Pure Appl. Chem., Vol. 83, No. 10, pp. 1937–1969, 2011.
doi:10.1351/PAC-REC-11-05-01

Extension of ThermoML: The IUPAC standard for thermodynamic data communications (IUPAC Recommendations 2011)*,§

Michael Frenkel[1,‡], Robert D. Chirico[1], Vladimir Diky[1], Paul L. Brown[2], John H. Dymond[3], Robert N. Goldberg[4], Anthony R. H. Goodwin[5], Heiko Heerklotz[6], Erich Königsberger[7], John E. Ladbury[8], Kenneth N. Marsh[9], David P. Remeta[10], Stephen E. Stein[11], William A. Wakeham[12], and Peter A. Williams[13]

[1]Thermophysical Properties Division, National Institute of Standards and Technology, Boulder, CO 80305, USA; [2]Rio Tinto Technology and Innovation, Bundoora, VIC 3083, Australia; [3]Chemistry Department, University of Glasgow, Glasgow G12 8QQ, UK; [4]Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA and Department of Chemistry and Biochemistry, University of Maryland, Baltimore County, Baltimore, MD 21250, USA; [5]Schlumberger Technology Corporation, Sugar Land, TX 77478, USA; [6]Department of Pharmacy, University of Toronto, Toronto, ON M5S 3M2, Canada; [7]Faculty of Science and Engineering, School of Chemical and Mathematical Sciences, Murdoch University, Murdoch, WA 6150, Australia; [8]University of Texas, M.D. Anderson Cancer Center, Houston, TX 77030, USA; [9]Department of Chemical and Process Engineering, University of Canterbury, Christchurch, New Zealand; [10]Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854, USA; [11]Chemical and Biochemical Reference Data Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA; [12]School of Engineering Sciences, University of Southampton, Highfield, Southampton SO17 1BJ, UK; [13]School of Natural Sciences, University of Western Sydney, Penrith South, NSW 1797, Australia

Abstract: ThermoML is an XML-based approach for storage and exchange of experimental, predicted, and critically evaluated thermophysical and thermochemical property data. Extensions to the ThermoML schema for the representation of speciation, complex equilibria, and properties of biomaterials are described. The texts of 14 data files illustrating the new extensions are provided as Supplementary Information together with the complete text of the updated ThermoML schema.

Keywords: communications; data; standards; standardization; thermochemistry; thermodynamics; ThermoML.

*Sponsoring body: IUPAC Committee on Printed and Electronic Publications: see more details on p. 1965.

§Trade names are provided in the text only to specify procedures adequately, and do not imply endorsement by the U.S. National Institute of Standards and Technology. Similar products by other manufacturers may be found to work as well or better.

‡Corresponding author

INTRODUCTION

While thermodynamic property data represent a key foundation for development and improvement of all chemical process technologies, a lack of standardization for communicating these data for many years has been a major obstacle in establishing efficient information-delivery processes from measurement to data-management system and from data-management system to engineering application. The challenges related to the establishment of robust communication channels are numerous and are associated primarily with the necessity to assure compliance with the myriad of existing and to be developed algorithmic languages, operating systems, and computational platforms. In addition, communication of thermophysical and thermochemical property data is further complicated by the nature of their metadata infrastructure, which includes more than 100 interrelated properties with associated variables, constraints, phases, and measures of uncertainty [1–4]. The combination of these challenges has made standardization of thermophysical and thermochemical property data communications an insurmountable task for many years in spite of a number of projects initiated between 1985 and 2000 to accomplish this goal [5–8].

In 2002, IUPAC established project 2002-055-3-024, "XML-based IUPAC Standard for Experimental and Critically Evaluated Thermodynamic Property Data Storage and Capture" [9], to create standardized mechanisms for thermodynamic data communications with XML (Extensible Markup Language) technology. This project was an activity of the Committee on Printed and Electronic Publications [10]. XML technology [11], fully developed within the last 10 years, provides significant advantages for the development of standards for data exchange, such as its native interoperability based on ASCII code, its modular nature, and its transparent readability by both humans and computers. From a practical standpoint, it is also critical that this technology is currently supported by both the software and hardware industries (see, e.g., IBM XML Toolkit [12] and Microsoft XML Downloads [13]). The project was successfully completed in 2006 with the establishment of ThermoML [14] as a standard for thermophysical and thermochemical property data communications [15].

The IUPAC Recommendations of 2006 [15] (ThermoML-06) provided a comprehensive summary describing the original formulation of ThermoML for representation of experimental data [16], extensions to the schema for representation of uncertainties [17], and further extensions for representation of predicted data, critically evaluated data, and fitting equations [18]. ThermoML covers essentially all thermodynamic and transport property data (more than 100 properties) for pure compounds, multicomponent mixtures, and chemical reactions (including change-of-state and equilibrium reactions).

ThermoML-06 played a key role in the establishment of one of the first global data delivery processes [19] that is now endorsed by five major journals publishing experimental thermophysical and thermochemical property data: the *Journal of Chemical and Engineering Data*, *Fluid Phase Equilibria*, the *Journal of Chemical Thermodynamics*, the *International Journal of Thermophysics*, and *Thermochimica Acta* [20], and an archive of published experimental data from these journals is now freely available on the Web in ThermoML format [21]. ThermoML was also a key component in the implementation of the concepts of Global Information Systems in Science (GISS) in application to the field of thermodynamics [2] and chemical process and product design on-demand [4].

In 2007, IUPAC initiated project 2007-039-1-024, "Extension of ThermoML—the IUPAC Standard for Thermodynamic Data Communications" [22], in order to broaden the scope of ThermoML to support storage and exchange of thermodynamic property data for (1) thermodynamic properties of biomaterials and (2) speciation and complex equilibria. The Task Group established for the project conducted two meetings. The first meeting, held in Boulder, Colorado (USA) in June 2009, resulted in approval of the ThermoML expansion for thermophysical and thermochemical properties of biomaterials. The second meeting, held in Tsukuba (Japan) in August 2010, led to the approval of the description of speciation and complex equilibria. Much of the material provided here as the IUPAC Recommendations for project 2007-039-1-024 was published previously in articles describing the formulation of ThermoML for representation of data for biomaterials [23] and extensions to the schema for representation of speciation and complex equilibria [24]. Every effort was made to ensure that information represented with the formats described in these earlier articles would remain valid within the new IUPAC standard version of ThermoML (Version 4.0). Several minor changes were made to improve consistency in tag names, and to eliminate unnecessary elements. These changes might invalidate files created with the earlier version of ThermoML, and could require minor adjustment in the file structure to bring it into compliance with the new schema definitions.

BASIC PRINCIPLES

Schema structure

As described for ThermoML-06 [15], the schema structure incorporates structural elements related to the basic principles of phenomenological thermodynamics with elements for representation of thermophysical and thermochemical properties, state variables, system constraints, phases, and units. Meta- and numerical data records are grouped into "nested blocks" of information corresponding to data sets. Metadata records precede numerical data information, providing a robust foundation for generating "header" records for any relational database where ThermoML-formatted files might be incorporated. Elements for comprehensive representation of uncertainties are included with all definitions and descriptions in full accord with the *Guide to the Expression of Uncertainty in Measurement*, ISO (International Organization for Standardization), October, 1993 [25] (the GUM), and the *U.S. Guide to the Expression of Uncertainty in Measuremen*t [26].

Tagging

IUPAC 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. Many tags are fully self-explanatory, and few abbreviations are used, with the goal of minimizing the time needed by users to understand the schema and to convert data formatted with ThermoML with customized software or commercial XML parsers.

Modularity

ThermoML was designed to take advantage of the modular nature of XML schemas so that the scope could be expanded easily into new areas [15]. The extensions described here take full advantage of this feature, which allows additions to be made without changing the basic schema structure, elements, and subelements. Only the extensions necessary for the representation of properties for biomaterials and systems involving speciation and complex equilibria are described here. In nearly all cases, elements described earlier [15] remain unchanged. Recently, some standard terms have been established in the field of polymer science [27,28] that necessitated several changes to elements for polymer sample descriptions. These changes are noted clearly in the text.

Units

There is only one unit allowed for each property represented in ThermoML, and these are SI-based [29,30]. Unit conversions are outside the scope of ThermoML. Unit tagging is explicitly propagated to every numerical property value in a ThermoML file through the property name, thus minimizing the likelihood of unit misinterpretation.

Conventions for names of elements in the ThermoML schema

Element names (or "*tags*") include special characters related to the type of information to be stored. A name beginning with "e" indicates an *enumeration* element (with values selected from a predefined list), "s" designates *string* elements (text strings), "n" specifies *numerical* elements (integer or floating), "yr" designates elements characterizing the *year*, "date" specifies *date* elements, and "url" indicates elements specifying addresses on the World Wide Web. Elements shown as dotted boxes in the figures are optional, while those shown as solid-lined boxes are mandatory. A *complex* element is an element that includes subelements. Complex elements illustrated without their internal structure are identified by "+" at the right-hand edge of the box. Multiple elements of the same type are often needed within the schema to specify such elements as multiple authors for a given citation or multiple property values for a given data type. These multiple elements are identified in the figures by lower and upper limits listed below the relevant boxes. The only limits used for repeated elements are "0...∞" for optional elements and "1...∞" for mandatory elements. A switch symbol in a figure indicates that only one of the subelements can be selected. In addition, an element can have associated *attributes* that provide additional information about the contained information. Prior to the extensions described here, *attributes* were not used in ThermoML. Here, they are used for one element to avoid unnecessary duplication.

Scope of ThermoML

As noted earlier, ThermoML covers essentially all experimentally determined thermodynamic and transport property data for pure compounds, multicomponent mixtures, and chemical reactions (including change-of-state and equilibrium) with full allowance for data provenance. Specification of the data source (bibliographic information), method of property generation (experimental, predicted, critically evaluated), and multiple uncertainty assessments (with specified assessors) are included. The list of all properties within the scope of ThermoML-06 is provided with the schema description [15].

GENERAL IMPLEMENTATION OF EXTENSIONS RELATED TO BIOTHERMODYNAMIC DATA IN ThermoML

Scope of extensions for biothermodynamic data

For ThermoML to be applicable to biothermodynamic data, extensions were required in three general areas. First, unambiguous identification of biochemical compounds and biological materials (such as proteins and enzymes) posed a major new challenge. This required addition of new schema elements for specific identification numbers, such as the Enzyme Commission (EC) number [31] and the Protein Data Bank (PDB) identifier [32]. Second, new properties specific to the field of biothermodynamics, such as the *apparent equilibrium constant* for a biochemical reaction, were needed. These can include the effects of dissociation, denaturation, partial unfolding, local dynamic changes, solvent binding, and protonation events that may occur on formation of a biomolecular complex. Also, properties associated with structural changes within specific molecules or groups of molecules in a complex solution, such as denaturation of proteins or phase transitions observed with lipid membrane interactions, required further schema extensions. Finally, solvents for biochemical reactions and properties must be carefully characterized. These are commonly much more complex than the reaction solvents that are presently accommodated in ThermoML. Extensions to represent important variables, such as pH, pMg, buffer composition, cofactors, etc., had to be considered for full specification of the biochemical systems.

A key challenge within the current project is the necessity to ensure that clear and consistent data definitions and nomenclature are used throughout, and that the definitions and nomenclature are consistent to the fullest extent possible with existing IUPAC recommendations and standards. Extensions to ThermoML involved two distinct types of thermodynamic measurements: those involving properties of reactions (bond making and/or breaking), and those involving physical properties (phase transition properties, heat capacities, etc.) Specifically, the extensions to ThermoML described here are designed for representation of results from four common types of experiments in the field of biothermodynamic property measurements. These are

1. properties of enzyme-catalyzed reactions,
2. reaction properties determined with titration calorimetry,
3. properties determined with differential scanning calorimetry (DSC), and
4. solubilities in complex media.

An extensive database of thermodynamic properties for enzyme-catalyzed reactions has been compiled by Goldberg et al. [33]. Full representation of the information in this database was a key goal of this project.

Joint recommendations for nomenclature and tables in biochemical thermodynamics were issued by IUPAC and the International Union of Biochemistry and Molecular Biology (IUBMB) in 1994 [34]. Some aspects of the 1994 recommendations have been revised recently under the auspices of IUBMB [35]. The recommendations of 1994 [34] and 2011 [35] revised and extended recommendations published in 1976 [36] and 1985 [37] and continue today as the foundation for the reporting of reaction data for biothermodynamics. The recommendations of 1994 and 2011 specifically address biochemical reactions that consist of species in equilibrium with each other. These reactions do not balance elements that are assumed fixed, such as hydrogen at constant pH. This approach leads to specification of *apparent equilibrium constants* written in terms of sums of species, and to calculation of *transformed* thermodynamic functions for reactions. This formalism results in the addition of several variables that extend those typically specified in chemical engineering applications (temperature, pressure, density, composition). For example, the *apparent equilibrium constant* is a function of temperature, pressure, ionic strength, pH, and pMg (various metal ions can be involved, but Mg^{2+} is used here as an example), and can be contrasted with the *thermodynamic equilibrium constant*, which is a function of temperature only.

Data representations for most physical properties of biochemical compounds of defined composition are fully realized in the ThermoML-06 schema. An important group of properties that is outside the scope of ThermoML-06 involves “phase transitions” and conformational changes (e.g., denaturation) in biological systems, such as proteins and nucleic acid sequences. These biochemical properties are typically associated with specific chemical species within a mixture, such as a protein in an aqueous solution. Furthermore, several properties and variables beyond the transition temperature and enthalpy are needed to fully characterize biochemical transitions. Additional properties, such as the zero Gibbs energy temperature, T_G, or the heat capacity change at a transition temperature, are used to characterize the temperature dependence of the heat capacity associated with the biochemical transition. Additional variables are analogous to those required for reactions and include ionic strength and pH. IUPAC recommendations for reporting such experimental data were discussed by Hinz and Schwarz [38].

IMPLEMENTATION OF STRUCTURAL ELEMENTS RELATED TO BIOTHERMODYNAMIC DATA IN ThermoML

ThermoML consists of four major blocks, which are shown in Fig. 1 together with the element **Version** [complex], which specifies the ThermoML version number. All elements of the four major blocks in ThermoML-06 were described previously [15]. The four major blocks are:

1. *Citation* (description of the source of the data)
2. *Compound* (characterization of the chemical system)
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 change of state or in chemical equilibrium)

The general locations of extensions described in this paper for biothermodynamic data are indicated with black arrows in Figs. 1–4. (General locations of extensions for speciation and complex equilibria are indicated with gray-filled arrows, and some necessary miscellaneous extensions are indicated with white-filled arrows.) Arrows that point to the right indicate that extensions are present within the subelements of the complex element. Arrows that point to the left indicate specific schema extensions (new elements) described in the text. Detailed schema figures and the text of ThermoML were created with the software package XML SPY [39]. Some elements include the text "tml:" to the left of the element tag. This is an artifact of the software used to produce the figures and should be ignored.

