Windows-Based Guided Data Capture Software for Mass-Scale Thermophysical and Thermochemical Property Data Collection[†]

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

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

Received May 7, 2002

Guided data capture software (GDC) is described for mass-scale abstraction from the literature of experimental thermophysical and thermochemical property data for organic chemical systems involving one, two, and three components, chemical reactions, and chemical equilibria. Property values are captured with a strictly hierarchical system based upon rigorous application of the thermodynamic constraints of the Gibbs phase rule with full traceability to source documents. Key features of the program and its adherence to scientific principles are described with particular emphasis on data-quality issues, both in terms of data accuracy and database integrity.

I. INTRODUCTION

The Thermodynamics Research Center (TRC) rejoined the National Institute of Standards and Technology (NIST) in 2000 with a goal of capturing from the world's literature essentially all experimental data available for thermophysical and thermochemical properties of organic chemical compounds. The purpose of this comprehensive collection is to serve as the basis for implementation of the Dynamic Data Evaluation concept.[1] The Guided Data Capture software (GDC) described here is key to the achievement of these goals.

The enormous growth of published thermophysical and thermochemical property data (doubling almost every 10 years) makes it practically impossible to use traditional (static) methods of data evaluation. The new concept of Dynamic Data Evaluation requires a large electronic database capable of storing essentially all of the published "raw/observed" experimental data with detailed descriptions of metadata and uncertainties. The combination of this electronic database with artificial intellectual (expert-system) software provides the means to generate recommended property values dynamically or "to order". This concept contrasts sharply with static compilations, which must be initiated far in advance of need. Capture of metadata and uncertainties for the "raw/observed" values allows propagation of reliable data-quality limits to the recommended values and, subsequently, to all aspects of chemical process design.

Establishment of a comprehensive data depository is one of the major challenges in implementation of the Dynamic Data Evaluation concept. The SOURCE data system[2,3] was designed and built to be such a depository for experimental thermophysical and thermochemical properties for organic chemical compounds reported in the world's scientific literature. The scope of the data system includes more than one hundred defined properties for pure compounds, binary and ternary mixtures, and reacting systems. The SOURCE now contains over 1 000 000 numerical values for roughly 17 000 pure compounds, 10 000 binary and ternary mixtures, and 4000 reaction systems. Though extensive, a major data-capture program has been initiated at NIST to make this collection comprehensive. To this end, NIST/TRC has established a Data Entry Facility as the focal point for its data collection projects.

Personnel of the Data Entry Facility are responsible for managing all contributions to the SOURCE including those from in-house compilers and from NIST/TRC collaborators worldwide. TRC operates a large in-house data-capture effort staffed chiefly by undergraduate students of chemistry and chemical engineering. Collaborators from outside NIST/TRC are involved with focused data-capture projects such as those related to specific compound types, properties, lingual sources, or contributions to the TRC Tables project.[4] Expansion of these collaborations to include authors of articles published in major peer-reviewed journals is in progress, as indicated in the recent editorial in the Journal of Chemical and Engineering Data.[5]

Information from original data sources is not entered directly into the SOURCE but is captured or "compiled" in the form of batch data files (coded ASCII text). This allows application of extensive completeness and consistency checks during the capture process before the data are loaded into the SOURCE. Due to the complexity of the properties and chemical systems involved, extensive expertise has traditionally been required for data compilation. Moreover, expertise in data and measurements is needed to assess uncertainties for each property value. In establishment of the Data Entry Facility two major concerns were identified: (1) how to ensure quality of captured information with technically sound but inexperienced data compilers and (2) how to minimize errors before the data are introduced into the SOURCE. To meet these goals, interactive Guided Data Capture software (GDC), written in Microsoft Visual Basic, was developed.

* Corresponding author e-mail: chirico@boulder.nist.gov.

[†] Contribution of the National Institute of Standards and Technology. Not subject to copyright in the United States. Products or companies are named solely for descriptive clarity and neither constitute nor imply endorsement by NIST or by the U.S. government.

