1344 *J. Chem. Eng. Data* **2003,** *48,* 1344−1359

ThermoML—An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 2. Uncertainties

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

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

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. Extensions to the ThermoML schema for the expression of uncertainties are described. Basic principles, scope, and description of all new structural elements are discussed. Representation of upper and lower limits for property values is also addressed. 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). Properties of polymers and radicals and some properties of ionic systems are not represented at present. The present role of ThermoML in global data submission and dissemination is discussed with particular emphasis on cooperation between major journals in the field and the Thermodynamics Research Center (TRC) at the National Institute of Standards and Technology. The text of several data files illustrating the expression of uncertainties in ThermoML format for pure compounds, mixtures, and chemical reactions are provided as Supporting Information, as well as the complete updated ThermoML schema text.

Background

The basic principles, scope, and description of all structural elements of ThermoML were discussed in the first paper[1] in this series. 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) with a primary focus on molecular compounds. Properties of polymers and radicals and some properties of ionic systems are not represented at present. Expansion of ThermoML to cover these systems is under development. Representation of quantities for the expression of uncertainty was not considered in the first paper because the complex issues involved require extensive discussion. Extension of the ThermoML schema to include representation of these quantities is described here. A third paper in this series is planned for description of extensions to ThermoML for the representation of critically evaluated and predicted data.

The expression of uncertainty requires clear definition of a variety of quantities and terms. Definitions and descriptions of all quantities related to the expression of uncertainty in this paper conform to the *Guide to the Expression of Uncertainty in Measurement*, ISO (International Organization for Standardization), October, 1993.[2] These ISO recommendations were adopted with minor editorial changes as the *U.S. Guide to the Expression of Uncertainty in Measurement*.[3] Reference 2 is commonly referred to by its abbreviation; the *GUM*. Reference 3 is assumed equivalent to ref 2 and is referred to as the *Guide* in this paper. The historical development of these recommendations beginning in 1977 is described in the *Guide*. The recommendations have been summarized in *Guidelines for the Evaluation and Expression of Uncertainty in NIST Measurement Results*,[4] which is available via free download from the Internet (http://physics.nist.gov/cuu/).

The definitions given in this paper are for the convenience of the reader and are not meant to modify in any way those given in the internationally accepted guides.[2,3] In this paper, application of the recommendations given in the *Guide* to particular aspects of experimental thermodynamic property data will be discussed, and additions to the ThermoML schema will be described fully. The present paper describes uncertainty data structures primarily in application to thermodynamic data obtained experimentally. Many of the concepts described could also be applied to predicted and critically evaluated data, which will be the principal subject of the next paper in this series. The

---

* Corresponding author. Phone: (303)-497-4126. Fax: (303)-497-5044. E-mail: chirico@boulder.nist.gov.

10.1021/je034088i CCC: $25.00

Table 1. Relationships between Quantities Used for the Expression of Uncertainty in ThermoML

state functions (measurand, Y)	measurement result, y	standard uncertainty (1σ),[a] u	coverage factor,[b] k	expanded uncertainty, U	level of confidence,[b] L (%)
variable(s), Y_V	y_V	$u_V = f(v_1, v_2, v_3, ...)$	k_V	$U_V = u_V k_V$	$L_V = f(k_V)$
constraint(s), Y_C	y_C	$u_C = f(c_1, c_2, c_3, ...)$	k_C	$U_C = u_C k_C$	$L_C = f(k_C)$
property, Y_P	y_P	$u_P = f(p_1, p_2, p_3, ...)$	k_P	$U_P = u_P k_P$	$L_P = f(k_P)$
for *properties* Y_P (only)	measurement result, y	combined standard uncertainty (1σ) [c]	combined coverage factor	combined expanded uncertainty	combined level of confidence
property, Y_P	y_P	$u_{comb} = f(u_V, u_C, u_P)$	k_{comb}	$U_{comb} = u_{comb} k_{comb}$	$L_{comb} = f(k_{comb})$

[a] All components of uncertainty are included except those of other state functions. [b] For many practical situations with assumed normal distributions, a coverage factor k near 2 corresponds to a level of confidence L near 95%. [c] Standard uncertainties of variables and constraints are propagated to the uncertainty of the property.

complete ThermoML schema is available on the Internet (http://www.trc.nist.gov/ThermoML.xsd).

Basic Principles and Definitions

All definitions of quantities related to the expression of uncertainty are quoted or closely adapted from the *Guide*.[3] Table 1 shows the general mathematical relationships between the quantities used for the expression of uncertainty, which are used explicitly in ThermoML. Uncertainties are represented for *variables*, *constraints*, and *properties*. The terms listed in the headings of Table 1 (*measurand*, *standard uncertainty*, *coverage factor*, *combined standard uncertainty*, etc.) are taken from the *Guide* and form the basis for quantitative specification of uncertainties. These terms are defined in the following sections.

Measurand: The particular quantity subject to measurement.

The objective of a measurement is to determine the value of the measurand Y. Within ThermoML, measurands are state functions (constraints, variables, and properties), as shown in column 1 of Table 1. The value of a measurand is by definition unknowable. Consequently, the meanings of the phrases "true value of the measurand" and "value of the measurand" are the same. The result of a measurement y is an estimate of the value of the measurand Y and is complete only when accompanied by a statement of the uncertainty of that estimate. The measurement result y for each state function is shown symbolically (y_C, y_V, y_P) in column 2 of Table 1.

The *Guide* establishes general rules for expression of uncertainty, which are designed to be applicable to a wide variety of applications and levels of accuracy "from the shop floor to fundamental research". Consequently, some information in the *Guide* is not applicable to thermodynamic property measurements, such as discussions of uncertainties associated with incomplete definition of the measurand. All measurands specified within ThermoML are fully specified through their thermodynamic definitions. The use of the GDC (Guided Data Capture) software[5] for data capture ensures full property specification through rigorous implementation of the Gibbs phase rule. The role of ThermoML and GDC in global data submission and dissemination is described later in this paper.

The *Guide* distinguishes between two uses of the word *uncertainty*, and this same distinction is used here. The word uncertainty means doubt, and therefore, in its broadest sense *uncertainty of measurement* means doubt about the validity of the result of a measurement. Because of the lack of different words for this general concept of uncertainty and the specific quantities that provide quantitative measures of the concept, for example, the standard deviation, it is necessary to use the word uncertainty in these two different senses. In the present paper, as in the *Guide*, the word uncertainty without adjectives refers to the general concept of uncertainty and to any or all quantitative measures of that concept. When a specific measure is intended, appropriate adjectives are used.

The formal definition of uncertainty listed in the *International Vocabulary of Basic and General Terms in Metrology*[6] (commonly abbreviated VIM) and used by the *Guide* follows.

Uncertainty (of measurement): a parameter, associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand.

NOTES

1. The parameter may be, for example, a standard deviation (or a given multiple of it), or a half-width of an interval having a stated level of confidence.

2. Uncertainty of measurement comprises, in general, many components. Some of these components may be evaluated from the statistical distribution of the results of series of measurements and can be characterized by experimental standard deviations. The other components, which can also be characterized by standard deviations, are evaluated from assumed probability distributions based on experience or other information.

3. It is understood that the result of the measurement is the best estimate of the value of the measurand, and that all components of uncertainty, including those arising from systematic effects, such as components associated with corrections and reference standards, contribute to the dispersion.

This general definition of *uncertainty of measurement* is not associated with a numerical value. As noted earlier, numerical uncertainties will always include modifying words, such as *standard* uncertainty, *expanded* uncertainty, *combined* uncertainty, and so forth. It is very important to note that this definition includes all possible contributions to its value. It is therefore clearly distinguishable from other quantities, which are commonly reported (and defined later), such as repeatability, reproducibility, or deviations from a fitted curve. Such quantities may be components of an estimated uncertainty but can provide only lower limits for the *uncertainty of measurement*.

A second important point is expressed in note 2 (above), which describes how *all* uncertainties are expressed as standard deviations. These standard deviations can be derived from probability distributions based on (1) "the statistical analysis of series of observations" (designated *Type A*) or (2) other means (designated *Type B*). Because the uncertainty of measurement includes all sources of uncertainty, it is probable that components of *Type A* and *Type B* will be included in an uncertainty estimate. A *Type B* estimate for a component of the uncertainty is "evaluated by scientific judgment based on all of the available information"[3] concerning the variability of the quantity considered. To clarify the meaning of a *Type B* estimate, the *Guide* lists some items that might be included in that pool of information: (i) previous measurement data, (ii) experience with or general knowledge of the behavior and properties of relevant materials and instruments, (iii) manufacturers' specifications, (iv) data provided in calibration and other certificates, and (v) uncertainties assigned to reference data taken from handbooks.

To base the uncertainty of measurement entirely upon *Type A* evaluations, it would be necessary to include numerous repetitions of all steps leading to the measurement, including procurement and purification of samples, calibration of instruments, construction and operation of instruments and apparatus, and so forth. This is clearly not practical, particularly in the field of experimental thermodynamics. Consequently, estimates of uncertainties will almost always include some, if not primarily, *Type B* evaluations.

A third important aspect involves application of this definition in ThermoML. In ThermoML, uncertainties are represented for all thermodynamic state functions (i.e., variables, constraints, and properties) and not just for those designated as properties. This is reflected in Table 1 in that uncertainties for all state functions are represented. For many experiments, such as phase equilibrium studies, the identification of the terms variable and property with the measured temperatures, pressures, and phase compositions is arbitrary. Similarly, for a pure component, "vapor pressure data" might be considered saturation pressures at given temperatures or boiling temperatures at given pressures. The issue of propagation of uncertainties from variables and/or constraints to the designated property is discussed separately later in this paper.

The *standard uncertainty* differs from the more general term *uncertainty of measurement* in that the value of the uncertainty is expressed as a specified number of standard deviations (i.e., one).

Standard Uncertainty u: Uncertainty of the result of a measurement expressed as a standard deviation.

This quantity u_x is represented in column 3 of Table 1 as an expression

$$u_x = f(x_1, x_2, x_3, ...) \quad (1)$$

where the x_i symbols represent various uncertainty components that are appropriately weighted to estimate u_x. For example, the estimated uncertainty for a temperature value might be a function of the method and traceability of the sensor calibration, the instrument used to read its response, estimated gradients in the apparatus, effects of thermal inertia, and so forth. A well-designed experiment will improve the quality of these estimates, but some scientific judgment is always involved.