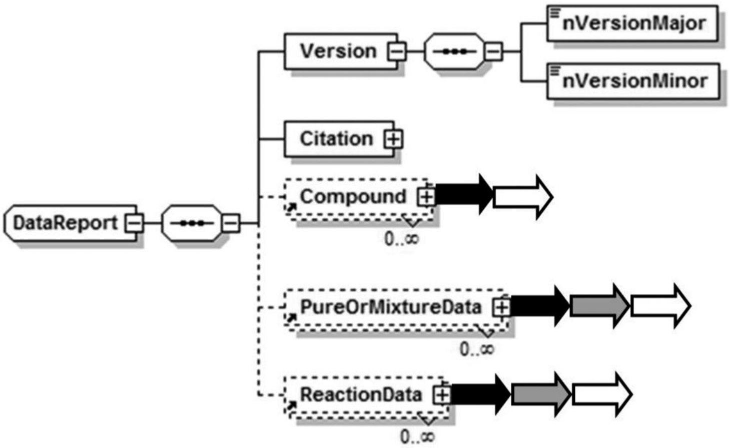

Fig. 1 Major components of the ThermoML schema. The arrows indicate locations of extensions to the schema: black for biomaterials, gray-filled for speciation and complex equilibria, and white-filled for other miscellaneous extensions described in this report.

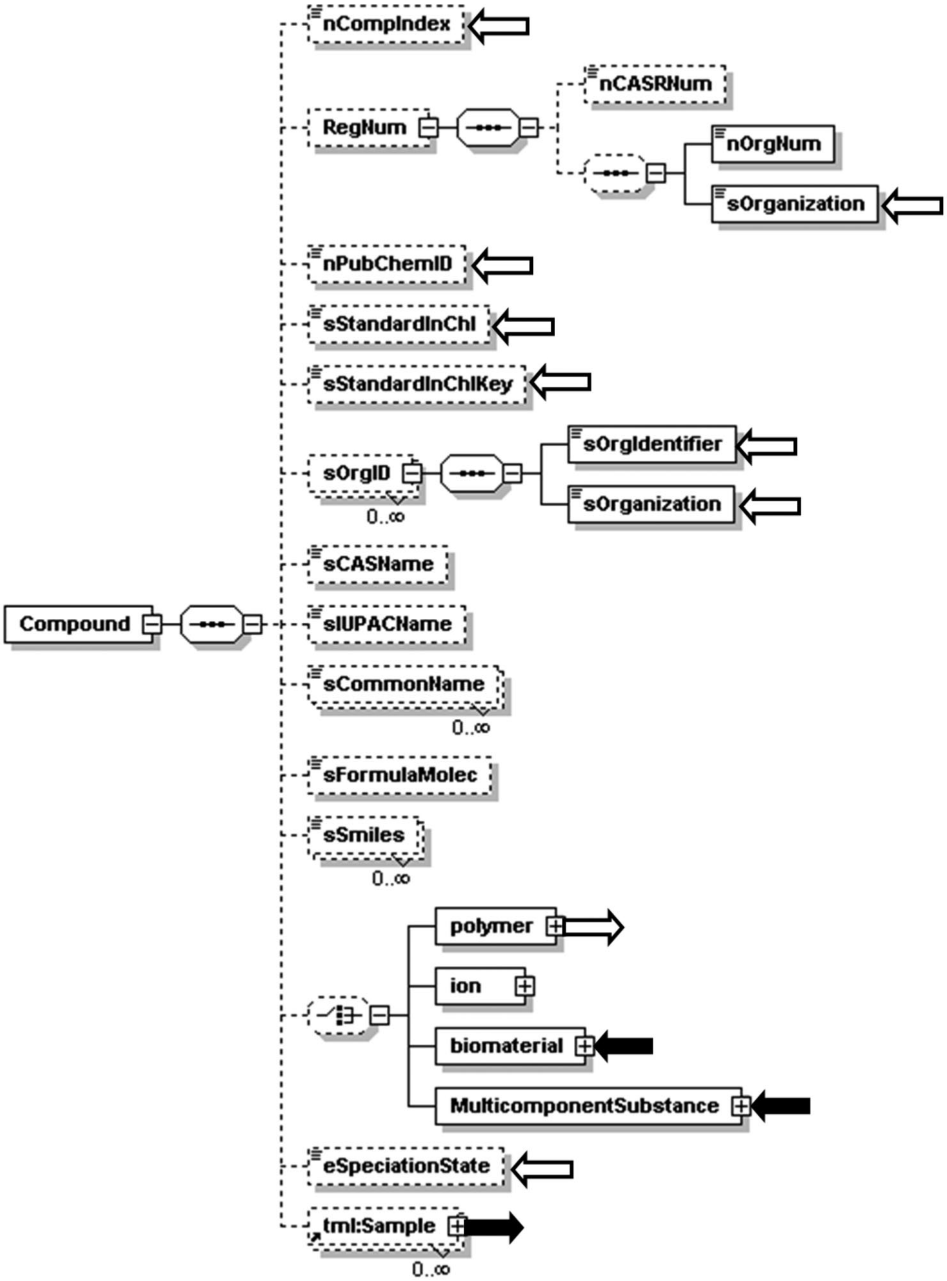

Fig. 2 Structure of the *Compound* block. The arrows indicate the locations of extensions described in the text.

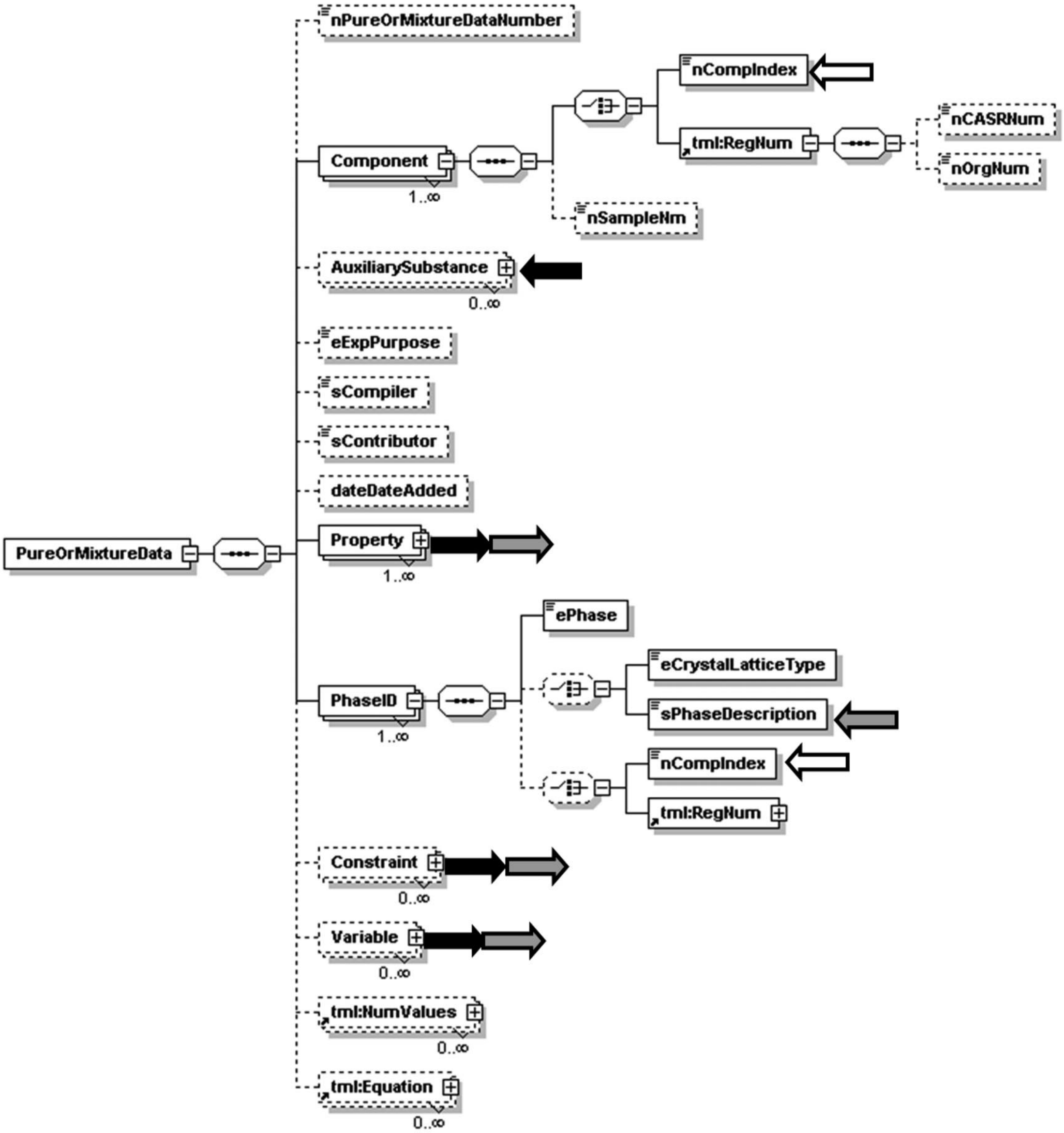

Fig. 3 Structure of the *PureOrMixtureData* block. The arrows indicate the locations of extensions described in the text.

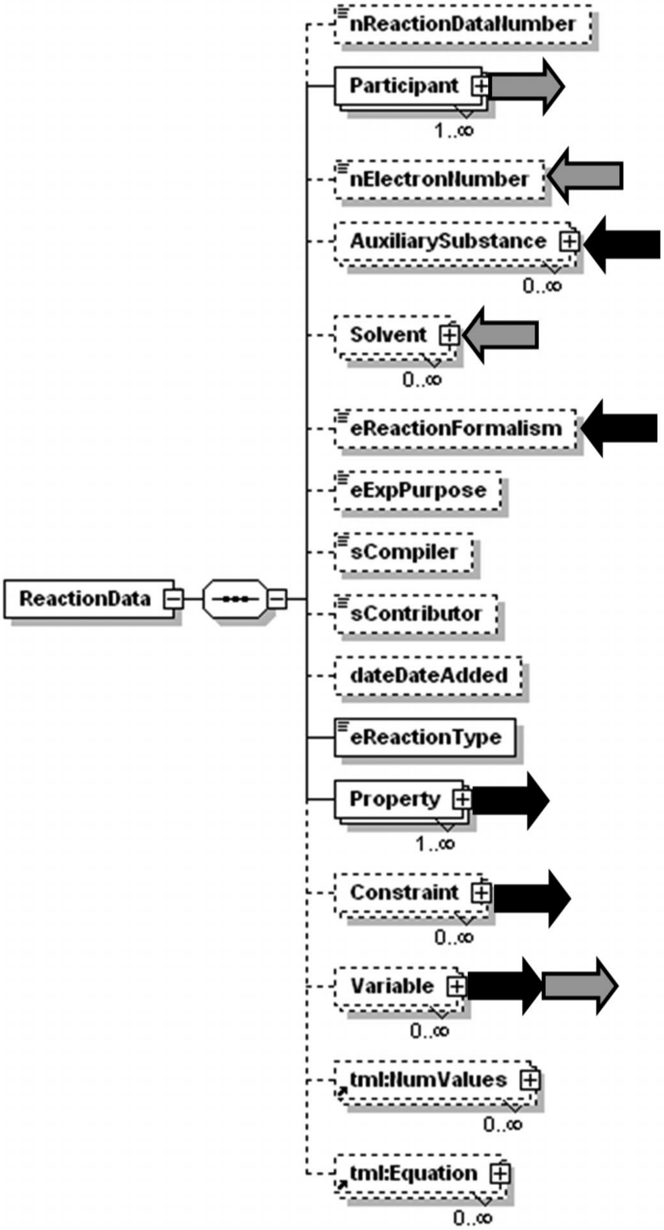

Fig. 4 Structure of the *ReactionData* block. The arrows indicate the locations of extensions described in the text.

Structural elements related to biothermodynamic data in the *Compound* block

Elements of the *Compound* block are shown in Fig. 2. Unique identification of bio-related compounds and materials is a major challenge. Extensions for specification of biomaterials are contained within the new **biomaterial** [complex] and **MulticomponentSubstance** [complex] elements, and within a subelement of the **Sample** [complex] element.

Two identification numbers that are widely used and accepted within the biothermodynamics community are included as the subelements of **biomaterial** [complex] (Fig. 5). The subelement

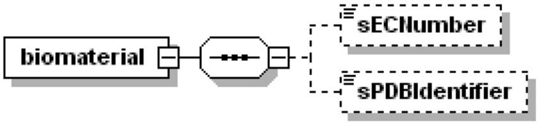

Fig. 5 Structure of the **biomaterials** [complex] element in the *Compound* block. Subelements are **sECNumber** [string] (the EC number) and **sPBDIdentifier** [string] (the PDB identifier).

sECNumber [string] contains the Enzyme Commission (EC) number. Enzymes are assigned with numerical identification numbers under the auspices of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) in consultation with the IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN). Details are provided on the web site maintained by the committee [31]. The subelement **sPDBIdentifier** [string] is the PDB identifier. These numbers are maintained by the Research Collaboratory for Structural Bioinformatics (RCSB), a nonprofit consortium dedicated to a better understanding of the function of biological systems through study of the 3D structure of biological macromolecules. RCSB members work cooperatively through joint grants and subsequently provide free public resources and publications to assist others and further the fields of bioinformatics and biology. The PDB is publicly available online [32].

The structure of the new element **MulticomponentSubstance** [complex] is shown in Fig. 6. The element has two attributes: *type* and *compositionBasis*. The enumeration list associated with the attribute *type* is (*alloy*, *clathrate*, *complex*, *crystal*, *solution*). The attribute *compositionBasis* has three possible choices: *mass fraction*, *mole fraction*, and *number of molecules*. The composition basis *number of molecules* is used to define the stoichiometry of a complex, such as A_2B, which is composed of two molecules of A and one molecule of B. Numerical values associated with composition basis are stored in the **nAmount** [numerical, floating] element (Fig. 6). The element **RegNum** [complex] or **nCompIndex** [numerical, integer] associates the particular component of the complex with a compound that is fully specified within the *Compound* block (Fig. 2).