10.1021/ci025534t This article not subject to U.S. Copyright. Published 2003 by the American Chemical Society
Published on Web 11/13/2002

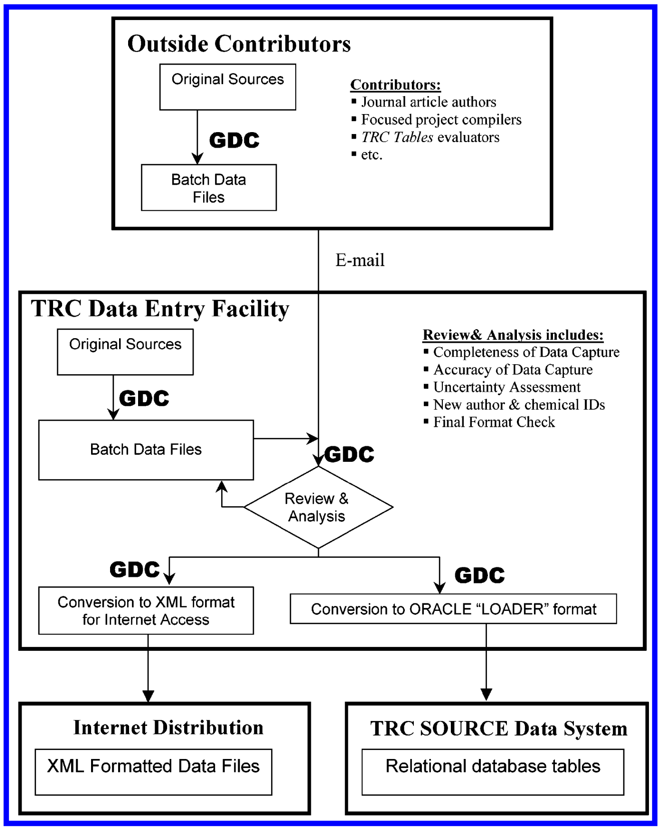


Figure 1. Applications of Guided Data Capture software (GDC) within the context of data capture and operations of the NIST/TRC Data Entry Facility.

The program guides data capture and provides convenient review and editing mechanisms. Undergraduate students involved in in-house data capture played, and continue to play, a key role in the testing of the GDC.

Figure 1 shows the applications of the GDC in the context of data management within the TRC Data Entry Facility. Although direct entry of data into the SOURCE data system is possible through the ORACLE Forms utility, this option is only practical for use in manual incorporation of minor data corrections. Data capture through batch data files is more appropriate for the distributed nature of the capture process and review procedures. Batch data files are sent to the TRC Data Entry Facility or generated in house, where they are reviewed and processed, and their data are added to the SOURCE. The GDC is the tool used to prepare, review, and edit the batch data files.

With the development of collaborations with major peer-reviewed journals for the capture of experimental data as they are published, an additional role for the GDC evolved. In addition to the creation of batch data files for loading into the SOURCE archive, the GDC simultaneously creates a separate text document coded in XML[8] format for easy access and use by the general scientific and data management community. (The ThermoML formats, designed for XML representation of thermophysical and thermochemical property data, are currently being developed at TRC in cooperation with industry.) These formatted text documents will be made available on the Internet together with a full description of the XML definitions and schema. This output from the program is indicated also in Figure 1.

In a recent paper,[7] TRC data-quality-assurance (DQA) policies were detailed with a full description given of the approach applied to the SOURCE data system. The paper described DQA policies related to six steps: (1) literature collection, (2) information extraction, (3) data-entry preparation, (4) data insertion into the SOURCE, (5) anomaly detection, and (6) database rectification. The initial steps (1−4) can be very labor intensive and represent key components of the entire data-system operation. These steps are the focus of the present paper. Aspects of steps 5 and 6 were discussed in the earlier paper.[7]

The GDC serves as a data-capture expert by guiding extraction of information from the literature, ensuring the completeness of the information extracted, validating the information through data definition, range checks, etc., and guiding uncertainty assessment to ensure consistency between compilers with diverse levels of experience. A key feature of the GDC is the capture of information in close accord with customary original-document formats and leaving transformation to formalized data records and XML formats within the scope of the software procedures. It will be shown that the GDC completely relieves the compiler of the need for knowledge related to the structure of the SOURCE data system or XML formats, thereby eliminating common errors related to data types, length, letter case, and allowable codes. The users of the GDC are scientists with varying levels of experience but with competence in the fields of chemistry and chemical engineering. Some information important for critical data evaluation, but rarely reported in the original publications, is not captured with GDC by design. This information should be taken into account during the data normalization stage of implementation of the Dynamic Data Evaluation concept.[1]

GDC was developed to serve as a powerful and comprehensive tool to be used for both TRC in-house data capture operations as well as a data-collection aid for authors of scientific and engineering publications. The software is available for free downloading via the TRC Web site (www.trc.nist.gov). Comprehensive documentation for the software is included.