In a broader sense, u_x could be considered a *combined standard uncertainty* (defined later) in that it "combines" uncertainties from various sources. In ThermoML, the term *combined standard uncertainty* is reserved for uncertainties derived by propagation of uncertainties in variables and constraints to those for the designated property. More detail is provided below.

As noted in a footnote of Table 1, the standard uncertainties associated with state functions are defined to be independent and do not include uncertainty components associated with propagation of uncertainty from one state function to another. For example, if the density of a single-component gas (the property) is reported as a function of temperature and pressure (the variables), it is important to avoid including the effect of uncertainty in temperature upon the uncertainty in the pressure. This is to avoid overestimation (or "double counting") of uncertainties, when they are propagated to the designated property in a subsequent step.

In Table 1, the standard uncertainty is listed independently for the variables, the constraints, *and* the property. The appropriateness of this type of reporting can be shown using typical results for vapor−liquid equilibrium (VLE) experiments. In the reporting of VLE results, pressures p, temperatures T, and phase-composition values x (liquid) and y (vapor) are reported. Uncertainties associated with each quantity (p, T, x, and y) are often reported independently. In the case of overdetermined systems, all the data points are separated into data sets characterized with independent variables only. Furthermore, a "property" is often not specified explicitly because the designation is arbitrary. To accommodate this type of reporting, ThermoML includes representation of standard uncertainties for the variables, the constraints, *and* the property. A second uncertainty (the *combined* uncertainty) is defined only for the property and includes propagation of uncertainties from the variables and constraints to the property. The combined uncertainty is represented separately, as shown in the bottom of Table 1. Further discussion is given below with the definition of the *combined standard uncertainty*.

Expanded Uncertainty U: the quantity defining an interval about the result of a measurement that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand.

1. The fraction may be viewed as the *coverage probability* or *level of confidence* of the interval.

2. To associate a specific *level of confidence* with the interval defined by the *expanded uncertainty* requires explicit or implicit assumptions regarding the probability distribution characterized by a measurement result and its *standard uncertainty*. The *level of confidence* that may be attributed to this interval can be known only to the extent to which such assumptions may be justified.

Coverage Factor k: Numerical factor k is a multiplier of the *standard uncertainty* in order to obtain an *expanded uncertainty*.

NOTE: A coverage factor k is typically in the range 2 to 3.

The relationship between the *standard uncertainty*, the *coverage factor*, and the *expanded uncertainty* for the state functions in ThermoML is shown in the upper section of Table 1.

$$\textit{Expanded uncertainty } U_x = \textit{Standard uncertainty } u_x \cdot \textit{Coverage factor } k_x \quad (2)$$

Propagation of uncertainties from the variables and constraints to the property is not included in the *expanded uncertainty* U_x.

Combined Standard Uncertainty u_{comb}: the standard uncertainty of the result of a measurement when that result is obtained from the values of a number of other quantities, equal to the positive square root of a sum of terms, the terms being the variances or covariances of these other quantities weighted according to how the measurement result varies with changes in these quantities.

In ThermoML, the *combined standard uncertainty* u_{comb} is included only for the quantity designated as the property.

Table 2. Values of the Coverage Factor k_L that Produce an Interval Having the Level of Confidence L Assuming a *Normal* Distribution

level of confidence, L (%)	coverage factor, k_L
68.27	1
90	1.645
95	1.96
95.45	2
99	2.576
99.73	3

The *combined coverage factor* k_{comb} and the *combined expanded uncertainty* U_{comb}, which also apply only to the designated property, are defined through the equation

$$U_{\text{comb}} = u_{\text{comb}} k_{\text{comb}} \tag{3}$$

This relationship is also shown in the lower section of Table 1.

The final important term used in the expression of uncertainty in ThermoML is the *Level of Confidence L*, which is represented as a percentage.

Level of Confidence L: an estimate of the probability that the measurand is within a specified range $y \pm U$.

[Note: In the *Guide*, the symbol p is used for *Level of Confidence* (or *Coverage Probability*). To avoid confusion with other terms such as *property* or *pressure*, the symbol L is used here. The term *confidence interval* is not used because in statistics it has a specific definition, which requires that all components of the uncertainty be obtained from *Type A* evaluations. As noted earlier, the results of experimental thermodynamic studies will very rarely meet this criterion.] The *Guide* includes an extensive discussion (Annex G) of the relationship between the level of confidence and the coverage factor. There is little value in repeating that discussion here, but a summary of several useful points follows.

Examples of the well-known correspondence between L and k for a normal distribution are given in Table 2. In general, to obtain the value of the *coverage factor* k_L that produces an interval corresponding to a specified *level of confidence L* requires detailed knowledge of the probability distribution characterizing the measurement results. However, as noted in the *Guide* (section G.6.6), for many practical measurements in a broad range of fields, some general approximations can be made. These emphasize the approximate nature of the uncertainty evaluation process and the impracticality of trying to distinguish between intervals having levels of confidence that differ by only a few percent. The following recommendations are given in the *Guide*: (i) adopt $k = 2$ and assume that $U = 2u_c(y)$ defines an interval having a level of confidence of approximately 95%, (ii) or for more critical applications, adopt $k = 3$ and assume that $U = 3u_c(y)$ defines an interval having a level of confidence of approximately 99%.

These practical suggestions should be of use to many authors reporting results of studies in experimental thermodynamics. Nonetheless, the *coverage factor* and *level of confidence* are represented separately in ThermoML to allow for those conditions that define an alternative mathematical relationship. It should be emphasized that the relationship between L and k_L shown in Table 2 is for an assumed normal distribution, and that for other types of distributions (rectangular, triangular, etc.) other relationships are obtained.

The following series of definitions describe terms that are commonly reported in the literature as reproducibilities, repeatabilities, or deviations from a fitted curve. These uncertainty-assessment components are referred to collectively in this paper as *precisions*. These represent components of an uncertainty assessment but do not meet the criteria for the *uncertainty of measurement*, which includes all sources of uncertainty. It was decided to include some precisions in ThermoML because certain of these quantities can be well defined and may be useful to data evaluators in subsequent assessments. The *International Vocabulary of Basic and General Terms in Metrology*[6] (commonly abbreviated *VIM*) does not give a definition for precision because of the many definitions that exist for this word. This is consistent with the usage in this paper.

Repeatability (of results of measurements) [VIM 3.6]: Closeness of the agreement between the results of successive measurements of the same measurand carried out under the same conditions of measurement.

NOTES

1. These conditions are called repeatability conditions.
2. Repeatability conditions include (i) the same measurement procedure, (ii) the same observer, (iii) the same measuring instrument used under the same conditions, (iv) the same location, and (v) repetition over a short period of time.
3. Repeatability may be expressed quantitatively in terms of the dispersion characteristics of the results.

Because *repeatability* can be expressed quantitatively, is readily defined, and is commonly reported in the thermodynamic literature, this quantity is represented in ThermoML. It must be recognized clearly by the user of this information that repeatability is only one component of many that can be used in estimating the broader *uncertainty of measurement*.

Reproducibility (of results of measurements) [VIM 3.7]: Closeness of the agreement between the results of measurements of the same measurand carried out under *changed* conditions of measurement.

The definition of *reproducibility*[3] includes an extensive list of what *might* be changed, including the principle of measurement, the method of measurement, the observer, the location, the reference standard, the measuring instrument, and so forth. To implement this definition in ThermoML, an extensive list of what was and what was not varied would be required. Consequently, representation of reproducibility is not included in ThermoML.

Several representations of precisions, which are not defined in the *Guide*, are included in ThermoML. These are included because they are commonly reported in the literature of experimental thermodynamics.

Root-mean-square (rms) deviation from a fitted curve (for a data set): A description of the fitted curve is included in ThermoML in addition to the numerical root-mean-square (rms) value. The deviation is represented in ThermoML for properties only and not for variables or constraints. This quantity is associated with entire data sets and is not associated with individual numerical values (i.e., data points). In ThermoML, the rms deviation can be expressed as an absolute or relative value (percentage).

The value for the *rms deviation from a fitted curve* δ_{rms} is defined by the following equation:

$$\delta_{\text{rms}} = [\sum (x_i - x_{\text{curve}})^2/n]^{0.5} \tag{4}$$

The summation is over all data points. The symbols in the equation represent the number of data points n and the deviation $(x_i - x_{\text{curve}})$ of data point i from the fitted curve.

Deviation from a fitted curve (for each data point in a data set): Again, a description of the fitted curve is included in ThermoML. The numerical value of the deviation from the curve for each data point of a data set is represented. Deviations are expressed in absolute terms only because relative (percentage) values can be derived readily.

Device Specifications: Device specifications from manufacturers or calibration certificates are often used and reported as part of an uncertainty assessment by experimentalists. The numerical value for this uncertainty component, level of confidence, and the information source should be specified. Elements corresponding to each are present in the extensions to ThermoML described in this paper.

Again, it must be recognized clearly by the user of this information that the *repeatabilities*, *deviations from a fitted curve,* and *device specifications* are components in an array of information that can be used in estimating the *uncertainty of measurement*.

Precision: As noted earlier, this term has numerous conflicting meanings, and it is not expressed quantitatively in the authoritative literature.[3]

This term is not represented in ThermoML.

Accuracy (of Measurement) [VIM 3.5]: Closeness of agreement between the result of a measurement and the measurand. “Accuracy is a qualitative concept.”[3]

Accuracy cannot be represented quantitatively and is not represented in ThermoML.

When determination of a measurand is impossible, values that represent only an upper or lower limit are reported at times in the thermodynamic literature. For example, this may be due to sample decomposition or to a limited experimental range available with a particular apparatus. Such limiting values are related indirectly to uncertainties in that the sign of the deviation from the measurand is known without an estimate of its magnitude. Limiting values are represented in ThermoML for properties only and not for constraints or variables. Details of the quantities represented are described later.

Implementation of Uncertainty Definitions in ThermoML

The *Guide* provides specific recommendations for the reporting of uncertainties (Chapter 7: Reporting Uncertainty), which are accommodated fully in ThermoML. Specifically, section 7.2.3 of the *Guide* lists recommendations for reporting of the *expanded uncertainty U*. The recommendations are as follows:

(1) Give a full description of how the *measurand Y* is defined.

(2) State the result of the measurement as $Y = y \pm U$ and give the units of y and U.

(3) Include the *relative expanded uncertainty* $U/|y|$, $|y| \neq 0$, when appropriate.

(4) Give the value of k used to obtain U [or, for the convenience of the user, give both k and u].

(5) Give the approximate *level of confidence* associated with the interval $y \pm U$ and state how it was determined.