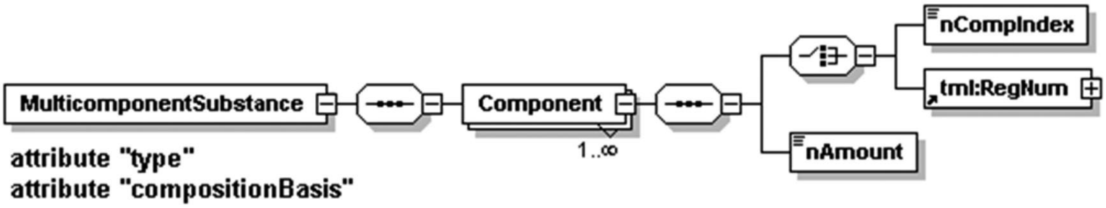

Fig. 6 Structure of the **MulticomponentSubstance** [complex] element within the *Compound* block. This element has two attributes: *type* and *composition Basis*.

The final extensions within the *Compound* block (Fig. 2) involve addition of elements for the **Sample** [complex]. Figure 7 shows the structure of the **Sample** [complex] block, including the new subelement **ComponentSample** [complex]. This new element is used to define a particular sample of a component within a multicomponent substance, such as a hydrate or other complex. The subelements of **ComponentSample** [complex] are **RegNum** [complex] or **nCompIndex** [numerical integer] and **nSampleNm** [numerical, integer].

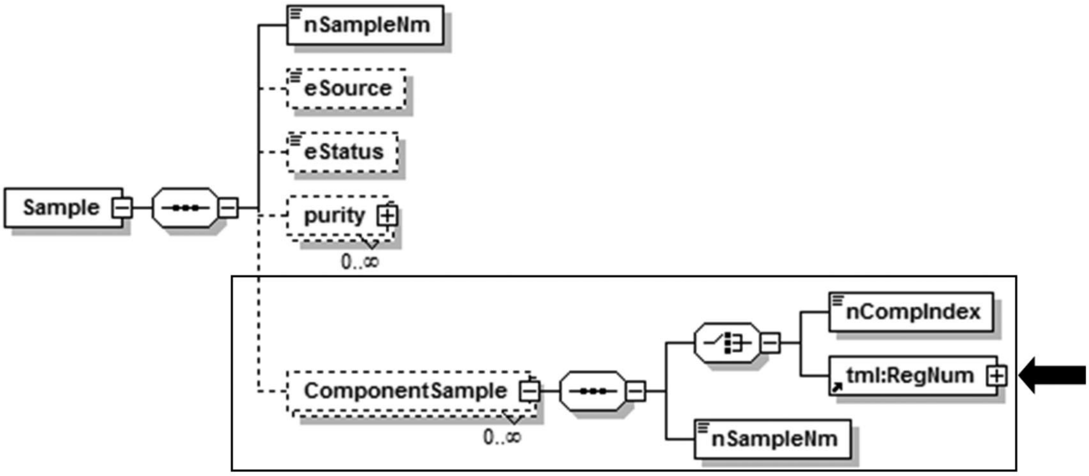

Fig. 7 Structure of the **Sample** [complex] element within the *Compound* block. The arrow indicates the location of extensions described in the text.

Structural elements related to biothermodynamic data in the *PureOrMixtureData* block

Extensions for biothermodynamic data are added in four locations within the *PureOrMixtureData* block (Fig. 3). **AuxiliarySubstance** [complex] is a new element that allows identification of substances that are part of the chemical system under consideration but are not directly associated with the particular pure-compound or mixture property. This extension allows, for example, the identification of a buffer used in a denaturation study of a protein by differential scanning calorimetry.

The location of the **AuxiliarySubstance** [complex] element in the schema is shown in Fig. 3. The **AuxiliarySubstance** [complex] element consists of three subelements (Fig. 8). As always, **RegNum** [complex] is a compound registry number that contains the further subelements **nCASNum** [numerical, integer] for the Chemical Abstracts Registry number and an identification number that may be assigned by a user organization **nOrgNum** [numerical, integer]. **RegNum** [complex] or **nCompIndex** [numerical, integer] can be used to associate the particular auxiliary substance with a compound that is fully specified within the *Compound* block (Fig. 2). **eFunction** [enumeration] represents the function of the auxiliary substance in the chemical system (Buffer, Solvent, Inert), **sFunction** [string] allows specification of a function not included in the enumeration list, and **nSampleNm** [integer] is the sam-

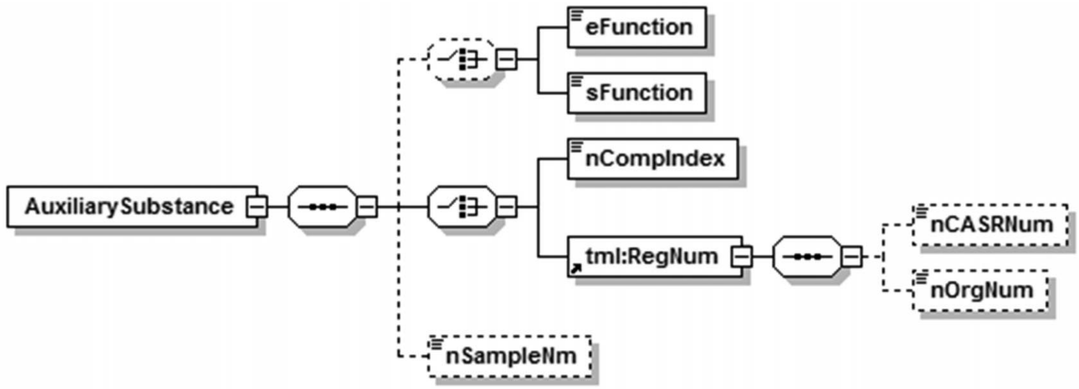

Fig. 8 Structure of the **AuxiliarySubstance** [complex] element within the *PureOrMixtureData* block.

ple number for the auxiliary substance. The existence of multiple samples for an auxiliary substance is very unlikely for any given data report; however, this element exists in numerous other analogous locations in ThermoML and is included here for consistency.

The second extension within the *PureOrMixtureData* block is within the **Property** [complex] element. This extension is used to define new properties within the ThermoML schema that are specific to biothermodynamic studies, as recommended by Hinz and Schwarz [38] for phase transitions in biological systems studied with differential scanning calorimetry. These are included as entries in an enumeration list, and further extensions can be added readily, if required. Other properties commonly reported in the biothermodynamic literature are already described within the ThermoML schema.

The **Property** [complex] element is expanded in Fig. 9, where the locations of extensions for biothermodynamic properties (**BioProperties** [complex], **eBioState** [enumeration], **sBioState** [string]) are indicated. The structure of **BioProperties** [complex], a subelement of **PropertyGroup** [complex] (upper right of Fig. 9), is shown in Fig. 10. The **ePropName** [enumeration] element allows selection of the biothermodynamic property (temperature of 1/2 conversion, K; peak temperature, K; zero Gibbs energy temperature, K; heat capacity change at transition, J K^{-1} mol^{-1}; van’t Hoff enthalpy of transition, kJ mol^{-1}). Property names are those recommended by Hinz and Schwarz [38]. The list of all properties enumerated in the **BioProperties** [complex] subelement of **PropertyGroup** [complex] is given at the top of Table 1. Methods enumerated within **eMethodName** are experimental in nature. The enumeration list for **eMethodName** [enumeration] within the **BioProperties** [complex] element is (DSC/DTA), which are the common abbreviations for differential scanning calorimetry and differential thermal analysis. **sMethodName** [string] can be used to identify other experimental methods. Methods associated with property prediction (**Prediction** [complex]) and critical evaluation (**CriticalEvaluation** [complex]) are represented separately to distinguish clearly between the three property sources: experiment, critical evaluation, and prediction. Full descriptions of these aspects of the schema were described with ThermoML-06 [15].

eBioState [enumeration] is a subelement of **PropPhaseID** [complex] (center of Fig. 9) and allows specification of states that are specific to biothermodynamic properties (Native, Denatured). For a transition from one state to another, a property is associated with two phases; the first is the initial state and the second is the final. **sBioState** [string] allows specification of other states not included in the enumeration list. For example, lipid membrane systems have been shown to exhibit a wide variety of states, including subgel, gel, and ripple states [40]. Research in this area is very active currently, and a consensus on specific terminology for the various states has not yet been established.

Table 1 Property groups that are modified with the present extensions to ThermoML[a].

Property group	Property	Unit
BioProperties	***Temperature of 1/2 conversion***	K
	Peak temperature	K
	Zero-Gibbs energy temperature	K
	Heat capacity change at transition	J K^{-1} mol^{-1}
	van’t Hoff enthalpy of transition	kJ mol^{-1}
ActivityFugacityOsmoticProp	***Mean ionic activity***	dimensionless
	Mean ionic activity coefficient	dimensionless
	Activity	dimensionless
	Activity coefficient	dimensionless
	Osmotic pressure	kPa
	Osmotic coefficient	dimensionless

(*continues on next page*)

Table 1 (*Continued*).

Property group	Property	Unit
Transport Prop	***Transport number***	dimensionless
	Viscosity	Pa s
	Kinematic viscosity	m s^{-1}
	Fluidity	Pa^{-1} s^{-1}
	Thermal conductivity	W m^{-1} K^{-1}
	Electrical conductivity	S m^{-1}
	Molar conductivity	S m^2 mol^{-1}
	Thermal diffusivity	m^2 s^{-1}
	Self-diffusion coefficient	m^2 s^{-1}
	Binary diffusion coefficient	m^2 s^{-1}
	Tracer diffusion coefficient	m^2 s^{-1}
ReactionStateChangeProp	***Potential difference of an electrochemical cell***	V
	Enthalpy of process	kJ
	Molar enthalpy of reaction	kJ mol^{-1}
	Specific internal energy of reaction at constant volume	J g^{-1}
	Molar internal energy of reaction at constant volume	kJ mol^{-1}
	Molar Gibbs energy of reaction	kJ mol^{-1}
	Molar entropy of reaction	J K^{-1} mol^{-1}
ReactionEquilibriumProp	Thermodynamic equilibrium constant	dimensionless
	Natural logarithm of the thermodynamic equilibrium constant	dimensionless
	Decadic logarithm of the thermodynamic equilibrium constant	dimensionless
	Equilibrium constant in terms of molality	(mol $kg^{-1})^n$
	Natural logarithm of the equilibrium constant in terms of molality	dimensionless
	Decadic logarithm of the equilibrium constant in terms of molality,	dimensionless
	Equilibrium constant in terms of amount concentration (molarity)	(mol $dm^{-3})^n$
	Natural logarithm of the equilibrium constant in terms of amount concentration (molarity)	dimensionless
	Decadic logarithm of the equilibrium constant in terms of amount concentration (molarity)	dimensionless
	Equilibrium constant in terms of partial p	$(kPa)^n$
	Natural logarithm of the equilibrium constant in terms of partial p	dimensionless
	Decadic logarithm of the equilibrium constant in terms of partial p	dimensionless
	Equilibrium constant in terms of mole fraction	dimensionless
	Natural logarithm of the equilibrium constant in terms of mole fraction	dimensionless
	Decadic logarithm of the equilibrium constant in terms of mole fraction	dimensionless

[a]Properties in italicized type are additions to the schema of ThermoML-06.

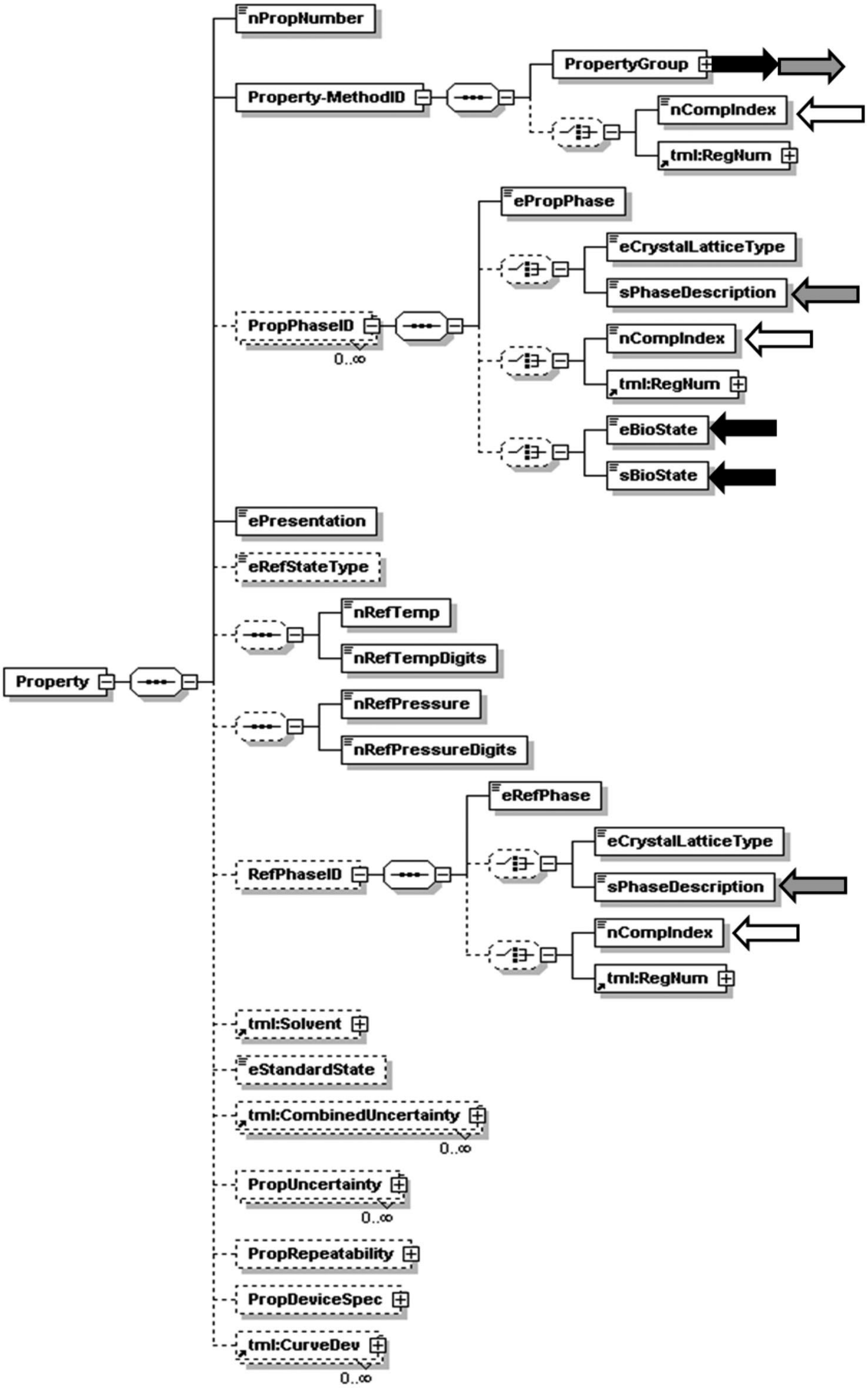

Fig. 9 Structure of the **Property** [complex] element within the *PureOrMixtureData* block. The arrows indicate the locations of extensions described in the text.