The present paper describes key features of the GDC. Its adherence to scientific principles is described with particular emphasis on data-quality issues. Features will be described that can readily detect inconsistencies and errors in reported data (erroneous compound identifications, typographical errors, etc.), resulting in improved integrity of the captured data over that given in the original sources.

II. ALGORITHMIC STRUCTURE

The general structure of the information to be captured is based upon answers to the following sequential queries. Each step must be complete before access to the next is possible.

(1) What is the bibliographic source?

(2) What chemical compounds were studied?

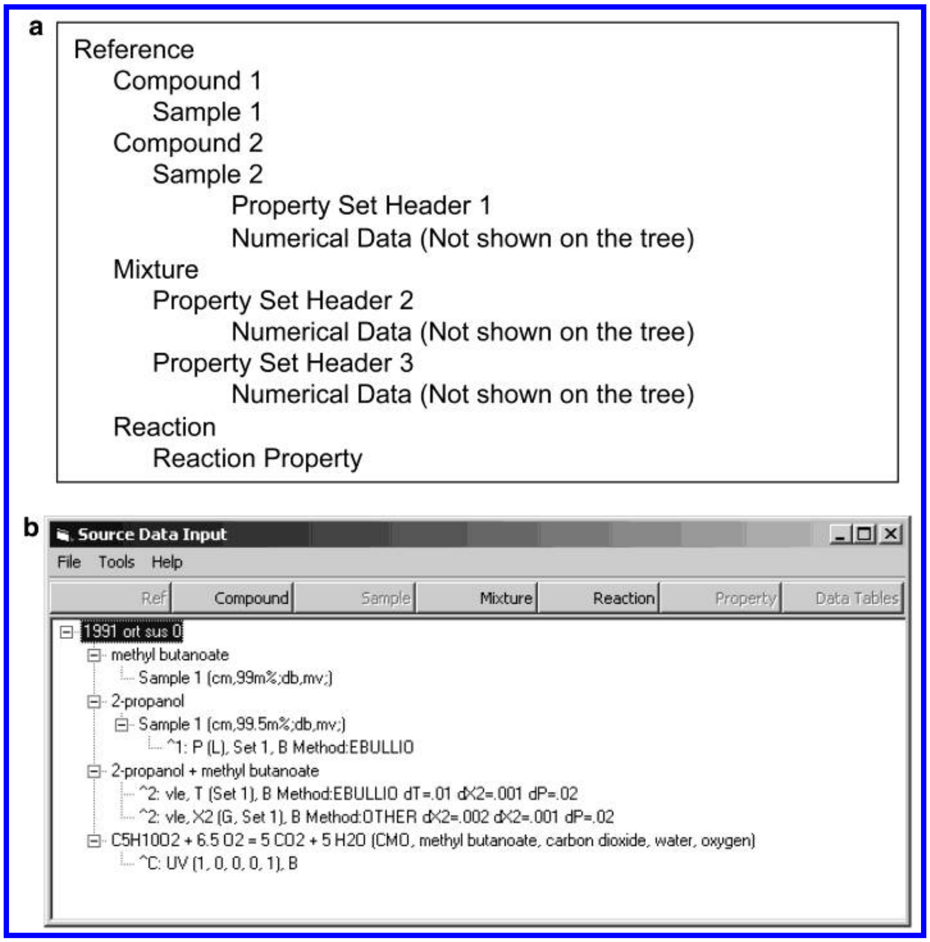
(3) What was the nature (source and purity) of the particular chemical samples?

(4) What mixtures or reactions involving the samples were studied?

(5) What properties were measured?

(6) How were the properties measured?

