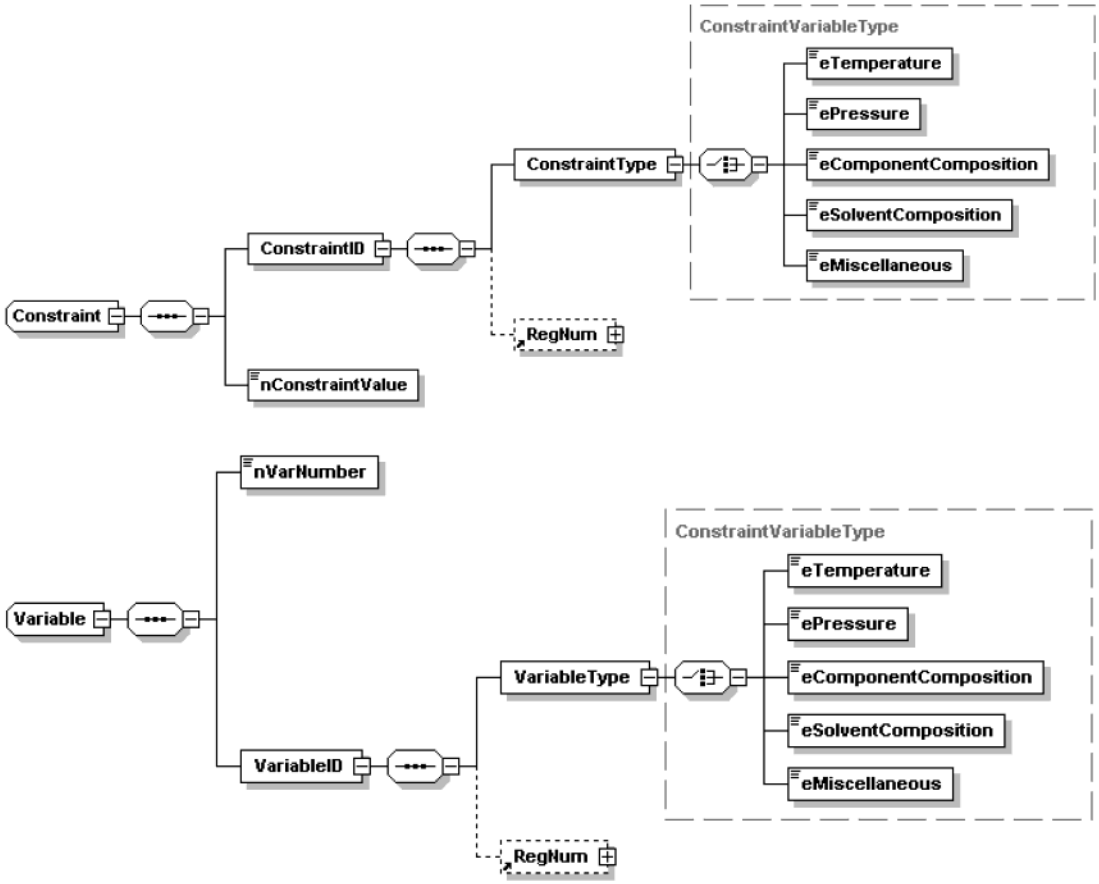

Figure 13. Structures of the **Variable** and **Constraint** elements of the "ReactionData" block.

The element **Property** [complex, Figure 12] is similar in structure to that used in "*PureOrMixtureData*". However, instead of the 10 property groups used in the "*PureOrMixtureData*" structure, the **Property** [complex] block here has only two: **EnthalpyInternalEnergy** [complex], which describes thermochemical properties for change-of-state reactions such as combustion with oxygen, and **ReactionEquilibriumProp** [complex], which describes properties for reactions in equilibrium. Both property groups are characterized with **ePropName** [enumeration] for identification of properties and **eMethodName** [enumeration], which specifies the experimental methods used.

EnthalpyInternalEnergy *[complex] (Figure 12)*

The element **ePropName** [enumeration] includes the following properties: (enthalpy of reaction, $kJ \cdot mol^{-1}$; internal energy, $J \cdot g^{-1}$; internal energy of reaction—mole basis, $kJ \cdot mol^{-1}$).

eMethodName [enumeration] includes the following experimental methods: (static bomb calorimetry, rotating bomb calorimetry, microbomb calorimetry, flame calorimetry).