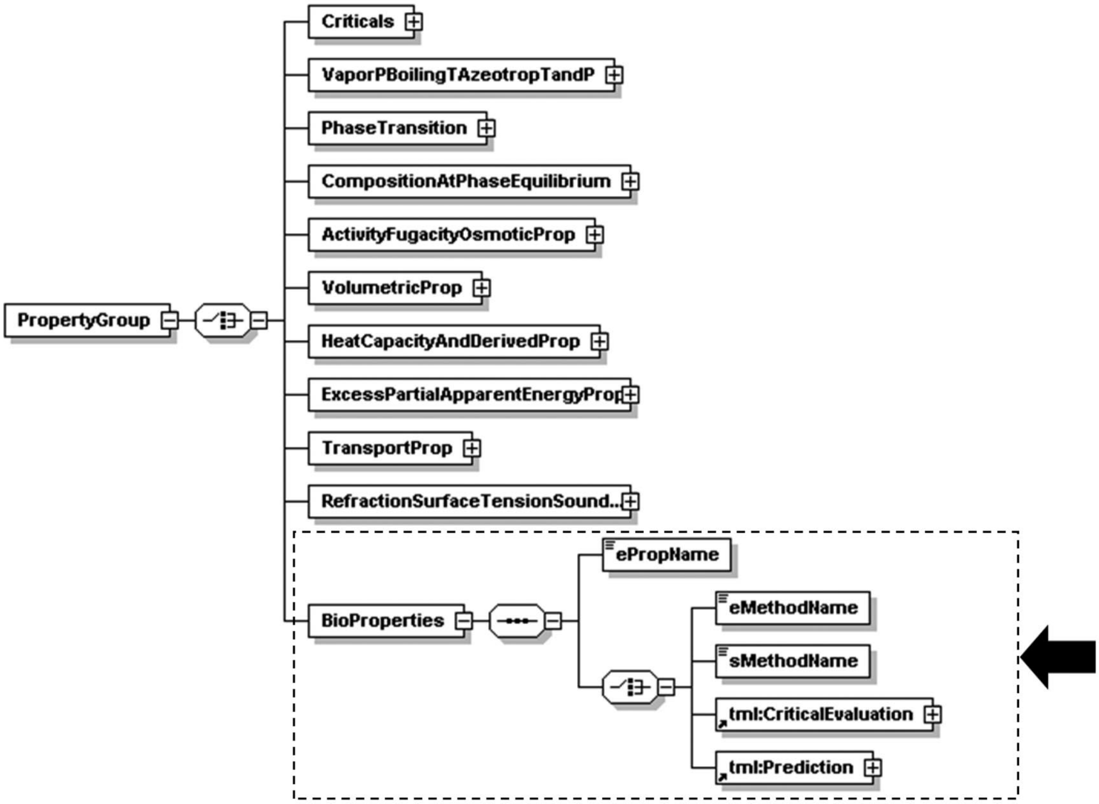
PropertyGroup
Criticals
VaporPBoilingTAzeotropTandP
PhaseTransition
CompositionAtPhaseEquilibrium
ActivityFugacityOsmoticProp
VolumetricProp
HeatCapacityAndDerivedProp
ExcessPartialApparentEnergyProp
TransportProp
RefractionSurfaceTensionSound..
BioProperties
ePropName
eMethodName
sMethodName
tml:CriticalEvaluation
tml:Prediction

Fig. 10 Structure of the **PropertyGroup** [complex] subelement of **Property** [complex] (Fig. 9) within the *PureOrMixtureData* block. The arrow indicates the location of extensions described in the text for biomaterials.

Structural elements related to biothermodynamic data in the *ReactionData* block

Extensions for biothermodynamic data are added in five locations within the *ReactionData* block (Fig. 4). The structure of the **AuxiliarySubstance** [complex] (Fig. 8) is the same as that within the *PureOrMixtureData* block. Within the *ReactionData* block, the enumerations for **eFunction** [enumeration] are specific for reactions (cofactor, buffer, inert). Catalyst and solvent specification are also important for biothermodynamic reaction data, but elements for these already exist elsewhere in the schema.

The **eReactionFormalism** [enumeration] element allows specification of the reaction formalism type (chemical, biochemical). If the *chemical formalism* is used, thermodynamic equilibrium constants depend only on temperature, and apparent equilibrium constants (i.e., those expressed in terms of sums of concentrations) further depend on the ionic strength, and thus, on the activity coefficients of the species participating in the reaction. In the *biochemical formalism*, thermodynamic equilibrium constants are, as a general rule, not measured. Instead, one measures apparent equilibrium constants, which are expressed by use of total concentrations of species in various forms of dissociation and complexation. Therefore, the apparent equilibrium constant depends on several factors, such as pH, pMg, ionic strength, etc. The number of these factors is not restricted.

New variable and constraint types for biothermodynamic data

Several variables that are commonly used in the reporting of biothermodynamic data, but were not present in ThermoML-06, are added now. The new subelement **eBioVariables** [enumeration] {pH; Ionic strength (molality basis), mol kg^{-1}; Ionic strength (amount concentration basis), mol dm^{-3}; pC (amount concentration basis); Solvent: pC (amount concentration basis)} is now included within the **Constraint** [complex] (Fig. 11) and **Variable** [complex] (Fig. 12) elements of the *PureOrMixtureData* block, as well as in the analogous elements of the *ReactionData* block (Figs. 13 and 14).

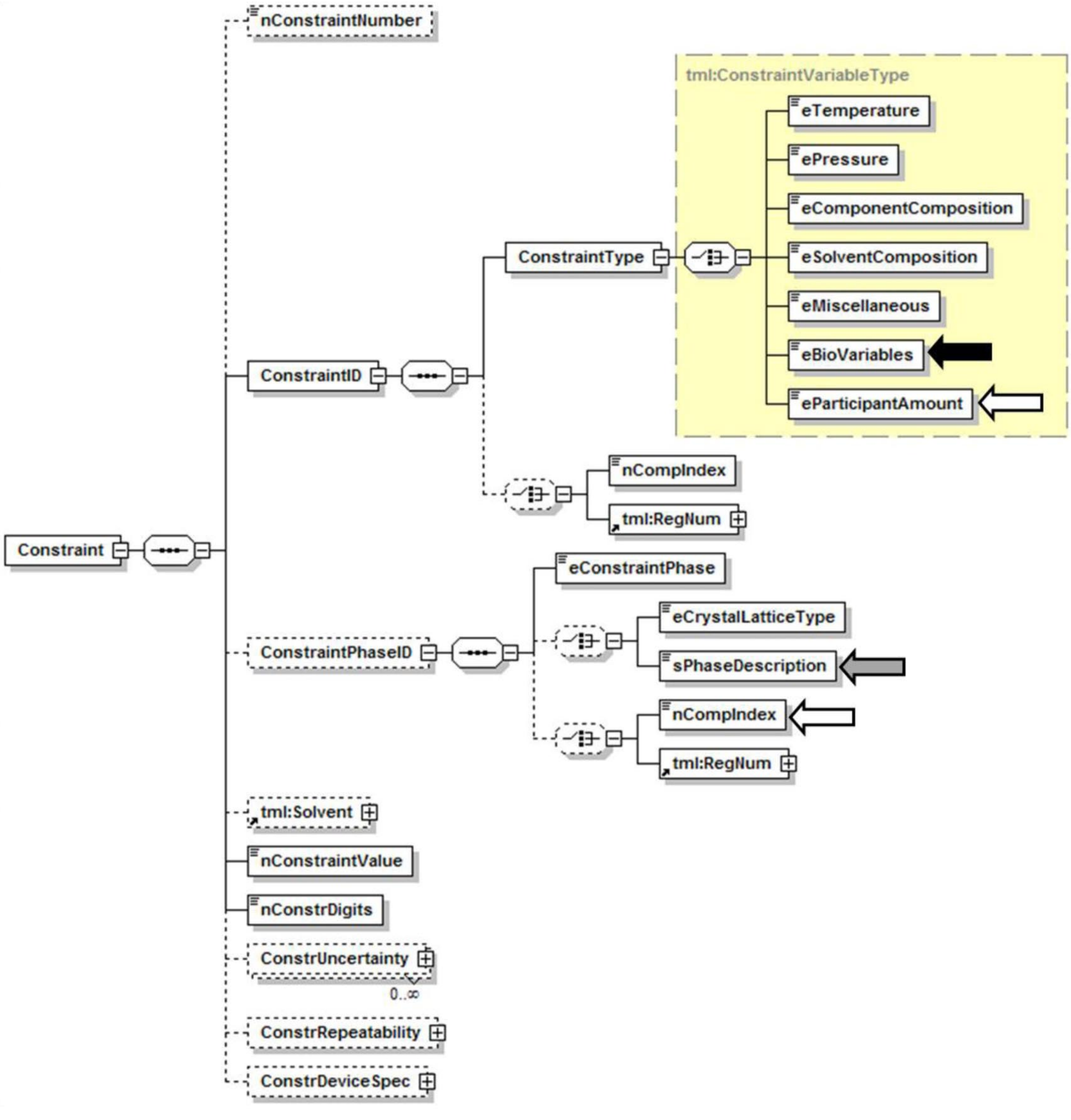

Fig. 11 Structure of the **Constraint** [complex] and **Variable** [complex] elements within the *PureOrMixtureData* block. The arrows indicate the locations of extensions described in the text.

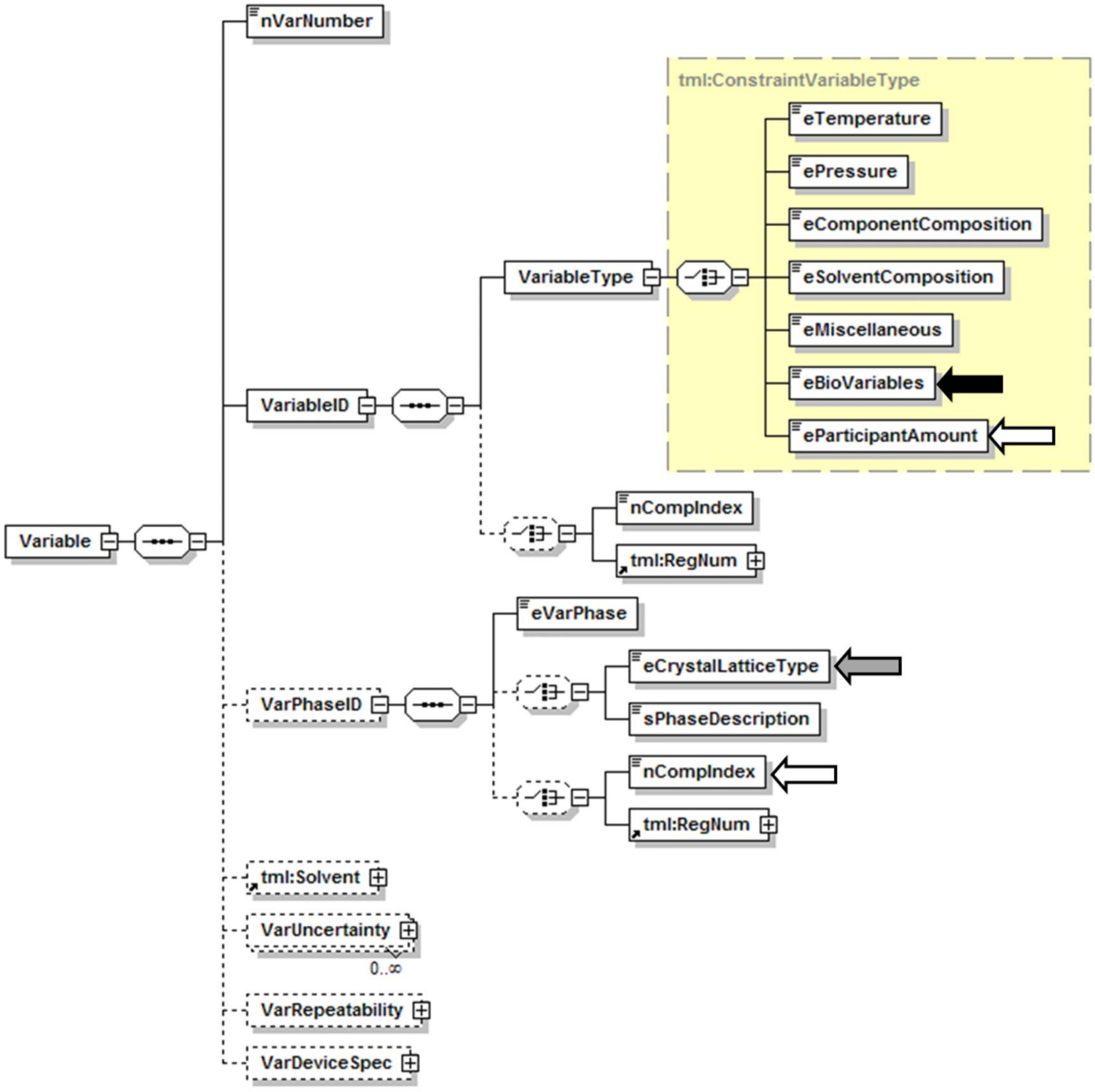

Fig. 12 Structure of the **Variable** [complex] element within the *PureOrMixtureData* block. The arrows indicate the locations of extensions described in the text.