(7) What were the numerical values and precisions obtained?



```
Reference
    Compound 1
        Sample 1
    Compound 2
        Sample 2
                Property Set Header 1
                Numerical Data (Not shown on the tree)
    Mixture
        Property Set Header 2
                Numerical Data (Not shown on the tree)
        Property Set Header 3
                Numerical Data (Not shown on the tree)
    Reaction
        Reaction Property
```

```
1991 ort sus 0
    methyl butanoate
        Sample 1 (cm,99m%;db,mv;)
    2-propanol
        Sample 1 (cm,99.5m%;db,mv;)
            ^1: P (L), Set 1, B Method:EBULLIO
    2-propanol + methyl butanoate
        ^2: vle, T (Set 1), B Method:EBULLIO dT=.01 dX2=.001 dP=.02
        ^2: vle, X2 (G, Set 1), B Method:OTHER dX2=.002 dX2=.001 dP=.02
    C5H10O2 + 6.5 O2 = 5 CO2 + 5 H2O (CMO, methyl butanoate, carbon dioxide, water, oxygen)
        ^C: UV (1, 0, 0, 0, 1), B
```

Figure 2. (a) Structure of a batch data file. (b) A screen capture of the tree representation within the GDC.

In contrast to the relational SOURCE data system, batch data files store information in a strictly hierarchical manner, as represented in Figure 2a. The leading record is the data source ("reference") identification with all following records associated with the identified source. The next level involves "compound", "mixture", and "reaction" specifications. For each compound, there are one or more sample descriptions including the sample source, the purification method, and the final purity along with the method of purity determination. Samples, mixtures, and reactions have associated measured properties with numerical values. In most cases, property data sets within the batch data file consist of a "header" and a table of numerical values for state variables, properties, and uncertainties. The header includes all metadata required to define the property and experimental method used as well as to give meaning to the table of numerical values through definition of variables and units.

The compiler's main interactions with the GDC involve a navigation tree, which provides a visual representation in accord with the hierarchical structure of the batch data file as it is created. A typical tree structure is shown in Figure 2b and represents the information shown at the top of the figure. Each node of the tree corresponds to a record in the batch data file structure. Management of records including deletion, addition, and editing is accomplished through interactions with the navigation tree. Numerical values are not shown explicitly in the tree, but may be accessed through the property-specification nodes.

III. CHECKS AND CONSTRAINTS IN THE CAPTURE PROCESS

Checks and constraints during data capture are demonstrated first with the example of a source containing experimental-property data for a pure compound. The sequence for capture of pure-compound data is shown as a flowchart in Figure 3. Key aspects of each step are described in the following paragraphs with examples of typical checks and constraints. Maintenance of the hierarchical structure and completeness of the captured information is ensured at each step.

Lists of established field values (journal title abbreviations, compound identifiers, properties, units, phases, experimental methods, etc.) are stored in a local database, which is a part of the GDC. Selection of field values by the data compiler is achieved through single-value or multiple-selection lists of the pre-defined values, which prevents many simple errors. All pre-defined lists are prioritized to speed access. Keyboard input is never used for direct input of coded information, which eliminates typographical errors.

A filled form is accepted by the software only after three checks are successful. The first is a check for completion of all required input elements. Required elements correspond to fields, such as identification of many state variables or phases, which for the purpose of data integrity and completeness, cannot be completed with default values. The second check is for valid numerical values entered through the keyboard for properties, state variables, or constraints.