(6) Give the information outlined in section 7.2.7 of the *Guide* or refer to a published document that contains it. (Note: section 7.2.7 makes specific recommendations related to reporting of the origins of all uncertainty estimates in the document text.)

Recommendation 1 was implemented previously in ThermoML[1] through the definition of variables, constraints, and properties based on the established laws of phenomenological thermodynamics. The extensions to ThermoML that are the focus of this paper directly address recommendations 2−5. Although not included explicitly, the *relative expanded uncertainty* (item 3) can be derived simply from values that are represented. Recommendation 6 involves detailed reporting suggestions that would be impractical to implement fully in ThermoML. These include listing the source of all uncertainty estimates used to estimate any standard uncertainty and providing partial derivatives or “sensitivity coefficients” related to key uncertainty components. Nonetheless, in all cases a text schema element is provided, which can be used for descriptions of the uncertainty estimation methods.

Schema Structure. The modular schema structure of ThermoML was described in the first paper of this series.[1] All new schema elements described here are extensions to ThermoML and do not modify the original structure in any way. ThermoML includes separate structures for the representation of variables, constraints, and properties. It will be seen that there is a one-to-one correspondence in elements for variables, constraints, and properties for representation of the uncertainty related quantities listed in the upper section of Table 1 (*standard uncertainty*, *coverage factor*, *expanded uncertainty*, and *level of confidence*). Some structural differences arise because constraints are defined once for a data set, while variables and properties are associated with individual numerical values or “data points.” Uncertainty elements for properties also include separate elements related to propagated uncertainties from the variables and constraints to the property (i.e., the *combined* quantities shown in the lower section of Table 1).

Units. By design, there is only one unit selected for each property covered by ThermoML. This unit is SI-based; however, for a number of properties the selected unit is a multiple of the SI unit to ease interpretation of numerical values. Unit tagging is explicit in a ThermoML file as part of each property name, thus minimizing the possibility of unit misinterpretation. As a natural extension to this approach, the units for the uncertainty values are the same as those of the quantity whose uncertainty is being represented. Relative uncertainties are represented as percentages. These are indicated clearly in the element names and descriptions.

Tagging. All tagging related to uncertainty is based upon terminology described in the *Guide*.[3] The names or “tags” for the elements of the schema include special characters related to the type of information to be stored. A name beginning with “e” designates an enumeration element (i.e., values selected from a predefined list), “s” designates a string element, and “n” specifies a numerical element (integer or floating). Elements identified by dotted boxes in a figure are optional, and those in solid-lined boxes are mandatory. Within ThermoML, all elements associated with the representation of uncertainty are optional. 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 figures. Within ThermoML, the only limits used are “0..∞” for optional elements and “1..∞” for mandatory elements. The “Choice” symbol, which appears as a switch (first seen in Figure 6), is used to designate elements that cannot be used simultaneously.

Extensions to the ThermoML Schema

ThermoML consists of four major blocks, which are shown in Figure 1 together with the new element **Version**

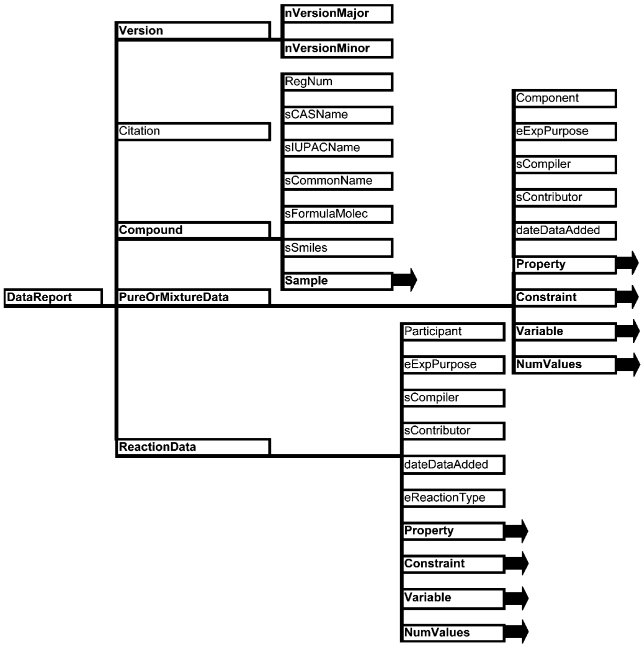

Figure 1. Major components of ThermoML. The arrows indicate locations of extensions to the schema for representation of uncertainties. New schema elements for storage of the ThermoML version number are included.

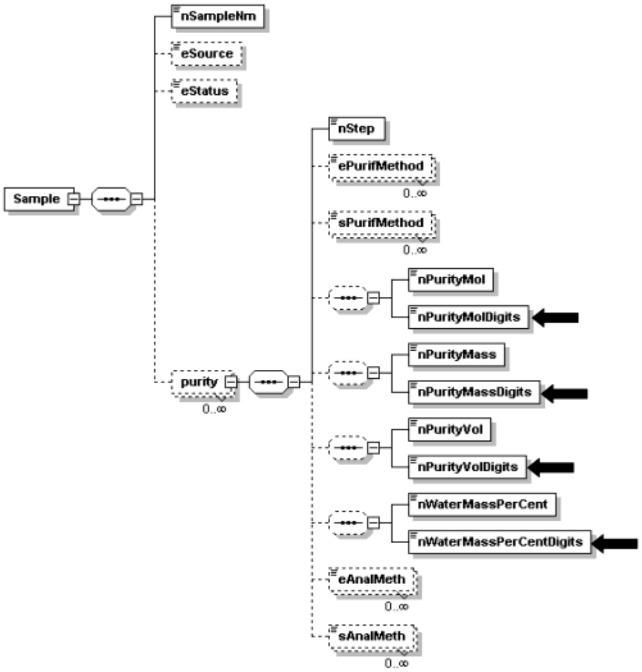

Figure 2. Structure of the **Sample** element of the "*Compound*" block. The arrows indicate new elements described in the text.

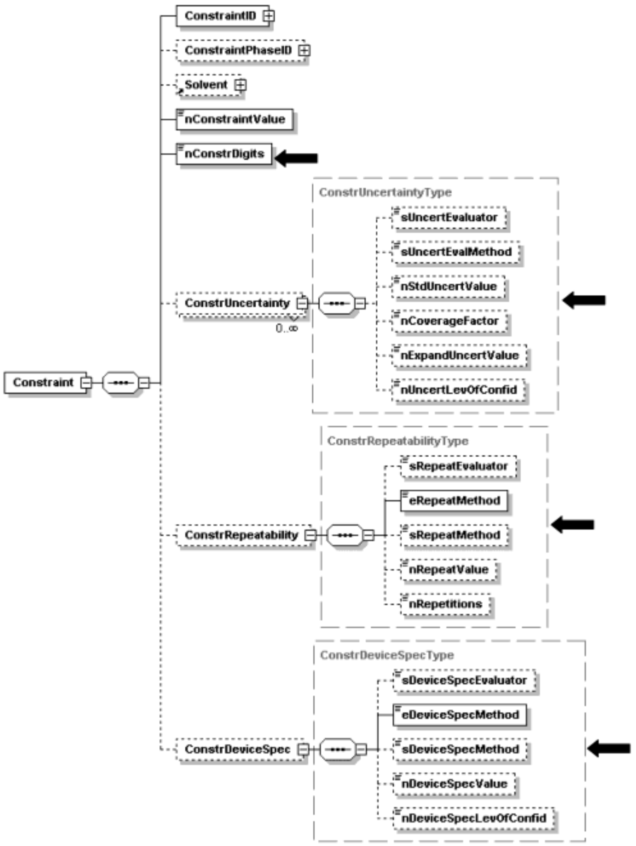

Figure 3. Structure of the **Constraint** element in the "*PureOrMixtureData*" block. The arrows indicate new elements described in the text.

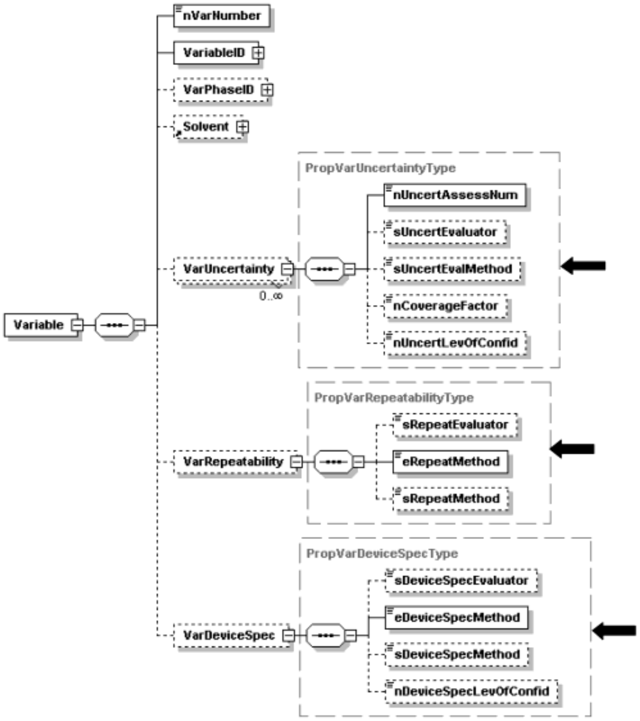

Figure 4. Structure of the **Variable** element in the "*PureOrMixtureData*" block. The arrows indicate new elements described in the text.

[complex]. The four major blocks were described previously.[1] A new element **Version** [complex] is mandatory and provides for storage of the ThermoML version designation. The subelements of **Version** [complex] are **nVersionMajor** [numerical, integer] and **nVersionMinor** [numerical, integer]. For example, if the version number of ThermoML were 2.1, the "major" element would store the value "2" and the "minor" element would store the value "1".