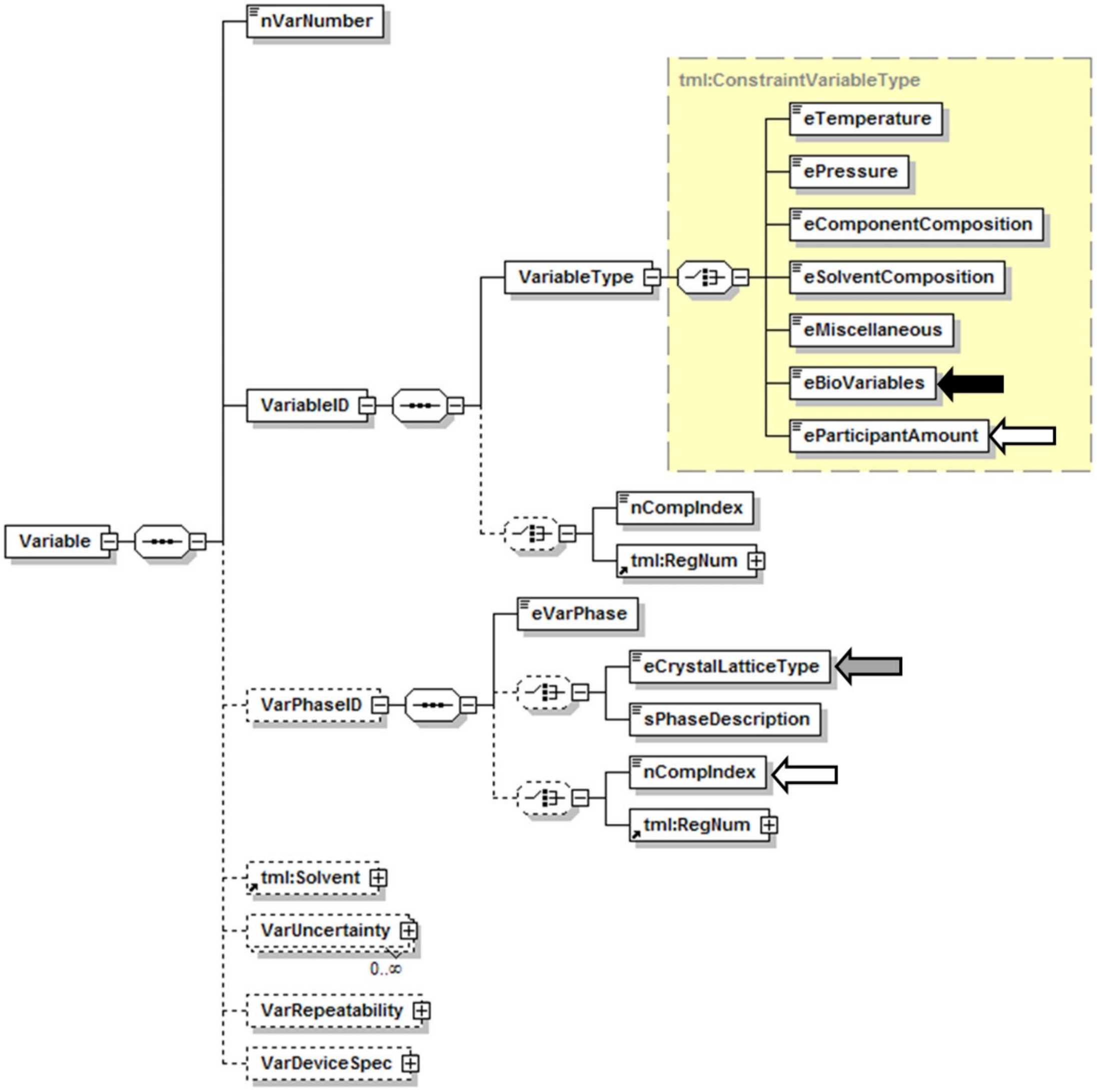

Fig. 13 Structure of the **Constraint** [complex] element within the *ReactionData* block. The arrows indicate the locations of extensions described in the text.

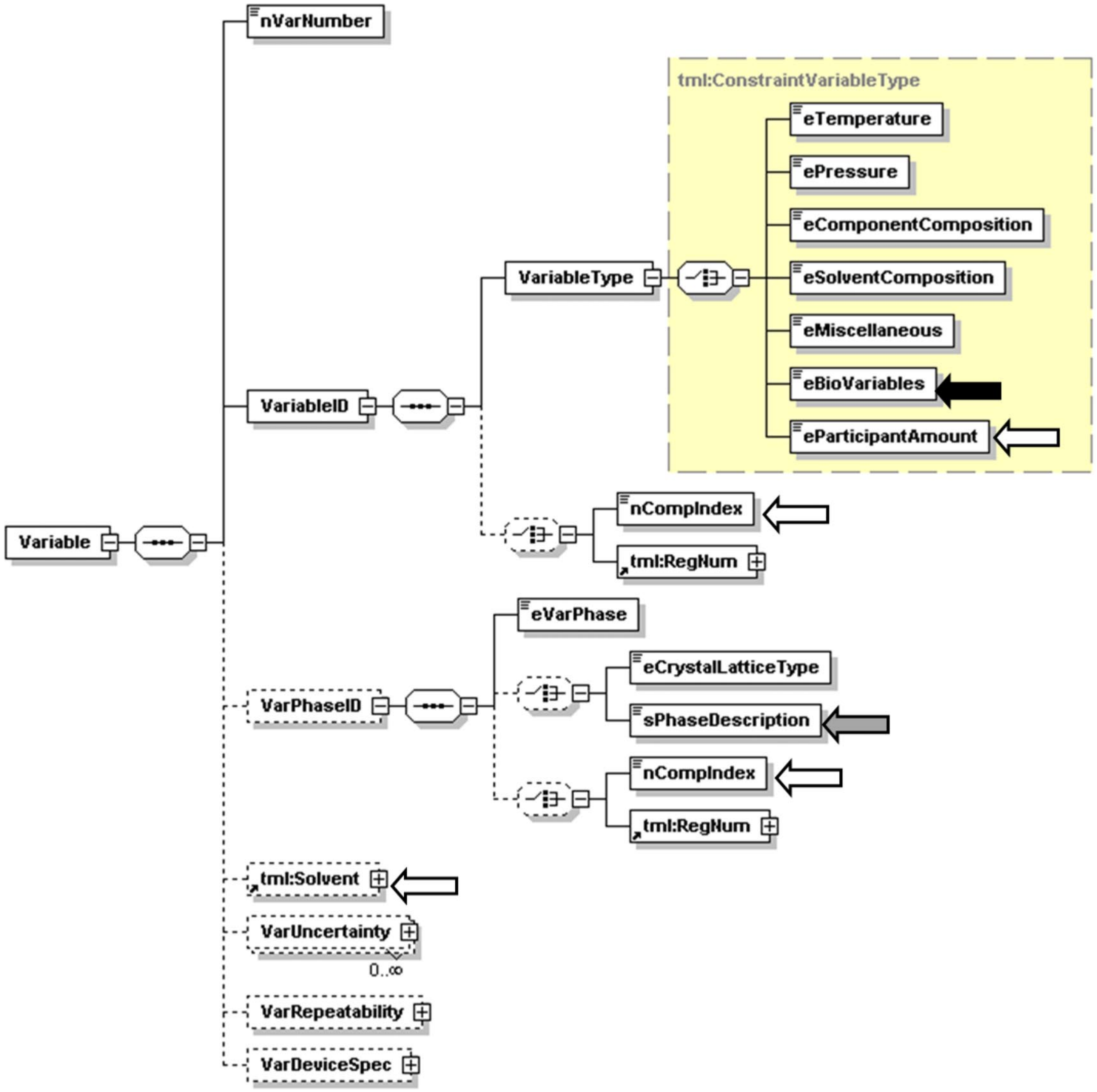

Fig. 14 Structure of the **Variable** [complex] element within the *ReactionData* block. The arrows indicate the locations of extensions described in the text.

Scope of extensions for speciation and complex equilibria

The primary goal of this work is to ensure that thermodynamic property data associated with chemical speciation and complex equilibria can be represented within the ThermoML schema. In particular, a series of extensive reviews of the thermodynamic properties for compounds and complexes of selected metals have been considered [41]. These reviews were completed under the auspices of the Organization for Economic Co-operation and Development (OECD) Nuclear Energy Agency [42]. Extensions necessary to adequately represent the information within ThermoML in these and related literature documents include additional properties, reaction specification details, reaction types, experimental methods, crystal phase specification details, and numerical representations. These may be classified into two broad categories: (1) those that require additional structural elements to be added to ThermoML, and (2) those that require extensions to existing enumeration lists that are used for such items as common reaction types (e.g., *hydrogenation* or *combustion with oxygen*) or common experimental methods (e.g., *static bomb calorimetry*). The ThermoML schema was expanded to represent properties of ions and reactions involving ions at the time of publication for ThermoML-06 [15]. Consequently, most of the schema elements necessary for representation of speciation and complex equilibria were included at that time. New elements and other extensions needed for a more complete representation of chemical speciation and complex equilibria are described in this manuscript. Examples (use cases) illustrating these new features are provided as ThermoML formatted files in the Supplementary Information.

IMPLEMENTATION OF STRUCTURAL ELEMENTS RELATED TO SPECIATION AND COMPLEX EQUILIBRIA IN ThermoML

Extensions described here are within the *PureOrMixtureData* and *ReactionData* blocks only. In all figures, the extensions related to speciation and complex equilibria are indicated with gray-filled arrows.

Structural elements for speciation and complex equilibria

For the description of new structural elements for speciation and complex equilibria, it is convenient to group the extensions by purpose, rather than by their location in the schema, as was done for new elements for biomaterials.

Crystal phase specification details

The element **eCrystalLatticeType** [enumeration] (Cubic, Tetragonal, Hexagonal, Rhombohedral, Orthorhombic, Monoclinic, Triclinic) was included in ThermoML-06 [15]. Many compounds have names of various types (geologic, colloquial, pharmaceutic, etc.) for specific crystalline forms, such as the “stishovite” form of SiO_2 [43]. Also, standard states for elements often include crystal specifications, such as “Cu(cr, cubic)” for copper, “Pu(cr, monoclinic)” for plutonium, “C(cr, graphite)” for carbon [41k]. It is not possible to enumerate a complete list of these common names, so the element **sPhaseDescription** [string] has been added.

Specification of the crystal lattice type can be necessary within many aspects of the representation of a property. Consequently, the new element **sPhaseDescription** [string] has been added in six locations within the ThermoML schema. Additions within the *PureOrMixtureProperty* block are shown in Figs. 2, 9, 11, and 12. In the *ReactionData* block (Fig. 4), the element **sPhaseDescription** [string] is a subelement of the **Participant** [complex] element (Fig. 15) and the **Variable** [complex] element (Fig. 14).

Application of **sPhaseDescription** [string] is demonstrated in Use Case 6 in the Supplementary Information, where the enthalpy of formation of 4-methylphenanthrene [44] based on the reaction

$$15\ \mathrm{C(cr,\ graphite)} + 6\ \mathrm{H_2(gas)} = \mathrm{C_{15}H_{12}(cr)} \qquad (1)$$

is represented, and the carbon crystal lattice type *graphite* is stored in the element **sPhaseDescription** [string]. Similarly, when representing the enthalpy of formation for sulfur-containing compounds [45], the standard state of sulfur at the temperature $T = 298.15$ K is often chosen to be *rhombic*, which is represented analogously.

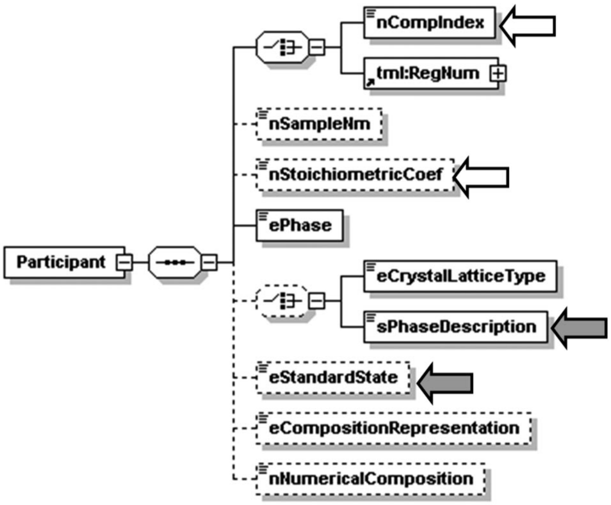

Fig. 15 Structure of the **Participant** [complex] element within the *ReactionData* block. The arrows indicate the locations of extensions described in the text.

Accommodation of multiple solvents for reaction data

It is common for thermodynamic studies that address speciation to involve measurements with electrochemical cells. A comprehensive description of electrochemical cells is outside the scope of ThermoML at this time. Nevertheless, the option to represent multiple solvents for electrochemical cells provides the means to unambiguously associate the property of interest with a particular experiment reported in a data source (journal article). The element **Solvent** [complex] was included in ThermoML-06 [15] and is now extended to include any number of solvents. This modification is within the *ReactionData* block (Fig. 4), where the notation 0…∞ below the **Solvent** [complex] element indicates that any number of solvents can be specified. Each solvent is associated with a separate liquid phase, such as *Solution 1*, *Solution 2*, etc.

Electron number of an electrochemical reaction

Specification of a reaction for an electrochemical cell includes the *electron number*, which is the number of electrons transferred in the balanced electrochemical reaction, and is a positive integer [46]. The element **nElectronNumber** [numerical, integer] is in the *ReactionData* block (Fig. 4).

Application of the element **nElectronNumber** [numerical, integer] is demonstrated in Use Case 7. This use case is based on an electrochemical study by Rakshit et al. [47] designed to determine the Gibbs energy of formation of $BaNdFeO_4$(cr) for which the electron number is 2 for the cell reaction

$$2\ BaNdFeO_4(cr) + CaF_2(cr) = CaO(cr) + BaNd_2Fe_2O_7(cr) + BaF_2(cr) \qquad (2)$$

Specification of standard state for individual reaction participants

The optional element **eStandardState** [enumeration] was included in ThermoML-06 [15], but only within the subelement **Property** [complex] of the *ReactionData* block (Fig 16), and thereby was associated with all participants of the given reaction. The element **eStandardState** [enumeration] is now included for all reaction participants (Fig. 15), with the dashed arrow in Fig. 16 indicating that the element was part of the ThermoML standard prior to the present extensions. The enumeration list for **eStandardState** [enumeration] remains unchanged: (Pure compound, Pure liquid solute, Standard molality [1 mol kg^{-1}] solute, Standard amount concentration [1 mol dm^{-3}] solute, Infinite dilution).