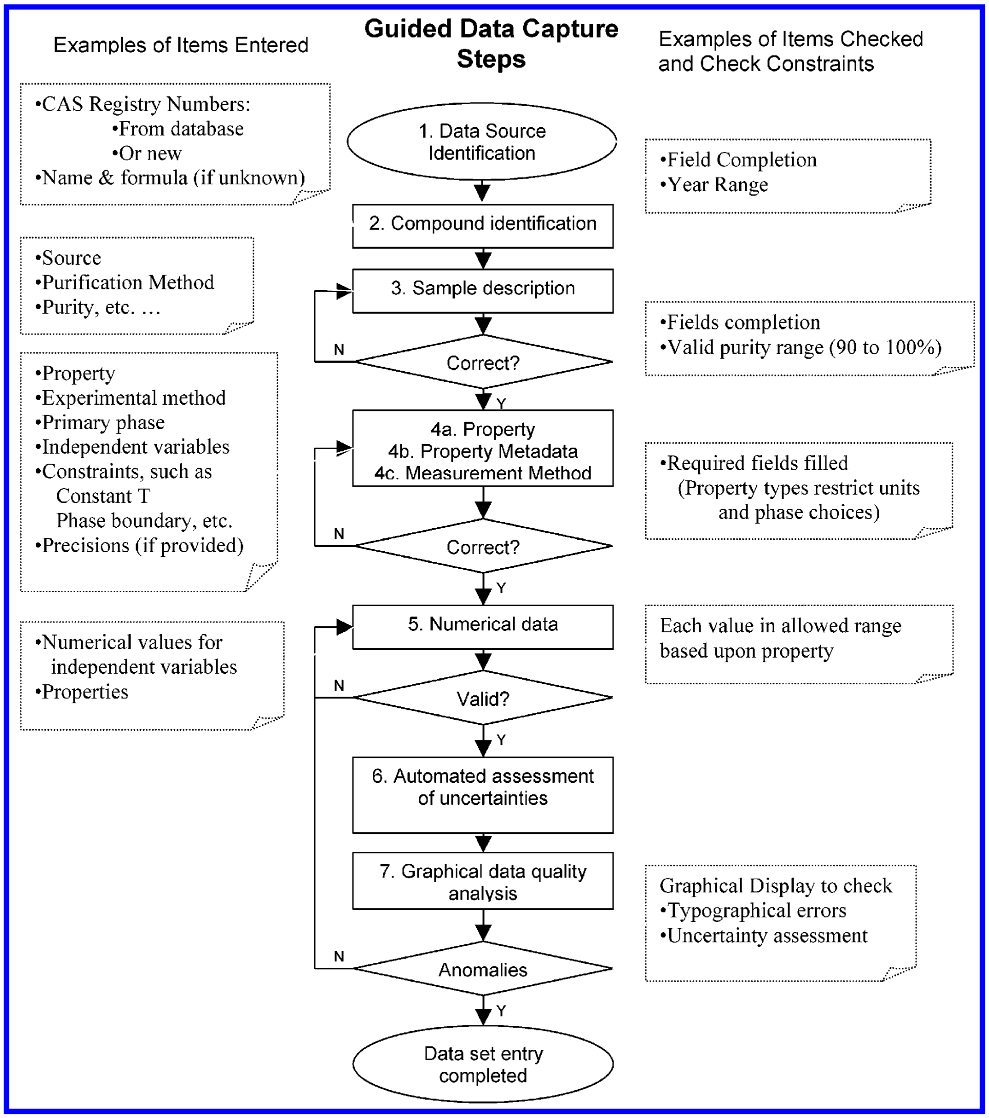


Figure 3. Sequence of operations for data capture in GDC.

"Constraint" in this context refers to an experimental constraint, such as constant temperature or constant mixture composition. The third check is a test for internal consistency and is described below. To provide visual confirmation of acceptance, a new tree node appears on the main screen after successful addition.

STEP 1 in the GDC flowchart (Figure 3) involves capture of the data-source information including standard bibliographic items (year, journal name, page, etc.), plus the document type (journal article, report, thesis, etc.), the title, all author names, key words, and abstract, if available. Authors' names and journal-title abbreviations are selected from extensive pre-defined lists or added manually, if not present. Completion of mandatory fields and the year range (1700 to present) are checked prior to acceptance.

STEP 2 involves identification of chemical compounds. Despite efforts by the International Union of Pure and Applied Chemistry and other standards organizations to develop consistent procedures for naming of chemical compounds, a wide variety of naming conventions, including common and trade names, are used worldwide. Consequently, misidentification of chemicals continues to be a common problem in nearly all chemical information archives. The GDC includes a local database of more than 100 000 chemical names and synonyms. In most cases, chemicals may be selected from this extensive pre-defined list based upon name (or partial name with wildcards), elemental formula, or Chemical Abstracts Registry Number (CASRN). Chemical entities are identified uniquely by CASRNs both in the SOURCE data system and in the batch data files. Temporary identification numbers are assigned automatically by GDC to new compounds for which CASRNs are not available. These numbers are easily distinguishable from CASRN by their numerical range.