The general locations of extensions described in this paper are indicated in Figure 1. Detailed schema figures (Figures 2−10) and the text of ThermoML were created with the software package XML SPY.[7] (The trade name is provided only to specify the procedure adequately and does not imply endorsement by the National Institute of Standards and Technology. Similar products by other manufacturers may be found to work as well or better.) The four

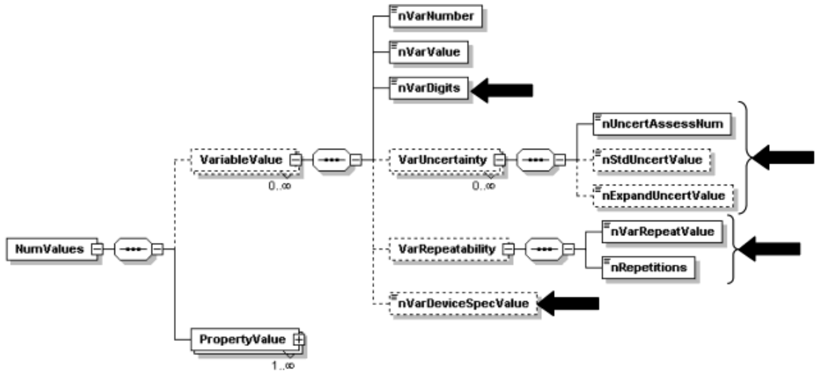

Figure 5. Structure of the **VariableValue** subelement within the **NumValues** element in the "*PureOrMixtureData*" block and in the "*ReactionData*" block. The arrows indicate new elements described in the text.

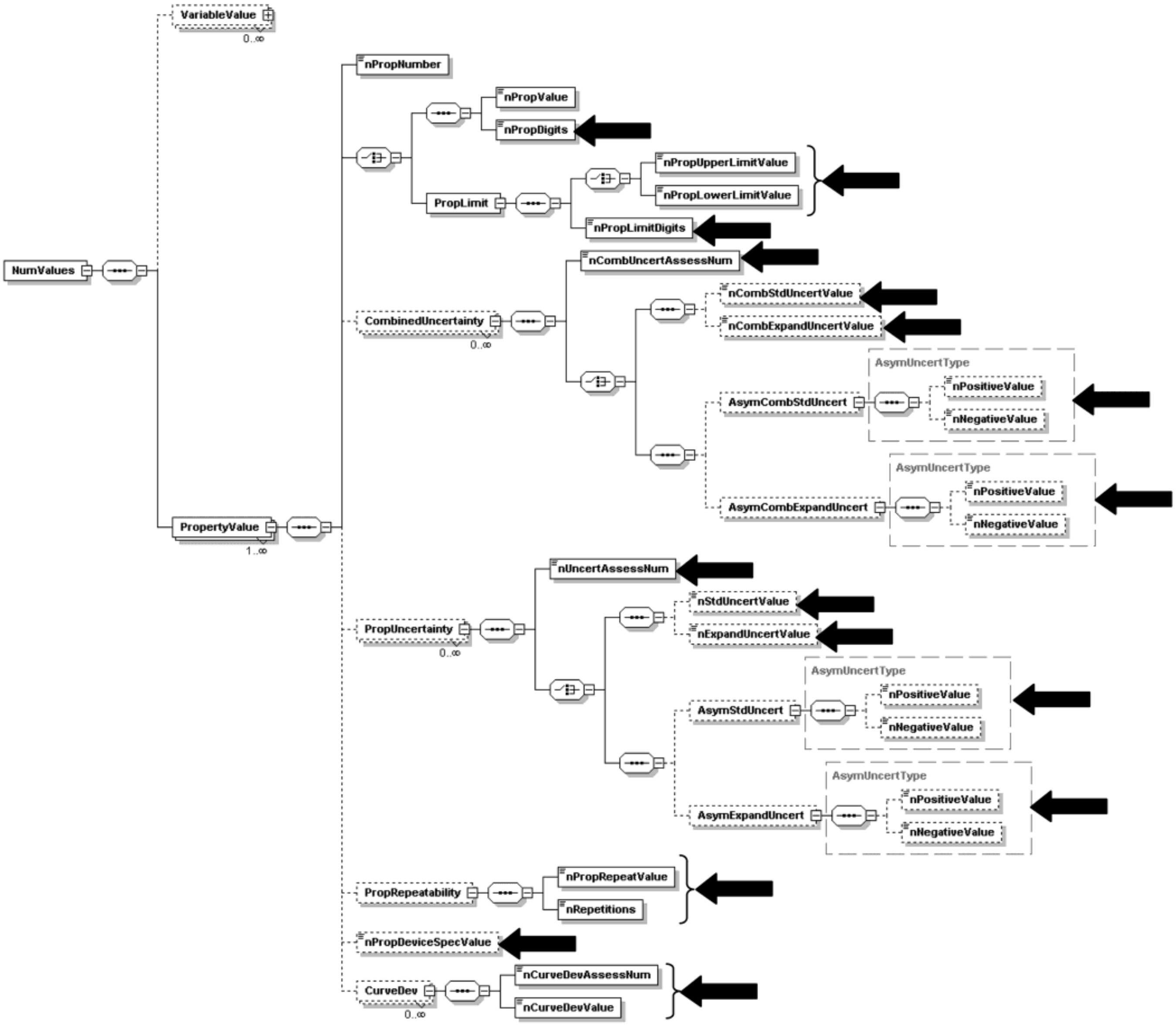

Figure 6. Structure of the **PropertyValue** subelement within the **NumValues** element in the "*PureOrMixtureData*" block and in the "*ReactionData*" block. The arrows indicate new elements described in the text.

major blocks are as follows:

(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 methods 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 change of state or in chemical equilibrium).

The extensions described in this paper are primarily in blocks 3 and 4, and they are associated with the numerical

Table 3. Names and Locations of ThermoML Elements for the Expression of Uncertainties

			location[b]	
abbreviated element name	full element name	measurand type[a]	set	value
**Uncertainty* [complex][c]	uncertainty			
nUncertAssessNum [numerical, integer]	uncertainty assessment number	P, V, C	•	•
sUncertEvaluator [string]	uncertainty evaluator	P, V, C	•	
sUncertEvalMethod [string]	uncertainty evaluation method	P, V, C	•	
nStdUncertValue [numerical, floating]	standard uncertainty value u_x	P, V, C		•
AsymStdUncert [complex][d]	standard uncertainty value u_x	P only	•	
nCoverageFactor [numerical, floating]	coverage factor k_x used to obtain $U_x = k_x u_x$	P, V, C	•	
nExpandUncertValue [numerical, floating]	expanded uncertainty value U_x	P, V, C		•
AsymExpandUncert [complex][d]	asymmetrical uncertainty	P only	•	
nUncertLevOfConfid [numerical, floating]	level of confidence (%) L_x associated with U_x	P, V, C	•	
CombinedUncertainty	combined uncertainty	P only		
nCombUncertAssessNum [numerical, integer]	combined uncertainty assessment number	P only	•	•
sCombUncertEvaluator [string]	combined uncertainty evaluator	P only	•	
eCombUncertEvalMethod [enumeration]	combined uncertainty evaluation method	P only	•	
sCombUncertEvalMethod [string]	combined uncertainty evaluation method	P only	•	
nCombStdUncertValue [numerical, floating]	combined standard uncertainty value u_{comb}	P only		•
AsymCombStdUncert [complex][d]	combined standard uncertainty value u_{comb}	P only		•
nCombCoverageFactor [numerical, floating]	coverage factor k_{comb} to obtain $U_{\mathrm{comb}} = k_{\mathrm{comb}} u_{\mathrm{comb}}$	P only	•	
nCombExpandUncertValue [numerical, floating]	combined expanded uncertainty value U_{comb}	P only		•
AsymCombExpandUncert [complex][d]	combined expanded uncertainty value U_{comb}	P only		•
nCombUncertLevOfConfid [numerical, floating]	level of confidence (%) L_{comb} associated with U_{comb}	P only	•	

[a] P = property, V = variable, C = constraint. [b] *Location* indicates the location of the element in the ThermoML schema. *Set* specifies elements associated with an entire data set, and *Value* specifies elements associated with individual numerical values. [c] The name of this complex element is different for a property (* = *Prop*), the variable (* = *Var*), and constraint (* = *Constr*). [d] Subelements of this complex element represent the positive and negative values of the asymmetrical uncertainty.

values for the variables, constraints, and properties, as shown in Figure 1. Numerical values of uncertainty are not provided for compound purities (block 2), but an extension to allow specification of the number of digits is now added. The number of digits specified should correspond approximately to the number of significant digits, but there is no strict requirement for this correspondence. A comprehensive uncertainty-specification scheme for purities would add unjustifiable complexity to the schema.

The following sections describe details of the additions made to ThermoML for representation of uncertainties for state functions (variables, constraints, and properties). Many of the new schema elements for constraints, variables, and properties have common definitions. To avoid unnecessary repetition of definitions, the following discussion is separated into two sections. The first involves definition and tagging of the new schema elements, and the second involves their precise location in the schema structure. All of the definitions apply to results for pure components, mixtures, and reactions. The elements for the representation of uncertainty are the same for blocks 3 and 4 (*PureOrMixtureData* and *ReactionData*). Extensions to ThermoML for the representation of uncertainties in blocks 3 and 4 differ only in terms of location.

Definition and Structure of New Schema Elements

New schema elements are summarized in Tables 3 and 4. Table 3 lists all elements associated with the specification of defined *uncertainties*, while Table 4 lists all elements associated with the specification of defined *precisions*. This separation is done to emphasize the conceptual difference between these two types of quantities. Column 1 in Tables 3 and 4 indicates the element name used in ThermoML. Column 2 clarifies the meaning of the abbreviations used in column 1. Column 3 specifies the type of measurand to which the names apply. Columns 4 and 5 specify the general location of the element in the ThermoML schema in terms of whether the element is associated with a "data set" or with each "data point." In the following sections, each element is defined and some specific guidance related to its usage is given.

ThermoML Elements for the Expression of Uncertainties

The elements listed in Table 3 are separated into those for representation of defined *uncertainties*, which are given for variables, constraints, and properties, and those for *combined uncertainties*, which are given for properties only. Elements for representation of the defined *uncertainties* are addressed first.