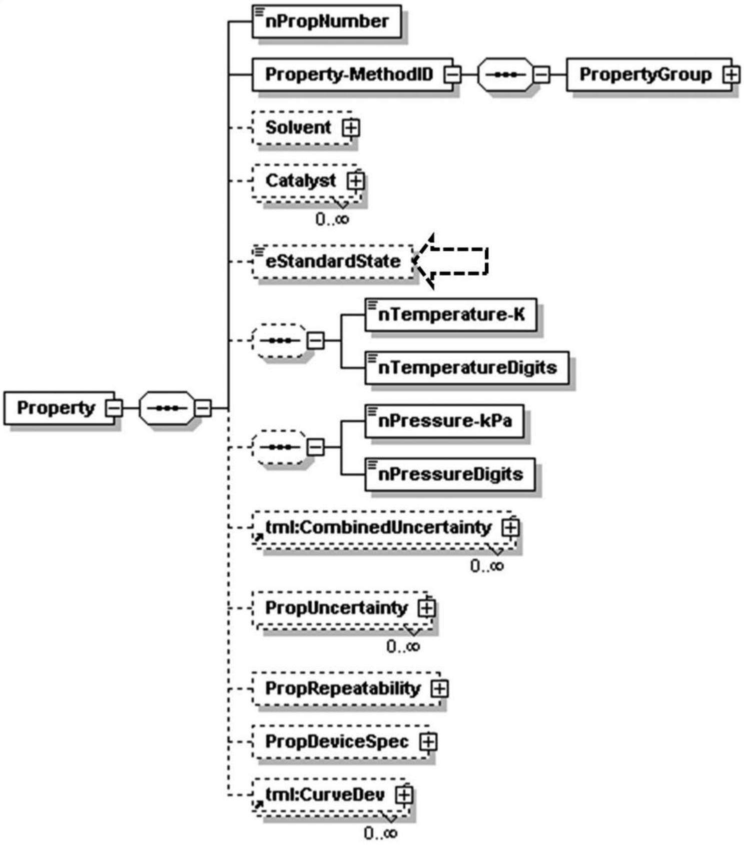

Fig. 16 Structure of the **Property** [complex] element within the *ReactionData* block. The arrow indicates the location of the **eStandardState** [enumeration] element in ThermoML-06.

The use of **eStandardState** [enumeration] is demonstrated in Use Case 8 for uniform application to all participants of a reaction (within **Property** [complex] of the *ReactionData* block; Fig. 16) and for individual assignment to each participant through representation as a subelement of **Participant** [complex] (Fig. 15). The use case is based on a study of the dissociation constants of some amines and alkanolamines by Hamborg and Versteeg [48]. Reaction 1 of the article (for piperazine) is represented using **eStandardState** [enumeration] for the reaction, whereas reaction 2 (also, for piperazine) is represented with **eStandardState** [enumeration] for each participant.

Extension of variable specification for reaction data

The optional element **VarPhaseID** [complex] was included in ThermoML-06 [15], but only within the *PureOrMixtureData* block. The element **VarPhaseID** [complex] is now included in the *ReactionData* block (Fig. 14), where its substructure has been expanded.

Extensions to enumeration lists for speciation and complex equilibria

Properties in ThermoML are categorized into 11 groups within the *PureAndMixtureProperty* block (Fig. 17) and 2 groups within the *ReactionProperty* block (Fig. 18). The **PropertyGroup** [complex] element includes the individual property groups as subelements for the *PureAndMixtureProperty* block

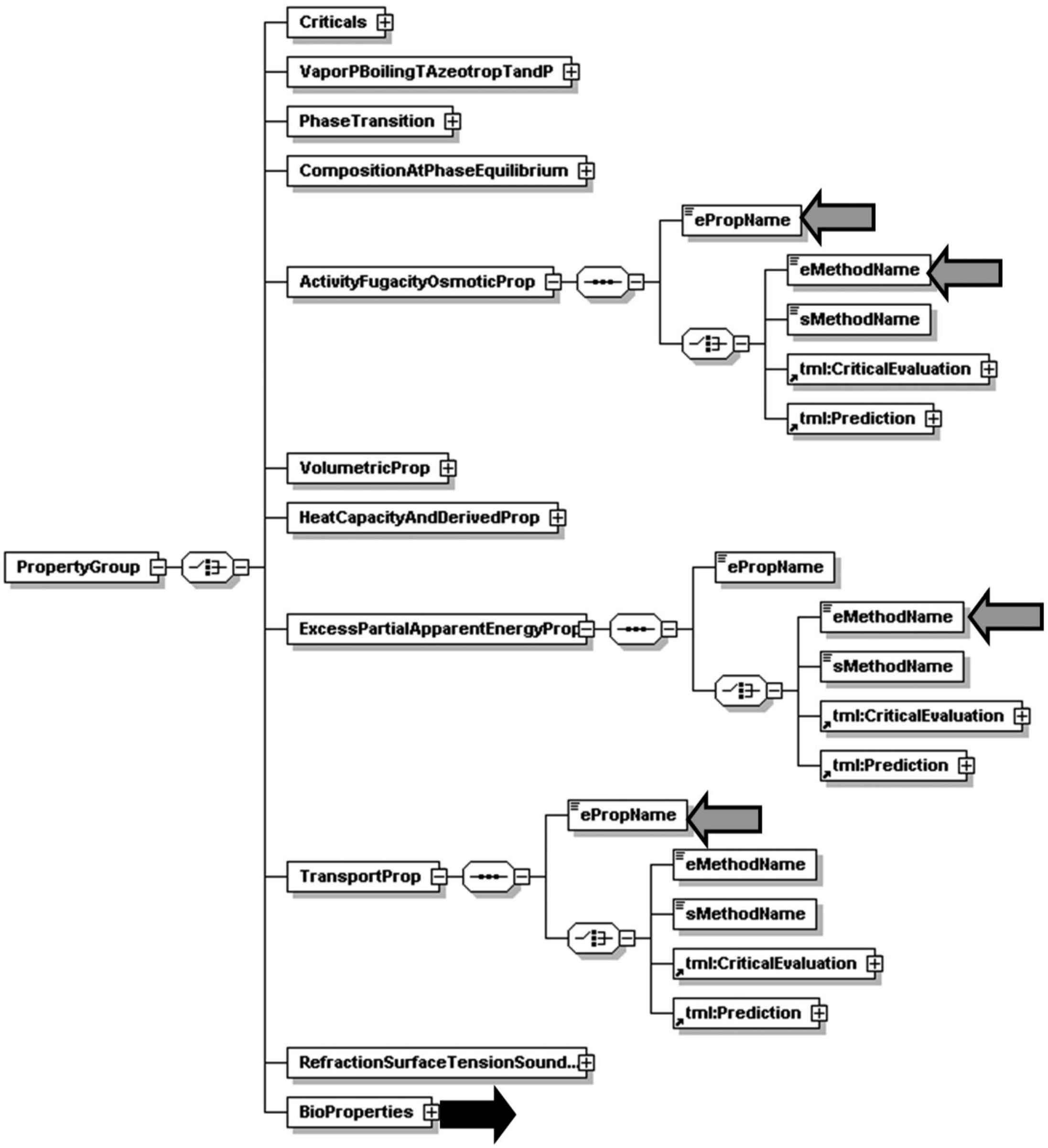

Fig. 17 Structure of the **PropertyGroup** [complex] subelement of **Property** [complex] (Fig. 9) within the *PureOrMixtureData* block. The arrows indicate the locations of extensions described in the text.

and *ReactionProperty* block. Names for specific properties are enumerated within the element **ePropName** [enumeration], which is a subelement of each property group (Figs. 17 and 18). Common experimental methods are enumerated for each property within the element **eMethodName** [enumeration]. The element **sMethodName** [string] allows storage of methods that are not enumerated. The complete list of property groups, properties, and experimental methods in ThermoML-06 were summarized in Table 1 of ref. [15]. Extensions described here for speciation and complex equilibria involve the additions of some new properties and experimental methods, but unlike the extensions for biomaterials, no new property groups were added.

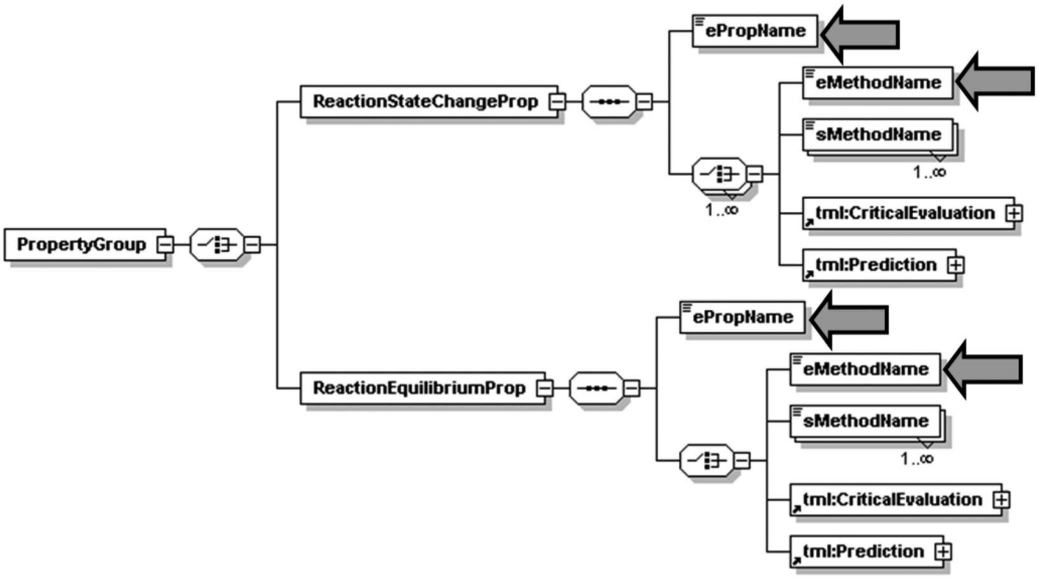

Fig. 18 Structure of the **PropertyGroup** [complex] subelement of **Property** [complex] (Fig. 16) within the *ReactionData* block. The arrows indicate the locations of extensions described in the text.

Reaction property: Potential difference of an electrochemical cell, V

Electrochemical cells are commonly used to study speciation, and the addition of this property allows direct representation of reported potential difference. These potential differences must be associated with reactions or transport processes and are included in the *ReactionStateChangeProp* group, as shown in Table 1. Representation of measured potential differences is included in Use Case 7 (Supplementary Information), which was described in the previous section.

Mixture property: Mean ionic activity and mean ionic activity coefficient

The mean ionic activity and mean ionic activity coefficient [46] are now added to the *ActivityFugacityOsmoticProp* group (Table 1). These properties were determined by Ciavatta et al. [49] for NaCl in (NaCl + water) through electrochemical measurements. Representation of these properties is demonstrated in Use Case 9 in the Supplementary Information.

Mixture property: Transport number

The transport number (formerly the transference number) [46] has been added to the TransportProp group (Table 1) of the *PureOrMixtureData* block. Transport numbers were determined for the component ions of $LaCl_3$ in aqueous solution by Longsworth and MacInnes [50]. The representation of a transport number in ThermoML for the La^{3+}(aq) ion is provided in the Supplementary Information (Use Case 10).

Logarithmic representation of equilibrium constants

Equilibrium constants can range over many orders of magnitude, and consequently, are often reported in the literature as the logarithm of the property. Numerical representation of these quantities can lead to awkward formulations with numerous zeroes, resulting in an increased likelihood of transcription errors. As equilibrium constants are at the core of speciation data, both natural logarithm and decadic logarithm representations have been added to ThermoML. ThermoML-06 [15] had five representations of equilibrium constants, and these included the thermodynamic equilibrium constant and four representations in terms of various measures of composition. Each of these can now be represented as the natural or decadic logarithm, as listed in Table 1. This extension is demonstrated in Use Case 8 with dissociation constants reported by Hamborg and Versteeg [48] in Table 10 of their publication.

Methods for properties of reactions

The extensive reviews published by the OECD Nuclear Energy Agency [41] include short descriptors for experimental methods used to determine properties associated with speciation and complex equilibria. These methods are listed in Table II.1 (page 10) of the first book in the series concerning the thermodynamics of uranium [41a]. Some of these methods were included in ThermoML-06, but many were not. New methods added to **eMethodName** [enumeration] for all reactions are: Anion exchange, Cation exchange, Colorimetry, Conductivity measurement, Coulometry, Cryoscopy, Distribution between two phases, Cell potential with glass electrode, Ion selective electrode, Rate of reaction, Molar volume determination, Polarography, Potentiometry, Proton relaxation, Cell potential with quinhydrone electrode, Cell potential with redox electrode, Spectrophotometry, Solubility measurement, Transient conductivity, Thermal lensing spectrophotometry, Solvent extraction, and Voltammetry.

Methods for properties of pure compounds and mixtures

Extensions are also now made in the enumerated methods for properties for the **ActivityFugacityOsmoticProp** [complex] and **ExcessPartialApparentEnergyProp** [complex] elements in the *PureOrMixtureData* block. The new method *Potential difference of an electrochemical cell* is added to the list for **eMethodName** [enumeration] within both of these complex elements (Fig. 17).

Enthalpy of ion formation

Values for the enthalpy of ion formation in the gas phase are most commonly calculated with measured ionization and appearance energies [51,52]. Sources for these values include the compilations by Wagman et al. (the NBS Tables) [51], Chase et al. (the "JANAF Tables") [53], and Rosenstock et al. [52]. To allow accommodation of this type of data in ThermoML, the electron is now represented in reactions as an ion with the name "electron", chemical formula "e", and charge "–1". For a reaction involving an electron, these quantities are assigned to the following elements in the *Compound* block: **sCommonName** [string], **sFormulaMolec** [string], and **nCharge** [numerical, integer].

Representations of the enthalpy of ion formation in ThermoML are demonstrated with values from the NBS Tables [51], and are included as Use Case 11 in the Supplementary Information. Formation properties of ions at infinite dilution are represented as properties of complete reactions involving H_2 and aqueous H^+ at infinite dilution, for which the reaction is

$$H_2(g) + S(cr, rhombic) + 2\ O_2(g) = 2\ H^+(aq, \infty) + SO_4^{2-}\ (aq, \infty) \tag{3}$$

The following reaction for formation of an ion in the gas phase is also included in Use Case 11:

$$S(cr, rhombic) = S^+(g) + e^-(g) \tag{4}$$

IMPLEMENTATION OF MISCELLANEOUS NEW STRUCTURAL ELEMENTS AND SCHEMA FEATURES

In this section, extensions to the schema are described that are not necessarily associated with biomaterials, speciation, or complex equilibria.

ThermoML namespace

In the cooperation between NIST and five journals, described in the introduction, the data files provided in the ThermoML Archive [21] were not linked to the namespace established on the IUPAC web site [14], as noted by Nic in his review of chemical XML formatting [54]. With the present release of ThermoML (version 4.0), all of the data files in the ThermoML Archive are now modified with this link. Use of this link also allows easy import of the ThermoML schema into other XML constructs. All XML files in the ThermoML Archive were validated with the new schema by replacement of the text

```
<DataReport xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="http://trc.nist.gov/ThermoML.xsd">
```

with

```
<DataReport xmlns="http://www.iupac.org/namespaces/ThermoML"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.iupac.org/namespaces/ThermoML ThermoML.xsd">.
```

This global change should be applied to all XML files downloaded from the ThermoML Archive and will allow validation of the old files against the new schema (version 4.0).