ReactionEquilibriumProp *[complex] (Figure 12)*

ePropName [enumeration] includes the following properties: (thermodynamic equilibrium constant; apparent

equilibrium constant in terms of molality, $(mol \cdot kg^{-1})^n$; apparent equilibrium constant in terms of molarity, $(mol \cdot dm^{-3})^n$; apparent equilibrium constant, in terms of partial pressure, $(kPa)^n$; apparent constant in terms of mole fraction). Here *n* represents the difference in the number of moles between the products and reactants. In the case of partial pressures, this difference is only for compounds in the gas phase.

eMethodName [enumeration] includes the following experimental methods: (static equilibration, dynamic equilibration, chromatography, IR spectrometry, UV spectroscopy, NMR spectrometry, titration, other).

Solvent [complex] and **Catalyst** [complex] (Figure 12) have essentially identical structures both characterized with **ePhase** [enumeration] and **RegNum** [complex], as described earlier.

Schema Validation: Extent and Strategy

The developed schema was validated extensively with data records in SOURCE.[3] Validation covered essentially all properties within the scope of ThermoML, including those of pure compounds, multicomponent mixtures, and chemical reactions. More than 5000 data sets from more than 3000 publications were used at the TRC Data Entry Facility to validate the schema. In addition, validation included data files submitted to TRC by authors of future publications submitted through the Advisory Board of the *Journal of Chemical and Engineering Data*, as well as data files submitted to TRC by its data collection contractors.

Role of ThermoML in Global Data Submission and Dissemination

The role of ThermoML in global submission and dissemination of experimental thermodynamic property data is represented in Figure 14. Guided Data Capture (GDC) software[11] has been developed at TRC for mass-scale abstraction from the literature of experimental thermophysical and thermochemical property data. Property values are captured with a strictly hierarchical system based upon rigorous application of the thermodynamic constraints of the Gibbs phase rule with full traceability to source documents. This software is freely available for download from the Internet.[12] Following the peer-review process, authors are requested by the journal editors to download and use the GDC software to capture the experimental property that has been accepted for publication. The output of the GDC software is a batch data capture file (plain text file), which is submitted directly to NIST/TRC. After additional consistency tests at the TRC Data Entry Facility, the files are converted into ThermoML format with software (TransThermo) developed at NIST/TRC. Upon release of the manuscript for publication, ThermoML files are posted on the TRC Web site for unrestricted public access. This procedure has been established formally by the Advisory Board of the *Journal of Chemical and Engineering Data*. Expansion of this operation to other journals in the field (*Journal of Chemical Thermodynamics*, *International Journal of Thermophysics*) is currently being discussed.

Use Cases and ThermoML Schema Text

Several examples illustrating the format of the data files created with the ThermoML format for pure compound and mixture data sets,[13] as well as for change-of-state[14] and equilibrium reactions,[15] are included as Supporting Information. The examples are based upon experimental studies published in the *Journal of Chemical and Engineering Data*. The complete text of the ThermoML schema is included also as Supporting Information and is available on the Web (http://pubs.acs.org) or through direct request to the authors.

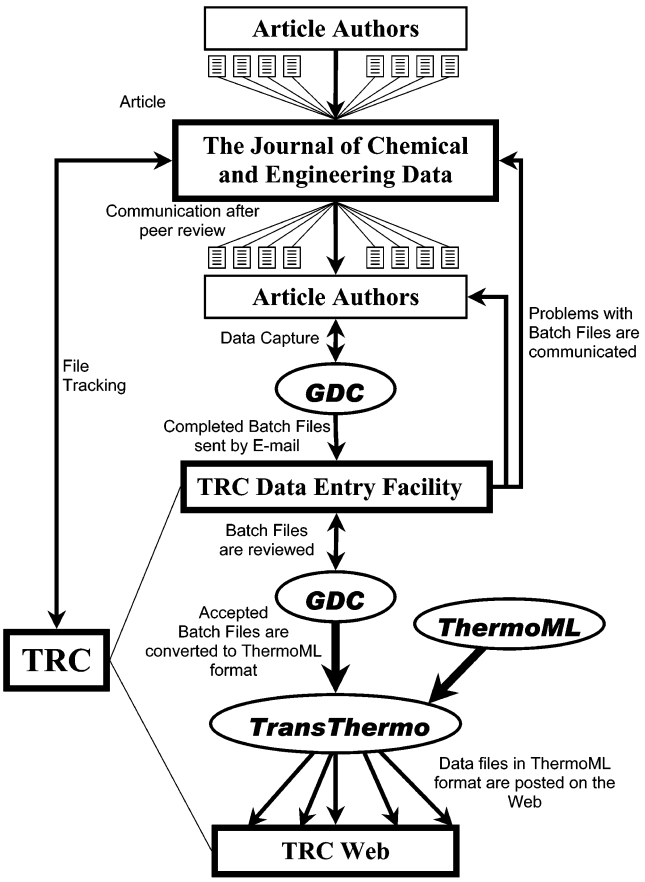

Figure 14. Information flow between article authors, the *Journal of Chemical and Engineering Data*, and NIST/TRC for data submission and dissemination.