***Uncertainty** *[complex]*

This complex element includes the subelements associated with expression of the *expanded uncertainty* values U_x for variables, constraints, and properties. The symbol * indicates that this element is present in the schema for variables (* = **Var**), constraints (* = **Constr**), and properties (* = **Prop**). The subelements of ***Uncertainty** [complex] are described in the following paragraphs.

nUncertAssessNum [numerical, integer]; the *uncertainty assessment number* is an integer used to identity a particular assessment of the uncertainty. ThermoML can accommodate multiple uncertainty assessments for the same data. For variables and properties, this number also serves to link uncertainty elements associated with individual data points (e.g., the *expanded uncertainty value*) with an element associated with the entire data set (e.g., the *coverage factor*). The *uncertainty assessment number* for a given assessment is associated with all of the subelements within ***Uncertainty** [complex], as listed in the upper section of Table 3.

sUncertEvaluator [string]; the *uncertainty evaluator* string is used to identify the individual or institution responsible for the assessment. For information reported explicitly in journal articles, the evaluator is the author(s). Multiple evaluations can be stored simultaneously with ThermoML.

sUncertEvalMethod [string]; the *uncertainty evaluation method* element allows storage of descriptive information related to the uncertainty assessment, such as the

Table 4. Names and Locations of ThermoML Elements for the Expression of Precisions

abbreviated element name[a]	full element name	measurand type[b]	location[c]	
			set	value
n*Digits [numerical, integer]	number of digits	P, V, C		●
**Repeatability* [complex]	repeatability			
sRepeatEvaluator [string]	repeatability evaluator	P, V, C	●	
eRepeatMethod [enumeration]	repeatability method	P, V, C	●	
sRepeatMethod [string]	repeatability method	P, V, C	●	
nRepeatValue [numerical, floating]	repeatability value	P, V, C		●
nRepetitions [numerical, integer]	number of repetitions	P, V, C	●	
**DeviceSpec* [complex]	device specification			
sDeviceSpecEvaluator [string]	device specification evaluator	P, V, C	●	
eDeviceSpecMethod [enumeration]	device specification method	P, V, C	●	
sDeviceSpecMethod [string]	device specification method	P, V, C	●	
nDeviceSpecValue [numerical, floating]	uncertainty based on device specification only	P, V, C		●
nDeviceSpecLevOfConfid [numerical, floating]	level of confidence for nDeviceSpecValue	P, V, C	●	
CurveDev [complex]	rms deviation from a fitted curve			
nCurveDevAssessNum [numerical, integer]	fitted curve assessment number	P only	●	●
sCurveDevEvaluator [string]	fitted curve evaluator	P only	●	
sCurveSpec [string]	fitted curve specification	P only	●	
nCurveRmsDevValue [numerical, floating]	fitted curve rms deviation value	P only	●	
nCurveRmsRelativeDevValue [numerical, floating]	fitted curve relative rms deviation value (%)	P only	●	
nCurveDevValue	deviations from the fitted curve	P only		●

[a] The names of some elements are different for a property (* = *Prop*), the variable (* = *Var*), and constraint (* = *Constr*). [b] P = property, V = variable, C = constraint. [c] *Location* indicates the location of the element in the ThermoML schema. *Set* specifies elements associated with a data set as a whole, and *Value* specifies elements associated with individual numerical values.

sources of key information. In this way, this element can be used to accommodate item number 6 of the recommendations for the reporting of uncertainties given in the *Guide* (*Guide* section 7.2.3).

nStdUncertValue [numerical, floating] is for storage of the numerical *standard uncertainty* value u_x, as defined earlier and shown in column 3 of Table 1. By definition, the value u_x represents one standard deviation.

AsymStdUncert [complex] is for storage of the *standard uncertainty* for an uncertainty that is asymmetrical about the property value. This element is included only for properties and not for variables, constraints, or defined precisions. A symmetrical uncertainty can by represented as $y_P \pm u_P$, where y_P is the measured value and u_P is the *standard uncertainty*. The values spanned by this range lie between $(y_P + u_P)$ and $(y_P - u_P)$. In contrast, the values spanned by an unsymmetrical range lie between $(y_P + u_{P+})$ and $(y_P - u_{P-})$, where $u_{P+} \neq u_{P-}$. Subelements of **AsymStdUncert** [complex] are **nPositiveValue** [numerical, floating] for storage of u_{P+} and **nNegativeValue** [numerical, floating] for storage of u_{P-}.

nCoverageFactor [numerical, floating] is used for storage of the *coverage factor* k_x, as defined earlier and shown in column 4 of Table 1.

nExpandUncertValue [numerical, floating] is used for storage of the *expanded uncertainty* value U_x, as defined earlier and shown in column 5 of Table 1. It is recognized that simultaneous storage of **nStdUncertValue**, **nCoverageFactor**, and **nExpandUncertValue** is redundant because of the simple relationship $u_x k_x = U_x$. It is a recommendation of the *Guide* that all three be given to avoid any possible ambiguity.

AsymExpandUncert [complex] is for storage of the *expanded uncertainty* for an uncertainty that is asymmetrical about the property value. This element is included only for properties. The values spanned by the unsymmetrical range lie between $(y_P + U_{P+})$ and $(y_P - U_{P-})$, where $U_{P+} \neq U_{P-}$. Subelements of **AsymExpandUncert** [complex] are **nPositiveValue** [numerical, floating] for storage of U_{P+} and **nNegativeValue** [numerical, floating] for storage of U_{P-}.

nUncertLevOfConfid [numerical, floating] is used for storage of the *Level of Confidence* L_x associated with U_x. The level of confidence is always expressed in ThermoML as a percentage.

CombinedUncertainty *[complex]*

This complex element includes the subelements associated with expression of the *combined expanded uncertainty* values U_{comb}. The *combined expanded uncertainty* is stored for properties only and includes propagation of uncertainties in the variables and constraints to that of the property, as discussed earlier and as indicated in the lower section of Table 1. Nearly all subelements are analogous to those for ***Uncertainty** [complex]. One additional element provides an enumeration list for description of the method of assessment.

nCombUncertAssessNum [numerical, integer]; the *combined uncertainty assessment number* is an integer used to identity a particular assessment of the combined uncertainty. Its use is as described for **nUncertAssessNum** [numerical, integer].

sCombUncertEvaluator [string]; the *combined uncertainty evaluator* string is for identification of the individual or institution responsible for the assessment of the *combined uncertainty*.

eCombUncertEvalMethod [enumeration] provides an enumeration list for specification of the general method for evaluation of the *combined uncertainty*. The enumerations are (Propagation of estimated standard uncertainties, Comparison with reference property values). In the field of experimental thermodynamics, it is common to test the reliability of an apparatus through measurements performed on reference materials with property values having well-established uncertainties. This approach is an alternative to that of determining uncertainties for all possible components and propagating these to the uncertainty in the property. The source of the reference property values should be provided in the corresponding string element, which follows.

sCombUncertEvalMethod [string]; the *combined uncertainty evaluation method* element allows storage of descriptive information related to the uncertainty assessment. This element can be used to accommodate item number 6 of the recommendations for the reporting of uncertainties given in the *Guide*. In addition, this element can be used for storage of information about the source of reference property values if "Comparison with reference property values" is chosen in **eCombUncertEvalMethod** [enumeration].

nCombStdUncertValue [numerical, floating] is used for storage of the combined standard uncertainty value u_{comb}, as defined earlier and shown in column 3 of the lower section of Table 1. The value u_{comb} represents one standard deviation by definition.

AsymCombStdUncert [complex] is for storage of the *combined standard uncertainty* for an uncertainty that is asymmetrical about the property value. This element is included only for properties. The values spanned by the unsymmetrical range lie between ($y_{\mathrm{P}} + u_{\mathrm{comb+}}$) and ($y_{\mathrm{P}} - u_{\mathrm{comb-}}$), where $u_{\mathrm{comb+}} \neq u_{\mathrm{comb-}}$. Subelements of **AsymCombStdUncert** [complex] are **nPositiveValue** [numerical, floating] for storage of $u_{\mathrm{comb+}}$ and **nNegativeValue** [numerical, floating] for storage of $u_{\mathrm{comb-}}$.

nCombCoverageFactor [numerical, floating] stores the *coverage factor* k_{comb} used to obtain $U_{\mathrm{comb}} = k_{\mathrm{comb}}u_{\mathrm{comb}}$.

nCombExpandUncertValue [numerical, floating] stores the *combined expanded uncertainty* value U_{comb}.

AsymCombExpandUncert [complex] is for storage of the *combined expanded uncertainty* for an uncertainty that is asymmetrical about the property value. This element is included only for properties. The values spanned by the unsymmetrical range lie between ($y_{\mathrm{P}} + U_{\mathrm{comb+}}$) and ($y_{\mathrm{P}} - U_{\mathrm{comb-}}$), where $U_{\mathrm{comb+}} \neq U_{\mathrm{comb-}}$. Subelements of **AsymCombExpandUncert** [complex] are **nPositiveValue** [numerical, floating] for storage of $U_{\mathrm{comb+}}$ and **nNegativeValue** [numerical, floating] for storage of $U_{\mathrm{comb-}}$.

nCombUncertLevOfConfid [numerical, floating] stores the *Level of Confidence* L_{comb} associated with U_{comb}. The level of confidence is stored as a percentage.

PropLimit *[complex]*

This complex element is a new subelement of **PropertyValue** [complex] and is for storage of property values reported as upper or lower limits for the measurand. This type of element is included for properties only and is not included for variables, constraints, or precisions. The subelements of **PropLimit** [complex] are **nPropUpperLimitValue** [numerical, floating] and **nPropLowerLimitValue** [numerical, floating] for storage of either numerical limiting value, and **nPropLimitDigits** [numerical, integer] for storage of the number of digits in the value.

ThermoML Elements for the Expression of Precisions

The elements listed in Table 4 are for the expression of precisions. The measurand types to which these may be associated are listed in column 3 of Table 4. These quantities are completely independent of those specified in Table 3. The term *combined* is not used in these elements. This term is applicable only to uncertainties, which include contributions from *all* sources. With the exception of **n*Digits** [numerical, integer] all of the following elements are optional in the schema, as is true for all elements associated with the specification of uncertainty. The general locations of the elements listed in Table 4 are indicated in columns 4 and 5. Detailed locations are provided later in this paper.

n*Digits [numerical, integer]. Every numerical element in ThermoML has an element of this type associated with it for specification of the number of reported digits. The symbol * indicates that this element is present in the schema for many different quantities. The number of decimal places associated with uncertainties and precisions is assumed equal to the number specified for the relevant variable, constraint, or property. By design, the number of digits is not specified as the number of significant digits. There are numerous occasions in the reporting of experimental thermodynamic results when it is valuable to report additional digits to avoid round-off errors in subsequent calculations. If uncertainties are reported as recommended in the *Guide*, the number of significant digits can be determined from these.

****Repeatability*** *[complex]*

The numerical quantity, *repeatability*, was defined earlier. The symbol * indicates that this element is present in the schema for variables (* = **Var**), constraints (* = **Constr**), and properties (* = **Prop**). The following elements are used for specification of this quantity for variables, constraints, and properties.

sRepeatEvaluator [string]; the *repeatability evaluator* string is used to identify the individual or institution responsible for the assessment of the repeatability. In most cases, this will be the author of the original publication.

eRepeatMethod [enumeration] provides an enumeration list for specification of the statistical definition of the *repeatability value*. The four enumerations are (Standard deviation of a single value (biased), Standard deviation of a single value (unbiased), Standard deviation of the mean, and Other). These terms are defined in most common texts in the field of statistics (cf. ref 8). Selection of the enumeration "Other" should be accompanied by a description of the method in the string element **sRepeatMethod** [string].