Extensions in the *Compound* block for improved substance specification

The element **RegNum** [complex] was included in ThermoML-06, with subelements for representation of the Chemical Abstracts Service Registry Number, **nCASRNum** [numerical, integer], and an identification number assigned by a user organization **nOrgNum** [numerical, integer] (Fig. 2). The element **sOrganization** [string] is added with the present extensions to allow identification of the organization that has assigned the value to **nOrgNum** [numerical, integer].

In ThermoML-06, the value for **RegNum** [complex] (i.e., either **nOrgNum** [numerical, integer] or **nCASRNum** [numerical, integer]) was used as the primary identifier for a particular chemical throughout a ThermoML data file. Because of this, the element **RegNum** [complex] was mandatory in ThermoML-06. In practice, we have found this to be unnecessarily restrictive. In the new schema described here, **RegNum** [complex] is now optional, and the new element **nCompIndex** [numerical, integer] is added (Fig. 2). This new element provides a mechanism to assign to each compound in a **Data Report** (Fig. 1), a separate integer that can be used within the ThermoML data file to link compound identities to their role in the data file (e.g., solvent, catalyst, subject of property measurement, reaction participant, etc.). To accomplish this, **nCompIndex** [numerical, integer] is now added in 22 locations in the schema, where only **RegNum** [complex] had appeared previously (cf. Figs. 2, 3, 9, 11–15). Both **RegNum** [complex] and **nCompIndex** [numerical, integer] are optional, which provides additional flexibility for representation and ensures that files formatted with ThermoML-06 remain valid under the new schema with minimal modification.

Since establishment of the ThermoML IUPAC standard in 2006, several new compound identifiers (PubChem "Compound ID's" [55], InChI strings, and InChIKeys [56]) have come into common usage by the scientific community. The element **nPubChemID** [numerical, integer] is now added to accommodate the PubChem Chemical ID numbers (Fig. 2). ThermoML-06 included an element for the representation of the IUPAC International Chemical Identifier, the InChI string [57]. In June 2010, IUPAC and the InChI Trust [58] announced the release of version 1.03 of the software that integrates the generation of the standard InChI string and non-standard, customized strings as well as the corresponding InChIKeys [56]. The elements **sStandardInChI** [string] for representation of the standard InChI string and **sStandardInChIKey** [string] for representation of the standard InChIKey, are now added (Fig. 2). Non-standard InChI strings and corresponding InChIKeys are not supported explicitly in ThermoML because this would add unnecessary complexity; however, these can be accommodated in the new element **sOrgID** [complex] (Fig. 2).

The elements **sOrgID** [complex] with subelements **sOrgIdentifier** [string] and **sOrganization** [string] are now added to accommodate all types of alphanumeric compound identifiers. The notation 0…∞ below the element **sOrgID** [complex] in Fig. 2 indicates that any number of these can be represented in a single *Data Report*. In ThermoML-06, the element **nOrgNum** [numerical, integer] was included, but the limitation to integer numbers was too restrictive.

Specification of an equilibrium mixture of species and single subspecies

Although unusual, properties can be reported for a particular species within an equilibrium mixture of species. For example, Kuznetsov et al. [59] reported partial pressures, determined by mass spectrometry, for individual species present in the saturated vapor above holmium trichoride. The element **eSpeciationState** [enumeration] is a new subelement in the *Compound* block (Fig. 2) that allows association of property values with the named species only or an equilibrium mixture of all subspecies or associated species of the named compound. The enumerations are *equilibrium* and *single species*. Some property data for single species reported by Kuznetsov et al. [59] are the subject of Use Case 13.

Improved substance specification for polymers in the *Compound* block

ThermoML-06 was established prior to publication of the most recent IUPAC recommendations for polymer terminology [27,28]. Consequently, important terms now recommended for polymers were not included. The element **polymer** [complex] is a subelement of the *Compound* block (Fig. 2), and its substructure is shown in Fig. 19. All previous subelements are now replaced to conform with the IUPAC recommendations [27,28,60]. The subelements **nNumberAvgMolMass** [numerical, floating], **nPeakAvgMolMass** [numerical, floating], **nViscosityAvgMolMass** [numerical, floating],

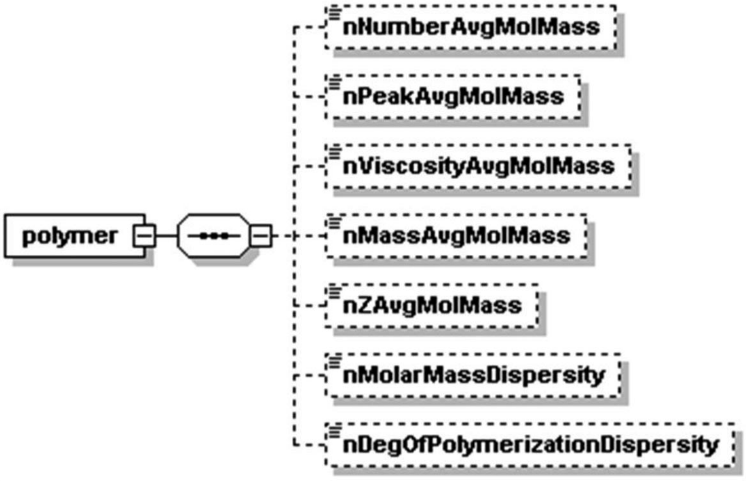

Fig. 19 Structure of the **polymer** [complex] element of the *Compound* block.

nMassAvgMolMass [numerical, floating], **nZAvgMolMass** [numerical, floating], **nMolarMassDispersity** [numerical, floating], and **nDegOfPolymerizationDispersity** [numerical, floating] are used, respectively, to represent the number average molar mass, the peak average molar mass, the viscosity average molar mass, the mass average molar mass, the Z-average molar mass, the molar mass dispersity, and the degree-of-polymerization dispersity.

Representation of enthalpies for the process of mixing solutions with solvent components in common

Another relatively uncommon type of property data involves measurement of enthalpies associated with the process of mixing, for example, (compound *A* + water) with (compound *B* + water) to yield (compound *A* + compound *B* + water). The study by Muñoz de Miguel et al. [61] of the enthalpies associated with the mixing of alkylureas with electrolytes in water is an example of such a study. Representation of these data in ThermoML required several additions and modifications to the schema: (1) the phase *Solution 4* was added to allow specification of all solutions to be mixed; (2) the element **Solvent** [complex] was added as a subelement to **Variable** [complex] in the **ReactionData** block (Fig. 14 lower left); (3) the subelement **eParticipantAmount** [enumeration] (amount, mol; mass, kg) was added to VariableType [complex] to specify how the quantity of the solutions are specified (Fig. 14, upper right); (4) the subelement of **Participant** [complex], **nStoichiometricCoef** [numerical, floating], in the *ReactionData* block is now optional (Fig. 15) to allow specification of participant amounts in terms of moles or mass (item 3); and (5) the property, *Enthalpy of process, kJ*, is now added to the enumeration list for **ePropName** [enumeration] within the **ReactionChangeStateProp** [complex] element of **PropertyGroup** [complex] in the *ReactionData* block (Fig. 18). The enthalpy is associated with the amount of the reaction participants represented in **eParticipantAmount** [enumeration] or **nStoichiometricCoef** [numerical, floating]. Representation of the experimental data reported by Muñoz de Miguel et al. [61] is provided as a use case in the Supplementary Information for this report (Use Case 14).

SUPPLEMENTARY INFORMATION

Use cases for representation of biothermodynamic data with ThermoML

Examples of files containing biothermodynamic data were created with the ThermoML formats and are included as Supplementary Information available online (doi:10.1351/PAC-REC-11-05-01). These use cases are based upon studies published in the peer-reviewed literature.

Use Case 1 involves experimental data for lysozyme unfolding in a differential scanning calorimeter [62]. Use Case 2 demonstrates representation of an enthalpy determination for an enzyme-catalyzed reaction [63]. Use Case 3 shows representation of binding constants determined with isothermal titration calorimetry [64]. Use Case 4 demonstrates representation of solubility data for some amino acids in salt solutions with various values of pH [65]. Use Case 5 involves data related to lipid polymorphism in a complex medium [66]. The experimental data were taken from the LIPIDAT.2 relational database of thermodynamic and associated information on lipid mesophase and crystal polymorphic transitions (LIPIDAT ID #10698). This database is freely available online [67].

Use cases for representation of speciation and complex equilibria

Six use cases are provided for demonstration of extensions for speciation and complex equilibria in the Supplementary Information. Use Case 6 demonstrates the element **sPhaseDescription** [string]. The example involves the enthalpy of formation for a hydrocarbon [44]. Use Case 7 demonstrates the element **nElectronNumber** [numerical, integer]. This use case includes an example of the reaction property *Potential difference of an electrochemical cell, V*. The use case is based on an electrochemical study designed to determine the Gibbs energy of formation of an inorganic compound [47]. Use Case 8 demonstrates the element **eStandardState** [enumeration] for all participants of a reaction at once and for each participant individually. It includes an example of the logarithmic representation of an equilibrium constant in ThermoML. This use case is based on a study of the dissociation constants of some amines and alkanolamines [48]. Use Case 9 provides examples of representation in ThermoML of the mixture properties *mean ionic activity* and *mean ionic activity coefficient*. The use case is based on a study of the (NaCl + water) system [49]. Use Case 10 provides an example of representation for the enthalpy of ion formation at infinite dilution and for a gaseous ion. This use case demonstrates representation of values from the NBS tables [51]. Use Case 11 provides an example of representation of transport numbers for an ion in solution. The particular chemical system used in the example is aqueous $LaCl_3$ [50].

Use cases for miscellaneous extensions

Use Case 12 demonstrates use of **eSpeciation** [enumeration] (Fig. 2). The example represents property data for the hypothetical ethanoic acid monomer, dimer, and equilibrium mixture in the gas phase reported by Chao et al. [68]. Use Case 13 demonstrates representation of a property for a single species within an equilibrium mixture. The example is based on partial pressures for various species of holmium chloride reported by Kuznetsov et al. [59]. Use Case 14 demonstrates representation of enthalpies for the process of mixing solutions with common solvent components. Enthalpies reported by Muñoz de Miguel et al. [61] are the basis for the example.

Text of the ThermoML schema

The complete text of the ThermoML schema (in text format) with all extensions described here, is included here as Supplementary Information or can be obtained free of charge through direct request to the corresponding author (M.F.). This material is available free of charge via the Internet at <http://pubs.acs.org>. The schema can also be downloaded from the ThermoML namespace located on an IUPAC web site [14].

MEMBERSHIP OF SPONSORING BODY

Membership of the Committee on Printed and Electronic Publications during the final preparation of this report (2011) was as follows:

***Chair*:** D. Martinsen (USA); ***Secretary*:** R. J. Lancashire (Jamaica); ***Members*:** S. M. Bachrach (USA), C. Batchelor (UK), R. Deplanque (Germany), B. Lawlor (USA), M. Nic (Czech Republic), C. Steinbeck (UK); ***Ex Officio*:** J. R. Bull (South Africa).

ACKNOWLEDGMENTS

The authors express their appreciation to the IUPAC Committee on Printed and Electronic Publications (CPEP) for providing travel funds for this project: 2007-039-1-024, “Extension of ThermoML—the IUPAC Standard for Thermodynamic Data Communications”. For M.F., R.D.C., V.D., R.N.G., and S.E.S., this work represents an official contribution of the U.S. National Institute of Standards and Technology and is not subject to copyright in the United States.

REFERENCES

1. M. Frenkel. *Pure Appl. Chem.* **77**, 1349 (2005).
2. M. Frenkel. *J. Chem. Eng. Data* **54**, 2411 (2009).
3. M. Frenkel. *J. Chem. Thermodyn.* **39**, 169 (2007).
4. M. Frenkel. *Comp. Chem. Eng.* **35**, 393 (2011).
5. R. C. Wilhoit, 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).
6. <http://www-i5.informatik.rwth-aachen.de/lehrstuhl/projects/gco/>. Web page contact: Alexander Kuckelberg (kuckelberg@informatik.rwth-aachen.de).
7. IUPAC project 024/1/99, "Standardization of physico-chemical property electronic data files", H. V. Kehiaian <http://www.iupac.org/web/ins/024-1-99>.
8. A. K. Dewan, D. L. Embry, T. J. Willman. DIPPR/AIChE Project 991 – Thermophysical Property Data Exchange, *Book of Abstracts of the 14th Symposium on Thermophysical Properties*, p. 169, Boulder, CO (2000).
9. IUPAC project 2002-055-3-024, "XML-based IUPAC Standard for Experimental and Critically Evaluated Thermodynamic Property Data Storage and Capture", M. Frenkel <http://www.iupac.org/projects/2002/2002-055-3-024.html>.
10. IUPAC Committee on Printed and Electronic Publication, CPEP <http://www.iupac.org/web/ins/024>.
11. C. Finkelstein, P. Aiken. *Building Corporate Portals with XML*, McGraw-Hill, New York (1999).
12. IBM XML Toolkit, <http://www-03.ibm.com/systems/z/os/zos/tools/xml/>.
13. Microsoft XML Downloads, <http://msdn.microsoft.com/en-us/data/bb190600.aspx>.
14. ThermoML - namespace for the XML-based IUPAC Standard for Thermodynamic Property Data <http://www.iupac.org/namespaces/ThermoML>.
15. M. Frenkel, R. D. Chirico, V. V. Diky, K. N. Marsh, J. H. Dymond, W. A. Wakeham, S. E. Stein, E. Königsberger, A. R. H. Goodwin. *Pure Appl. Chem.* **78**, 541 (2006).
16. 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).
17. R. D. Chirico, M. Frenkel, V. V. Diky, K. N. Marsh, R. C. Wilhoit. *J. Chem. Eng. Data* **48**, 1344 (2003).
18. M. Frenkel, R. D. Chirico, V. V. Diky, K. N. Marsh, J. H. Dymond, W. A. Wakeham. *J. Chem. Eng. Data* **49**, 381 (2004).
19. M. Frenkel, R. D. Chirico, V. Diky, C. Muzny, Q. Dong, K. N. Marsh, J. H. Dymond, W. A. Wakeham, S. E. Stein, E. Königsberger, A. R. H. Goodwin, J. W. Magee, M. Thijssen, W. M. Haynes, S. Watanasiri, M. Satyro, M. Schmidt, A. I. Johns, G. R. Hardin. *J. Chem. Inf. Model.* **46**, 2487 (2006).
20. (a) P. T. Cummings, T. de Loos, J. P. O'Connell, W. M. Haynes, D. G. Friend, A. Mandelis, K. N. Marsh, P. L. Brown, R. D. Chirico, A. R. H. Goodwin, J. Wu, R. D. Weir, J. P. M. Trusler, A. Pádua, V. Rives, C. Schick, S. Vyazovkin, L. D. Hansen. *Fluid Phase Equilibr.* **276**, 165 (2009); (b) P. T. Cummings, T. de Loos, J. P. O'Connell, W. M. Haynes, D. G. Friend, A. Mandelis, K. N. Marsh, P. L. Brown, R. D. Chirico, A. R. H. Goodwin, J. Wu, R. D. Weir, J. P. M. Trusler, A. Pádua, V. Rives, C. Schick, S. Vyazovkin, L. D. Hansen. *Int. J. Thermophys.* **30**, 371 (2009); (c) P. T. Cummings, T. de Loos, J. P. O'Connell, W. M. Haynes, D. G. Friend, A. Mandelis, K. N. Marsh, P. L. Brown, R. D. Chirico, A. R. H. Goodwin, J. Wu, R. D. Weir, J. P. M. Trusler, A. Pádua, V. Rives, C. Schick, S. Vyazovkin, L. D. Hansen. *J. Chem. Eng. Data* **54**, 2 (2009); (d) P. T. Cummings, T. de Loos, J. P. O'Connell, W. M. Haynes, D. G. Friend, A. Mandelis, K. N. Marsh, P. L. Brown, R. D. Chirico, A. R. H. Goodwin, J. Wu, R. D. Weir, J. P. M. Trusler, A. Pádua, V. Rives, C. Schick, S. Vyazovkin, L. D. Hansen. *J. Chem. Thermodyn.*

41, 575 (2009); (e) P. T. Cummings, T. de Loos, J. P. O'Connell, W. M. Haynes, D. G. Friend, A. Mandelis, K. N. Marsh, P. L. Brown, R. D. Chirico, A. R. H. Goodwin, J. Wu, R. D. Weir, J. P. M. Trusler, A. Pádua, V. Rives, C. Schick, S. Vyazovkin, L. D. Hansen. *Thermochim. Acta* **484**, vii (2008).