Acknowledgment

The authors express their deep appreciation to the following people for their valuable advice in the development and completion of the ThermoML format and in its implementation for data processing: Professor W. A. Wakeham (University of Southampton, U.K.), Dr. J. H. Dymond (University of Glasgow, U.K.), Dr. A. D. Dewan (Shell Global Solutions, Inc.; Houston, TX), Dr. M. A. Satyro (Virtual Materials Group, Inc.; Calgary, Canada), and Drs. J. W. Magee and W. M. Haynes (NIST, Boulder, CO). Our special thanks to Dr. D. G. Friend of NIST/Boulder for enhancing communications and cooperation between NIST/TRC and the DIPPR 991 research groups.

Supporting Information Available:

Several examples illustrating the use of ThermoML to represent experimental data for pure compounds, mixtures, and chemical reactions as well as complete text of ThermoML. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Wilhoit, R. C.; Marsh, K. N. *CODATA STANDARD Thermodynamics. Rules for Preparing a COSTAT Message for Transmitting Thermodynamic Data*; Report to CODATA Task Group on Geothermodynamic Data and Chemical Thermodynamic Tables; Paris, 1987.
(2) www.fiz-karlsruhe.de/dataexplorer/test/iucosped/dataexplorer.html.

(3) 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*, 215−226.
(4) Frenkel, M. Dynamic Compilation: A Key Concept for Future Thermophysical Data Evaluation. In *Report on Forum 2000: Fluid Properties for New Technologies—Connecting Virtual Design with Physical Reality*; Rainwater, J. C., Friend, D. G., Hanley, H. J. M., Harvey, A. H., Holcomb, C. D., Laesecke, A., Magee, J. W., Muzny, C., Eds.; NIST Special Publication 975,83; Gaithersburg, 2001.
(5) Wilhoit, R. C.; Marsh, K. N. Automation of Numerical Data Compilation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 17.
(6) Wilhoit, R. C. The DataFetch Library of Functions for the Retrieval and Interpretation of Thermophysical Data from the TRC SOURCE Database. *Int. J. Thermophys.* **2002**, *23*, 187−197.
(7) Finkelstein, C.; Aiken, P. *Building Corporate Portals with XML*; McGraw-Hill: New York, 1999.
(8) Dong, Q.; Yan, X.; Wilhoit, R. C.; Hong, X.; Chirico, R. D.; Diky, V. V.; Frenkel, M. Data Quality Assurance for Thermophysical Property Databases—Applications to the TRC SOURCE Data System. *J. Chem. Inf. Comput. Sci.* **2002,** *42*, 473−480.
(9) Whiting, W. B. Effects of Uncertainties in Thermodynamic Data and Models on Process Calculation. *J. Chem. Eng. Data* **1996**, *41*, 935−941.
(10) *XML SPY v. 4.4 u.* ALTOVA GmbH and ALTOVA, Inc.: 1998−2002.
(11) Diky, V. V.; Chirico, R. D.; Wilhoit, R. C.; Dong, Q.; Frenkel, M. Windows-Based Guided Data Capture Software for Mass-Scale Thermophysical and Thermochemical Property Data Collection. *J. Chem. Inf. Comput. Sci.* (in press).
(12) www.trc.nist.gov.
(13) Chylinski, K.; Fras, Z.; Malanowski, S. K. Vapor-Liquid Equilibrium in Phenol + 2-Ethoxyethanol at 363.15 K to 383.15 K. *J. Chem. Eng. Data* **2001**, *46*, 29−33.
(14) Hamilton, W. S.; Thompson, P.; Pustejovsky, S. The Enthalpies of Combustion and Formation of 2-Methyl-2-oxazoline and 2-Ethyl-2-oxazoline. *J. Chem. Eng. Data* **1976**, *21*, 428−429.
(15) Rihko, L. K.; Linnekoski, J. A.; Krause, A. O. Reaction Equilibria in the Synthesis of 2-Methoxy-2-methylbutane and 2-Ethoxy-2-methylbutane in the Liquid Phase. *J. Chem. Eng. Data* **1994**, *39*, 700−704.

Received for review November 19, 2002. Accepted November 20, 2002.

JE025645O