The *standard deviation of a single value (unbiased)* $\sigma_{\mathrm{unbiased}}$ for a series of observations with mean value $\bar{x}$ is calculated with the following equation.[8] The summation is over the number n of observations.

$$\sigma_{\mathrm{unbiased}} = [(n-1)^{-1}\sum(x_i - \bar{x})^2]^{0.5} \qquad (5)$$

The *standard deviation of a single value (biased)* σ_{biased} is calculated with the equation[8]

$$\sigma_{\mathrm{biased}} = [n^{-1}\sum(x_i - \bar{x})^2]^{0.5} \qquad (6)$$

The *standard deviation of the mean* σ_{mean} is calculated with the equation[9]

$$\sigma_{\mathrm{mean}} = [\{n(n-1)\}^{-1}\sum(x_i - \bar{x})^2]^{0.5} \qquad (7)$$

A discussion of the application of these formulas to particular experimental conditions is beyond the scope of this paper. The reader is referred to any common text in statistics for this information. Equation 7 is often applied in the analysis of results obtained with combustion bomb calorimetry.[9]

sRepeatMethod [string]; the *repeatability assessment method* can be used for storage of details related to the determination, such as the particular type of statistics used in determining the repeatability value. This element should always be used when "Other" is selected in **eRepeatMethod** [enumeration].

nRepeatValue [numerical, floating] is used for storage of the *repeatability value*. The units match those of the quantity being repeated.

nRepetitions [numerical, integer] is used for storage of the number of *measurement repetitions* n used in the calculation of the repeatability value.

****DeviceSpec*** *[complex]*

This complex element includes subelements for storage of components of uncertainty obtained as *device specifications* from manufacturers or certificates of calibration. The symbol * indicates that this element is present in the schema for variables (* = **Var**), constraints (* = **Constr**), and properties (* = **Prop**). These quantities are often used and reported as part of an uncertainty assessment by experimentalists, and they may be of value to subsequent data evaluators.

sDeviceSpecEvaluator [string]; the *device specification evaluator* string is used to identify the individual or institution responsible for assessment of the device specification. In most cases, this will be a manufacturing company or an institute or company providing calibration services.

eDeviceSpecMethod [enumeration] provides an enumeration list for identification of the *device specification method*. The three enumerations are (Specified by the manufacturer, Calibrated by the experimentalist, Calibrated or certified by a third party). Details related to the specification can be described in **sDeviceSpecMethod** [string].

sDeviceSpecMethod [string]; this element is used for storage of details related to the enumeration selected in **eDeviceSpecMethod** [enumeration]. Details might include particulars of the calibration method, identities and sources of reference materials, literature references to standard values, and so forth.

nDeviceSpecValue [numerical, floating] is used for storage of the numerical value of the uncertainty component arising from the device specification. The units match those of the state function being determined with the device.

nDeviceSpecLevOfConfid [numerical, floating] is used to store the *level of confidence* (percent) associated with **nDeviceSpecValue** [numerical, floating].

***CurveDev** [complex]*

This element allows storage of uncertainty information derived from fitting of curves to experimental property data. By definition, these quantities are associated with properties only. The information stored is the root-mean-square deviation of the experimental values from the fitted curve (for a *data set*) and the deviations from the fitted curve for each numerical value (i.e., for each *data point*), as indicated in Table 4.

nCurveDevAssessNum [numerical, integer]; the *deviation assessment number* is an integer used to identity a particular assessment. Its use is as described for **nUncertAssessNum** [numerical, integer]. The assessment number is needed to allow storage of results for fits with various equations. Identification of the particular equation is stored in **sCurveSpec** [string], as described below.

sCurveDevEvaluator [string]; the *curve deviation evaluator* element is used to identify the individual or institution responsible for the assessment.

sCurveSpec [string]; the *curve specification* element is used for storage of text that describes the fitted curve. The description might include a particular equation name (e.g., Antoine or Wagner for vapor pressures), an equation form (e.g., $C_{p,m} = a + bT$, for heat capacities of a liquid), or special conditions, such as specification of fixed parameters.

nCurveRmsDevValue [numerical, floating]; the *curve rms deviation value* is stored in this element and has the same units as the associated property. This value is associated with the data set as a whole. The numerical value δ_{rms} is defined in eq 4.

nCurveRmsRelativeDevValue [numerical, floating]; the *curve rms relative deviation value* is stored in this element as a percentage. This value is also associated with the data set as a whole. Calculation of this value is analogous to that shown in eq 4, but with the deviations expressed as $[100(x_i - x_{\text{curve}})/x_{\text{curve}}]$ rather than $(x_i - x_{\text{curve}})$.

nCurveDevValue [numerical, floating]; the *curve deviation value* $(x_i - x_{\text{curve}})$ is stored in this element and is the deviation of a particular numerical value (a *data point*) from the specified fitted curve. This value is associated always with an individual data point. The units are those of the property. Percentage values are not represented explicitly because they can be easily derived from the values provided.

Detailed Locations in ThermoML of Elements for the Expression of Uncertainties' Precisions

Figure 1 provides an overview of extension locations in the ThermoML schema. This figure can be used with subsequent figures to show the detailed connections between the new schema elements and the primary "*DataReport*" shown in Figure 1.

Extensions to the "Compound" Block. Extensions to the "*Compound*" block are shown in Figure 2. In ThermoML, purities are characterized in terms of mole percent, mass percent, volume percent, and mass percent of water present. Four new elements, which are used to store the number of digits associated with each purity type, are now added to the schema. The new elements, **nPurityMolDigits** [numerical, integer], **nPurityMassDigits** [numerical, integer], **nPurityVolDigits** [numerical, integer], and **nWaterMassPerCentDigits** [numerical, integer], are indicated by arrows in the figure. New schema elements are indicated by arrows in all of the detailed schema figures (Figures 2−10).

Extensions to the "PureOrMixtureData" Block. Extensions to this block are shown separately for constraints, variables, and properties. New elements for the expression of uncertainties and precisions for constraints are shown in Figure 3. All numerical values for constraints are associated with data sets rather than individual data points. Consequently, the element **nConstraintValue** [numerical, floating] is included within **Constraint** [complex], as seen in the figure. The number of digits for **nConstraintValue** [numerical, floating] is now represented in **nConstrDigits** [numerical, integer]. All elements for specification of the constraint uncertainty are within the element **ConstrUncertainty** [complex], as seen the figure. In contrast, new elements for the expression of uncertainty for variables and properties must be split between the location associated with the data set and that associated with the individual data points.

New elements for the expression of uncertainties and precisions for variables in the "*PureOrMixtureData*" block are shown in Figures 4 and 5. Figure 4 shows those elements associated with the data set as a whole, while Figure 5 shows the elements associated with the individual numerical data points. The element **nVarDeviceSpecValue** [numerical, floating] is associated with the individual data points because device specifications are sometimes given as a function of the size of the measured value (e.g., as a percentage) and are not constant for the entire data set. In addition, different devices may be used for measurements within a single data set. If different devices are used, it is preferable to identify a separate data set with each separate device.

New elements for the expression of uncertainties and precisions for properties in the "*PureOrMixtureData*" block

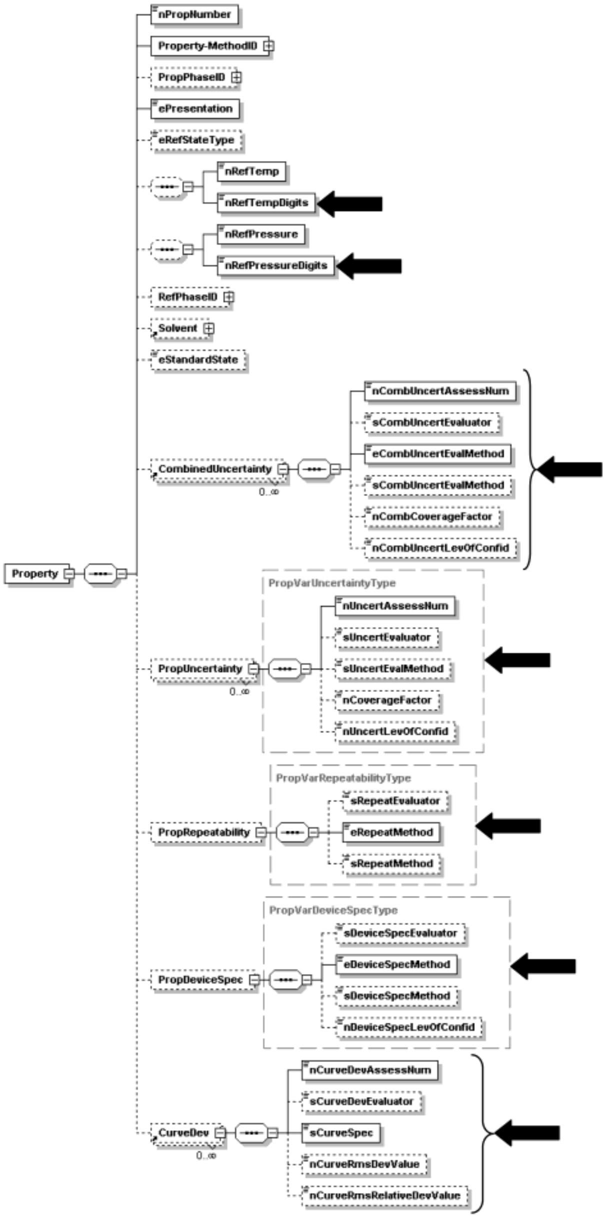

Figure 7. Structure of the **Property** element in the "*PureOrMixtureData*" block. The arrows indicate new elements described in the text.

are shown in Figures 6 and 7. Figure 7 shows those elements associated with the data set as a whole, while Figure 6 shows the elements associated with the individual numerical data points. The additional elements for properties associated with representation of *combined uncertainties* and *deviations from fitted curves* are apparent in the figures.

Extensions to the "ReactionData" Block. Extensions to this block are also shown separately for constraints, variables, and properties. Detailed locations for all of the new elements for the expression of uncertainties and precisions for constraints are shown in Figure 8. This is analogous to the structure shown in Figure 3 for the "*PureOrMixtureData*" block.