21. ThermoML Web Archive, <http://trc.nist.gov/ThermoML.html>.
22. IUPAC project 2007-039-1-024, "Extension of ThermoML: The IUPAC Standard for Thermodynamic Data Communications", M. Frenkel <http://www.iupac.org/web/ins/2007-039-1-024>.
23. R. D. Chirico, M. Frenkel, V. V. Diky, R. N. Goldberg, H. Heerklotz, J. E. Ladbury, D. P. Remeta, J. H. Dymond, A. R. H. Goodwin, K. N. Marsh, W. A. Wakeham. *J. Chem. Eng. Data* **55**, 1564 (2010).
24. M. Frenkel, V. Diky, R. D. Chirico, R. N. Goldberg, H. Heerklotz, J. E. Ladbury, D. P. Remeta, J. H. Dymond, A. R. H. Goodwin, K. N. Marsh, W. A. Wakeham, S. E. Stein, P. L. Brown, E. Königsberger, P. A. Williams. *J. Chem. Eng. Data* **56**, 307 (2011).
25. *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.
26. *U.S. Guide to the Expression of Uncertainty in Measurement*, ANSI/NCSL Z540-2-1997, NCSL International, Boulder, CO (1997).
27. R. G. Gilbert, M. Hess, A. D. Jenkins, R. G. Jones, P. Kratochvíl, R. F. T. Stepto. *Pure Appl. Chem.* **81**, 351 (2009).
28. IUPAC. *Compendium of Polymer Terminology and Nomenclature*, IUPAC Recommendations 2008 (the "Purple Book"). Edited by R. G. Jones, J. Kahovec, R. Stepto, E. S. Wilks, M. Hess, T. Kitayama, W. V. Metanomski, RSC Publishing, Cambridge, UK (2008); Chap. 3.
29. B. N. Taylor, A. Thompson. *The International System of Units (SI)*, NIST Special Publication 330, National Institute of Standards and Technology, Washington, DC (2008).
30. A. Thompson, B. N. Taylor. *Guide for the Use of the International System of Units (SI)*, NIST Special Publication 811, National Institute of Standards and Technology, Washington, DC (2008).
31. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). *Enzyme Nomenclature. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions They Catalyse.* <http://www.chem.qmul.ac.uk/iubmb/enzyme/>.
32. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. *Nucleic Acids Res.* **28**, 235 (2000).
33. (a) R. N. Goldberg, Y. B. Tewari, T. N. Bhat. *Bioinformatics* **20**, 2874 (2004); (b) NIST Standard Reference Database 74, <http://xpdb.nist.gov/enzyme_thermodynamics>.
34. (a) R. A. Alberty, A. Cornish-Bowden, Q. H. Gibson, R. N. Goldberg, G. Hammes, W. Jencks, K. F. Tipton, R. Veech, H. V. Westerhoff, E. C. Webb. *Pure Appl. Chem.* **66**, 1641 (1994); (b) R. A. Alberty, A. Cornish-Bowden, Q. H. Gibson, R. N. Goldberg, G. Hammes, W. Jencks, K. F. Tipton, R. Veech, H. V. Westerhoff, E. C. Webb. *Eur. J. Biochem.* **240**, 1 (1996).
35. R. A. Alberty, A. Cornish-Bowden, R. N. Goldberg, G. G. Hammes, K. Tipton, H. V. Westerhoff. *Biophys. Chem.* **155**, 89 (2011).
36. (a) I. Wadsö, H. Gutfreund, P. Privalov, J. T. Edsall, W. P. Jencks, G. T. Strong, R. L. Biltonen. *J. Biol. Chem.* **251**, 6879 (1976); (b) I. Wadsö, H. Gutfreund, P. Privalov, J. T. Edsall, W. P. Jencks, G. T. Strong, R. L. Biltonen. *Q. Rev. Biophys.* **9**, 439 (1976).

37. I. Wadsö, R. L. Biltonen. *Eur. J. Biochem.* **153**, 429 (1985).
38. H.-J. Hinz, F. P. Schwarz. *Pure Appl. Chem.* **73**, 745 (2001).
39. *XML SPY v. 4.4 u.* ALTOVA GmbH and ALTOVA, Inc., 1998–2002.
40. H. Heerklotz. *J. Phys.: Condens. Matter* **16**, R441 (2004).
41. (a) I. Grenthe, J. Fuger, R. J. M. Konings, R. J. Lemire, A. B. Muller, C. Nguyen-Trung, H. Wanner. *Chemical Thermodynamics, Vol. 1, Chemical Thermodynamics of Uranium*, OECD Nuclear Energy Data Bank (Eds.), North Holland Elsevier Science, Amsterdam (1992); (b) R. J. Silva, G. Bidoglio, P. B. Robouch, I. Puigdomènech, H. Wanner, M. H. Rand. *Chemical Thermodynamics, Vol. 2, Chemical Thermodynamics of Americium*, OECD Nuclear Energy Data Bank (Eds.), North Holland Elsevier Science, Amsterdam (1995); (c) J. A. Rard, M. H. Rand, G. Anderegg, H. Wanner. *Chemical Thermodynamics, Vol. 3, Chemical Thermodynamics of Technetium*, OECD Nuclear Energy Data Bank (Eds.), North Holland Elsevier Science, Amsterdam (1999); (d) R. J. Lemire, J. Fuger, H. Nitsche, P. Potter, M. H. Rand, J. Rydberg, K. Spahiu, J. C. Sullivan, W. J. Ullman, P. Vitorge, H. Wanner. *Chemical Thermodynamics, Vol. 4, Chemical Thermodynamics of Neptunium and Plutonium*, OECD Nuclear Energy Data Bank (Eds.), North Holland Elsevier Science, Amsterdam, Netherlands (2001); (e) R. Guillaumont, T. Fanghänel, J. Fuger, I. Grenthe, V. Neck, D. A. Palmer, M. H. Rand. *Chemical Thermodynamics, Vol. 5, Update on the Chemical Thermodynamics of Uranium, Neptunium, Plutonium, Americium, and Technitium*, OECD Nuclear Energy Data Bank (Eds.), North Holland Elsevier Science, Amsterdam, (2003); (f) H. Gamsjäger, J. Bugajski, T. Gajda, R. J. Lemire, W. Preis. *Chemical Thermodynamics, Vol. 6, Chemical Thermodynamics of Nickel*, OECD Nuclear Energy Data Bank (Eds.), North Holland Elsevier Science, Amsterdam (2005); (g) A. Olin, B. Nolang, E. G. Osadchii, L.-O. Öhman, E. Rosen. *Chemical Thermodynamics, Vol. 7, Chemical Thermodynamics of Selenium*, OECD Nuclear Energy Data Bank (Eds.), North Holland Elsevier Science, Amsterdam (2005); (h) P. L. Brown, E. Curti, B. Grambow. *Chemical Thermodynamics, Vol. 8, Chemical Thermodynamics of Zirconium*, OECD Nuclear Energy Data Bank (Eds.), North Holland Elsevier Science, Amsterdam (2005); (i) W. Hummel, G. Anderegg, I. Puigdomènech, L. Rao, O. Tochiyama. *Chemical Thermodynamics, Vol. 9, Chemical Thermodynamics of Compounds and Complexes of U, Np, Pu, Am, Tc, Se, Ni and Zr With Selected Organic Ligands*, OECD Nuclear Energy Data Bank (Eds.), North Holland Elsevier Science, Amsterdam (2005); (j) J. Bruno, D. Bosbach, D. Kulik, A. Navrotsky. *Chemical Thermodynamics, Vol. 10, Chemical Thermodynamics of Solid Solutions of Interest in Radioactive Waste Management*, OECD Nuclear Energy Data Bank (Eds.), OECD Publications, Paris (2007); (k) M. Rand, J. Fuger, I. Grenthe, V. Neck, D. Rai. *Chemical Thermodynamics, Vol. 11, Chemical Thermodynamics of Thorium*, OECD Nuclear Energy Data Bank (Eds.), OECD Publications, Paris (2007).
42. The Organisation for Economic Co-operation and Development (OECD) <http://www.oecd.org>.
43. N. A. Dubrovinskaia, L. S. Dubrovinsky. *Mater. Chem. Phys.* **68**, 77 (2001).
44. R. D. Chirico, I. A. Hossenlopp, A. Nguyen, W. V. Steele, B. E. Gammon. *J. Chem. Thermodyn.* **21**, 179 (1989).
45. M. A. V. Ribeiro da Silva, A. F. L. O. M. Santos, J. R. B. Gomes, M. V. Roux, M. Temprado, P. Jiménez, R. Notario. *J. Phys. Chem. A* **113**, 11042 (2009).
46. IUPAC. *Quantities, Units and Symbols in Physical Chemistry*, 3rd ed. (the "Green Book"). Prepared for publication by E. R. Cohen, T. Cvitaš, J. G. Frey, B. Holmström, K. Kuchitsu, R. Marquardt, I. Mills, F. Pavese, M. Quack, J. Stohner, H. L. Strauss, M. Takami, A. J. Thor, RSC Publishing, Cambridge, UK (2007). First corrected printing (2008).
47. S. K. Rakshit, S. C. Parida, S. Dash, Z. Singh, V. Venugopal. *Thermochim. Acta* **443**, 98 (2006).
48. E. S. Hamborg, G. F. Versteeg. *J. Chem. Eng. Data* **54**, 1318 (2009).
49. L. Ciavatta, V. Elia, E. Napoli, M. Niccoli. *J. Solution Chem.* **37**, 1037 (2008).
50. L. G. Longsworth, D. A. MacInnes. *J. Am. Chem. Soc.* **60**, 3070 (1938).

51. D. D. Wagman, W. H. Evans, V. B. Parker, R. H. Schumm, I. Halow, S. M. Bailey, K. L. Churney, R. L. Nuttall. "The NBS tables of chemical thermodynamic properties", *J. Phys. Chem. Ref. Data* **11**, Suppl. No. 2 (1982).
52. H. M. Rosenstock, K. Draxl, B. W. Steiner, J. T. Herron. "Energetics of gaseous ions", *J. Phys. Chem. Ref. Data* **6**, Suppl. No. 1 (1977).
53. M. W. Chase Jr. (Ed.). *NIST-JANAF Thermochemical Tables*, 4th ed., American Chemical Society, American Institute of Physics for the National Institute of Standards and Technology, Washington, DC (1998).
54. M. Nic. In *Chemical Information Mining: Facilitating Literature-Based Discovery*, D. L. Banville (Ed.), pp. 99–122, CRC Press, Boca Raton (2009).
55. E. Bolton, Y. Wang, P. A. Thiessen, S. H. Bryant. In *Annual Reports in Computational Chemistry*, Vol. 4, Chap. 12, American Chemical Society, Washington, DC (2008). See also <http://pubchem.ncbi.nlm.nih.gov/>.
56. InChI Version 1, Software Version 1.03 - implemented for both Standard and Non-standard (Customized) InChI/InChIKey <http://www.iupac.org/inchi/release103.html>.
57. IUPAC project 2000-025-1-800, "IUPAC International Chemical Identifier", A. D. McNaught <http://www.iupac.org/web/ins/2000-025-1-800>.
58. The InChI Trust <http://www.inchi-trust.org>.
59. A. Yu. Kuznetsov, L. S. Kudin, A. M. Pogrebnoi, M. F. Butman, G. G. Burdukovskaya. *Zh. Fiz. Khim.* **73**, 566 (1999).
60. P. Kratochvíl, U. W. Suter. *Pure Appl. Chem.* **61**, 211 (1989).
61. E. Muñoz de Miguel, C. Yanes, A. Maestre. *J. Chem. Eng. Data* **46**, 423 (2001).
62. H.-J. Hinz, F. P. Schwarz. *J. Chem. Thermodyn.* **33**, 1511 (2001).
63. Y. B. Tewari, J. Chen, M. J. Holden, K. N. Houk, R. N. Goldberg. *J. Phys. Chem. B* **102**, 8634 (1998).
64. N. A. Todorova, F. P. Schwarz. *J. Chem. Thermodyn.* **39**, 1038 (2007).
65. R. Carta. *J. Chem. Thermodyn.* **30**, 379 (1998).
66. A. Ortiz, F. J. Aranda, J. Villalaín, J. C. Gómez-Fernández. *Biochim. Biophys. Acta* **1122**, 226 (1992).
67. The LIPIDAT home page, <http://www.lipidat.tcd.ie/>.
68. J. Chao, K. R. Hall, K. N. Marsh, R. C. Wilhoit. *J. Phys. Chem. Ref. Data* **15**, 1369 (1986).