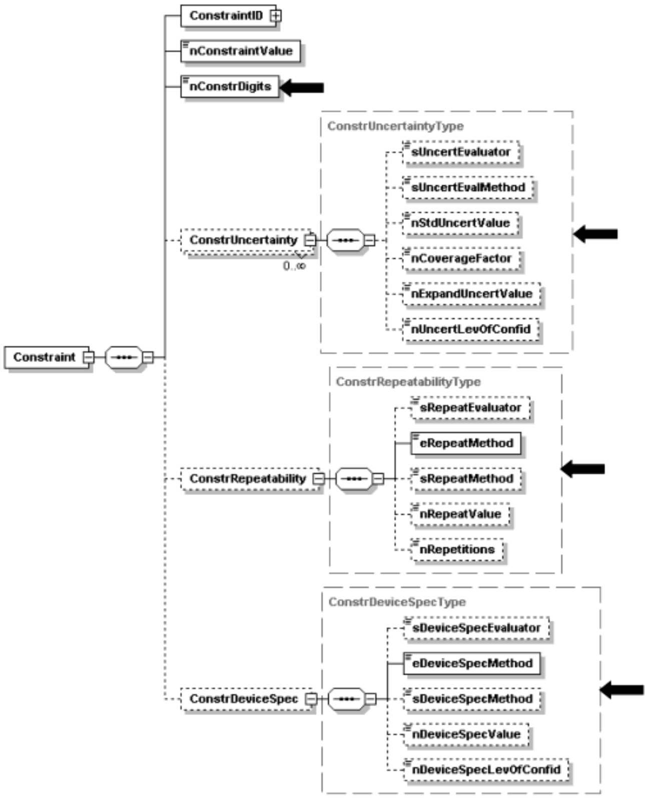

Figure 8. Structure of the **Constraint** element in the "*ReactionData*" block. The arrows indicate new elements described in the text.

Figure 9. Structure of the **Variable** element in the "*ReactionData*" block. The arrows indicate new elements described in the text.

New elements for the expression of uncertainties and precisions for variables in the "*ReactionData*" block are shown in Figures 5 and 9. The schema structure for the representation of numerical values for variables is the same in the "*PureOrMixtureData*" block and the "*ReactionData*" block. Figure 9 shows those elements associated with the data set as a whole, while Figure 5 shows the elements associated with the individual numerical data points.

New elements for the expression of uncertainties and precisions for properties in the "*ReactionData*" block are shown in Figures 6 and 10. The structure of the schema for the representation of numerical values for properties is the same in the "*PureOrMixtureData*" block and the "*ReactionData*" block. Figure 10 shows those elements

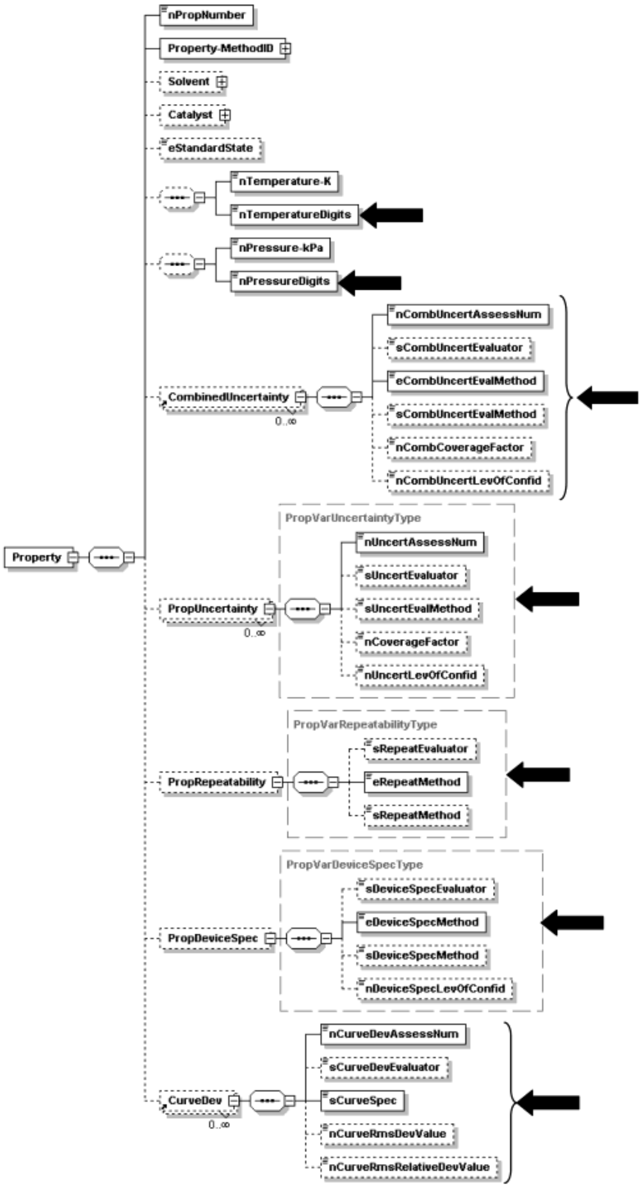

Figure 10. Structure of the **Property** element in the "*ReactionData*" block. The arrows indicate new elements described in the text.

associated with the data set as a whole, while Figure 6 shows the elements associated with the individual numerical data points.

Example 1: A Comparison of the *Expanded Uncertainty* and the *Combined Expanded Uncertainty*

The distinctly different meanings of the *expanded uncertainty* and the *combined expanded uncertainty* can be demonstrated with the following vapor−liquid equilibrium data set for a binary mixture. The reported state functions are the pressure ($p = 101.3$ kPa), the composition of the liquid phase x_1 (the mole fraction of component 1), and the boiling temperature T of the mixture. The experimental boiling temperatures are shown plotted as a function of the composition of the liquid phase in Figure 11 (upper graph).

The contributions to the uncertainties for each independent state function were analyzed by the authors, and the following values of the *standard uncertainty* u were estimated: $u_T = 0.1$ K, $u_p = 0.1$ kPa, and $u_{x_1} = 0.005$. Normal distributions were assumed in each case, and a coverage factor $k = 2$ was used to estimate each *expanded uncertainty*: $U_T = 0.2$ K, $U_p = 0.2$ kPa, and $U_{x_1} = 0.010$.

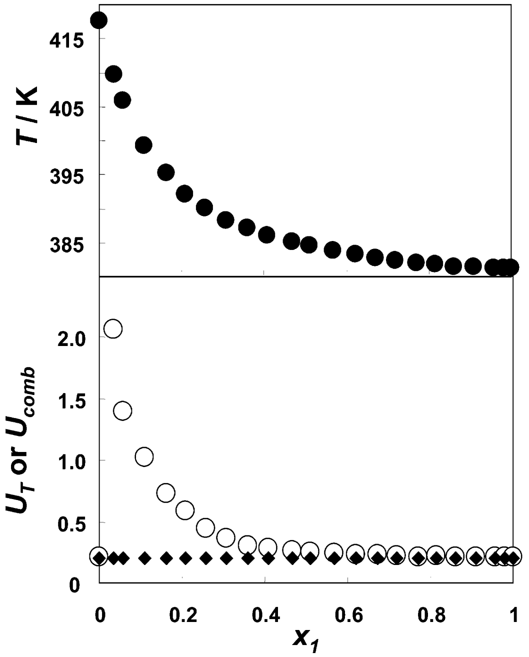

Figure 11. Experimental boiling temperatures T (●) for a two-component mixture, the *expanded uncertainty* U_T (◆), and the *combined expanded uncertainty* U_{comb} (○) plotted against the mole fraction of component 1 in the liquid phase, x_1. Calculations of U_T and U_{comb} are described in the text.

The *combined expanded uncertainty* U_{comb} can be calculated for any of the state functions, but it can be stored in ThermoML for only one: the designated property. In this example, the boiling temperature T is selected as the property, and the *combined expanded uncertainty* is calculated by propagation of U_p and U_{x_1} to calculate U_{comb}.

Figure 11 (lower graph) shows the two types of uncertainties (U_T and U_{comb}) associated with the temperature, as a function of x_1. U_T is independent of the other state functions and is constant. For $x_1 > 0.5$, U_T and U_{comb} are nearly the same. For $x_1 < 0.5$, however, U_{comb} increases rapidly (to a value approximately 10 times U_T) and passes through a maximum before approaching U_T again near $x_1 = 0$. The contribution of U_{x_1} to U_{comb} [i.e., $U_{x_1}(\partial T/\partial x_1)_p$] is the primary source of the large values for $0 < x_1 < 0.5$. The contribution from U_p is relatively small.

Example 2: Estimates of *Combined Expanded Uncertainties*

The estimation of uncertainties for experimental data is a key step in the data evaluation processes, which form the foundation for all recommended thermodynamic data. The NIST Thermodynamics Research Center (TRC) is one of the oldest data research centers in the United States, and for over 60 years has produced a great number of periodical compilations (the *TRC Tables*,[10] *International Data Series* for mixtures,[11] *TRC Books* of evaluated property data,[12] etc.) that have become major sources of recommended data for scientific research and industrial process design.

The following paragraphs outline an approach for the estimation of *combined standard uncertainties* developed at TRC and used by the personnel of the TRC Data Entry Facility. This example is provided to reinforce the concept that all contributions to uncertainty should be considered in the estimation of *uncertainty of measurement*. The estimates described are based on the information provided in a given document (e.g., a journal article, thesis, report, etc.) and do not involve comparisons with previously evaluated data. This is done so as not to bias the assessment. These estimates provide preliminary values, which are subject to further improvement based on the enforcement of single-property and multiple-property consistency requirements included in the TRC data quality-assurance program.[13]

Estimated uncertainties for experimental data have essentially two purposes: (1) to provide weighting factors needed to distinguish between duplicate determinations (i.e., relative uncertainties) and (2) to provide a basis for propagation of uncertainties to derived properties required for recommended values for chemical-process design, benchmark values for validation of results from ab initio or other computational methods, and numerous other applications (i.e., absolute uncertainties).

Estimation of uncertainties for experimental data is critical to a new approach for data evaluation that is under development at NIST. This approach, which is termed dynamic data evaluation (DDE), requires the development of large electronic databases capable of storing essentially all experimental data known to date with complete descriptions of the relevant metadata and uncertainties. The combination of these electronic databases with expert-system software leads to the ability to produce compilations of recommended values dynamically or “to order”. General descriptions of DDE have been published.[14,15] As part of the DDE implementation, TRC has developed a general and consistent scheme for estimation of uncertainties for a wide variety of properties and experimental methods.

Although almost all publications of experimental data include some discussion or estimation of “uncertainties”, these are rarely done in the structured and defined way recommended by the *Guide*. Inconsistencies abound between different journals and publications, and conflicting interpretations and implementations of many terms such as precision, accuracy, systematic error, random error, and so forth are common. Consequently, most “uncertainties” reported in the literature cannot be used for weighting of data or error propagation, particularly if the reported uncertainties are combined from different sources.

The scheme developed by TRC for estimation of the *combined standard uncertainty* u_{comb} for a given property p as a function of constraints c and variables v is based upon a summation of terms:

$$u_{\mathrm{comb}}^2 = u_p^2 + \sum\{u_c(\partial p/\partial c)_v\}^2 + \sum\{u_v(\partial p/\partial v)_c\}^2 \quad (8)$$

The partial derivatives $(\partial p/\partial c)_v$ and $(\partial p/\partial v)_c$ are calculated approximately upon the basis of the reported values if possible, or they are estimated upon the basis of approximate models for the property. The summations are over all constraints and variables, respectively. The *standard uncertainty* for the property u_p is rarely provided in a document and is estimated at TRC upon the basis of the following general relationship:

$$u_p^2 = \{u_{\mathrm{method}}^2 + \sum(f_m u_{\mathrm{method-details}}^2)\} + \{u_{\mathrm{sample}}^2 + \sum(f_s u_{\mathrm{sample-details}}^2)\} \quad (9)$$

This relationship involves two major contributions to u_p: uncertainties associated with the *experimental method*, and those associated with the *sample*.

The term u_{method} is a default contribution to u_p and is based on the particular experimental method only. For example, a heat capacity $C_{\mathrm{sat},m}$ determined with high-precision adiabatic calorimetry might have a default value for u_{method} of $0.002C_{\mathrm{sat},m}$, while the same property determined with a differential-scanning calorimeter might have a default value 10 times larger. Some details related to particular methods are also considered, such as the method of calibration for a vibrating-tube densimeter or the failure to report degassing procedures for a static vapor-pressure measurement. These adjustments are indicated as $u_{\mathrm{method-details}}$ in eq 9 and can increase or decrease u_p on the basis of the value of f_m, which is 1 or −1.

The term u_{sample} in eq 9 represents a default contribution to u_p related directly to the purity of the sample. Additional contributions to u_p related to the sample are indicated as $u_{\mathrm{sample-details}}$ in eq 9. The magnitude of $u_{\mathrm{sample-details}}$ is a function of several items, including the property, experimental method, special characteristics of the material (e.g., thermal stability or hygroscopicity), and the experimental conditions (e.g., pressure or temperature range). This formulation is required to take into account the fact that impurities do not affect all properties or experimental methods to the same extent. The value of f_s is 1 or −1.

Values for the standard uncertainties u_c and u_v (and u_p, if appropriate) are taken from the document, if provided and supported in the text. Default values are substituted for those not provided. Default values are based upon the general method used and are larger than those reported typically in the literature for the method. Incomplete reporting or the absence of this information in a document is considered indicative of the general quality of the work. Consequently, results reported with incomplete uncertainty descriptions are given uncertainties at TRC larger than those with well-supported estimates.

If estimates of u_{comb} are provided in the document, these are checked against the estimates calculated with eq 8. Large discrepancies are reviewed carefully and can form the basis for modification of default values. Because various indicators of precision (repeatabilities, deviations from fitted curves, etc.) provide only a lower limit for any uncertainty estimate, these are considered only if larger than the default uncertainties for the particular variable, constraint, or property. As a final step, the estimated combined standard uncertainty u_{comb} is multiplied by a coverage factor k_{comb} to estimate the combined expanded uncertainty U_{comb} corresponding to a level of confidence of approximately 95%. For most cases, k_{comb} equals 2; however, other values are used, if required to obtain the desired level of confidence. The U_{comb} values are stored together with the property values y_P and required metadata in the TRC SOURCE Data System.[18]

This approach to the estimation of *combined standard uncertainties* provides the basis for consistent evaluations of the numerous data types encountered by TRC as part of its thermodynamic data evaluation projects. This brief overview demonstrates that even approximate estimates of u_{comb} require careful consideration of a wide variety of contributions to the uncertainty.

Use Cases and ThermoML Schema Text

Examples illustrating the format of the data files created with the ThermoML formats for pure compound or mixture data sets,[16] as well as for chemical reactions,[17] are included as Supporting Information. The examples are based upon experimental studies published in the peer-reviewed literature.

Schema Validation

The developed schema was validated extensively with data records in SOURCE.[18] Validation covered essentially all properties within the scope of ThermoML, including pure compounds, multicomponent mixtures, and chemical reactions. In addition, the validation process included data files submitted to TRC by authors of articles submitted through the Editorial Board of the *Journal of Chemical & 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 was described.[1] Guided Data Capture (GDC) software[5] was developed at TRC for mass-scale abstraction from the literature of experimental thermophysical and thermochemical property data. This software is freely available for download from the Web.[19]

Following the peer-review process, authors are requested by the journal editors to download and use the GDC software to capture the experimental property data that has been accepted for publication. The output of the GDC software is an electronic data file (plain text file), which is submitted directly to TRC. The electronic data files are converted into ThermoML format with software (TransThermo) developed at TRC. Upon release of the manuscript for publication, the ThermoML files are posted on the public-domain TRC Web site with unrestricted public access. This procedure has been established formally by the *Journal of Chemical & Engineering Data*.[20] Expansion of this operation to other journals in the field is under discussion.

Version 1 of the GDC software included very limited tools for the specification and capture of uncertainty information. This software is now being expanded to include a more comprehensive representation of uncertainty based on the definitions and structures described here. The new version (Version 2) will also be available for free download from the Internet.[19] After release of this new version of GDC, posted ThermoML files will include uncertainty estimates provided by the article authors. Version 2 will implement tools for generation of ThermoML files upon completion of the data-capture process.

Acknowledgment

The authors express their appreciation to the following people for their valuable advice in the development of the extensions to the ThermoML format described in this paper: Drs. J. W. Magee, R. A. Perkins, D. G. Friend, and W. M. Haynes (NIST, Boulder, CO), Drs. S. E. Stein, T. C. Allison, and P. Linstrom (NIST, Gaithersburg, MD). Our special thanks to Dr. G. J. Rosasco for his support and help in coordination of the project. The authors also wish to express their appreciation to Dr. D. L. Embry (DIPPR, ConocoPhillips, Ponca City, OK), Dr. T. L. Teague (DIPPR, ePlantData, Inc., Houston, TX), Dr. P. R. Franchois (NIST, Boulder, CO), S. Frenkel (NIST, Boulder, CO), for their valuable cooperation in development of the ThermoML format, and Drs. W. A. Wakeham (University of Southhampton, U. K.), J. H. Dymond (University of Glasgow, U. K.), and A. R. H. Goodwin (Schlumberger-Doll Research, Ridgefield, CT) for valuable discussions concerning global data submission and dissemination.

Supporting Information Available: Examples are provided illustrating the use of ThermoML for representation of experimental data with uncertainties for pure compounds, mixtures, and chemical reactions. The complete current text of the ThermoML schema is included also as Supporting Information and is available on the Web (http://www.trc.nist.gov/ThermoML.xsd) or through direct request to the authors. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Frenkel, M.; Chirico, R. D.; Diky, V. V.; Dong, Q.; Frenkel, S.; Franchois, P. R.; Embry, D. L.; Teague, T. L.; Marsh, K. N.; Wilhoit, R. C. ThermoML—An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 1. Experimental Data. *J. Chem. Eng. Data* **2003**, *48*, 2−13.

(2) *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.

(3) *U.S. Guide to the Expression of Uncertainty in Measurement*, ANSI/NCSL Z540-2-1997, ISBN 1-58464-005-7. NCSL International: Boulder, CO, 1997.

(4) Taylor, B. N.; Kuyatt, C. E. *Guidelines for the Evaluation and Expression of Uncertainty in NIST Measurement Results*; NIST Technical Note 1297; NIST: Gaithersburg, MD, 1994.

(5) 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.* **2003**, *43*, 15−24.

(6) *International Vocabulary of Basic and General Terms in Metrology*, 2nd ed. (International Organization for Standardization, Geneva, Switzerland, 1993). This document (abbreviated VIM) was prepared by ISO Technical Advisory Group 4 (TAG 4), Working Group 1 (WG 1). 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. The individual members of WG 1 were nominated by BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, or OIML, and the document is published by ISO in the name of all seven organizations.

(7) *XML SPY*, v. 4.4 u; ALTOVA GmbH and ALTOVA, Inc.: 1998−2002.

(8) Ostle, B. *Statistics in Research: Basic Concepts and Techniques for Research Workers*; Iowa State University Press: Ames, IA, 1958.

(9) Cox, J. D.; Pilcher, G. *Thermochemistry of Organic and Organometallic Compounds*; Academic Press: London, 1970.

(10) *NIST/TRC Thermodynamic Tables. Hydrocarbon and Non-Hydrocarbon Series*; Thermodynamics Research Center, National Institute of Standards and Technology: Boulder, CO, 1942−2003.

(11) *International Data Series, Selected Data on Mixtures Ser. A. Thermodynamic properties of nonreacting binary systems of organic substances*; Thermodynamics Research Center, National Institute of Standards and Technology: Boulder, CO, 1979−2003.

(12) Descriptions of TRC Books are available on the Web at the following address: http://www.trc.nist.gov/publications/publications.htm.

(13) 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.

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

(15) Frenkel, M. Dynamic Compilation: A Key Concept for Future Thermophysical Data Evaluation. In *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; Gaithersburg, MD, 2001; pp 83−84.

(16) Duarte-Garza, H. A.; Magee, J. W. Isochoric $p-\rho-T$ and Heat Capacity C_v Measurements on $\{xC_3H_8 + (1-x)i\text{-}C_4H_{10}, x \approx 0.7, 0.3\}$ from 200 to 400 K at Pressures to 35 MPa. *J. Chem. Eng. Data* **1999**, *44*, 1048−1054.

(17) Simirsky, V. V.; Kabo, G. J.; Frenkel, M. L. Additivity of the Enthalpies of Formation of Urea Derivatives in the Crystalline State. *J. Chem. Thermodyn.* **1987**, *19*, 1121−1127.
(18) 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.
(19) http://www.trc.nist.gov.
(20) Marsh, K. N. New Process for Data Submission and Dissemination. *J. Chem. Eng. Data* **2003**, *48*, 1.

Received for review May 6, 2003. Accepted May 8, 2003.

JE034088I