Because of the existence of a variety of names for a given chemical compound, chemical synonyms are used in the local database to allow user discretion in selection of names for display in the navigation tree. For example, the common name "tetralin" can be selected rather than the more cumbersome "1,2,3,4-tetrahydronaphthalene". This allows correspondence between the name in the tree and that appearing in the source document. This feature aids in the elimination of chemical-identification errors.

Once a compound is identified fully in STEP 2, STEP 3 involves capture of the description for the particular chemical sample used in the reported measurements. This information is used in the uncertainty-assessment process described separately.[10] Except for purity values, which are captured as numerical percentages, all information related to the sample description (sample source, original purity, purification methods, final purity, and method of purity determination) is selected from lists. It is possible to enter two different final purities determined by different methods. This information is used also in the uncertainty-assessment process, where it is prioritized based on the method of determination. The sample-description form is checked for completeness and the existence of a valid purity value (greater than 90, but less than 100 percent) before acceptance. Acceptance of lower purities is possible, but confirmation of the low value is required.

Pre-defined lists are used for the selection of all experimental methods including those for sample purification, purity determination, and property determination. It is not possible for these lists to be exhaustive, and so “other experimental method” is always included as an option. If this option is selected, the user is requested to provide a brief description. These descriptions are monitored closely by personnel of the TRC Data Entry Facility as part of the review process shown in Figure 1. In this way, new methods are captured and added to the pre-defined lists, when necessary.

STEP 4 in the data-capture process involves full specification of the property measured and key information about the method used. These are closely connected and are designated in Figure 3 as “4a. Property”, “4b. Property Metadata”, and “4c. Measurement Method”. The SOURCE data system includes more than 120 thermophysical and thermochemical properties. Enforcement of scope is achieved through use of pre-defined lists for property selection. Property selection is eased through a two-step process in which properties are grouped into families such as volumetric properties, transport properties, excess properties, etc., before final selection. Selection of property metadata in STEP 4b involves selection of phases, units, constraints, and entry of constraint values.

After full specification of the property in STEPs 4a and 4b, the method used to determine the property is selected from a pre-defined list in STEP 4c. The list is restricted to those methods related to the selected property through use of the local database. For example, methods for density determination include “vibrating-tube densimeter”, “pycnometer”, “isochoric PVT apparatus”, etc. If needed, further operational details, such as those related to calibration method, are captured. As noted earlier, “Other experimental method” is a valid selection and is monitored carefully for significant new advances. Metadata include uncertainties reported for state variables (e.g., uncertainty in temperature for a vapor-pressure measurement). Capture of method information is a key component of subsequent uncertainty assessments.[8] These assessments represent an estimate of the overall data quality with all sources of uncertainty taken into account.

Units are selected from pre-defined lists that are coupled to the variable or property type (temperature, pressure, composition, etc.) through the local database. These include SI units and common engineering units (psia, Torr, BTU, calories, etc.). Capture of property values in the same units as the original document minimizes conversion errors. Exotic units may be used, but a conversion factor to a specified SI unit is required. Errors in specification of conversion factors are minimized through property-value range checks.

The property selected is used also to limit choices for specification of phases and state variables. For example, if “vapor pressure” is selected as the property, the second phase is defined to be “gas”. These phase restrictions are important because phase designations extend far beyond simple “solid”, “liquid”, or “gas” to include multiple crystalline phases, metastable phases, liquid crystals, multiple liquid phases for mixtures, the supercritical state, *etc.*

A key feature of the GDC involves the application of thermodynamic principles in the specification of the number of state variables for all data types captured. This is achieved through strict application of the Gibbs phase rule, which also serves as the basis for the structure of SOURCE.[2,3] The Gibbs phase rule and its application within the software are described in Section IV.

After completion of STEP 4 in the data-capture process (Figure 3), creation of the numerical data “header” is complete. As noted earlier, the header includes all information required to define the property and experimental method used and gives meaning to numerical values through definition of property, method, variables, and units. Integrity of the data is enforced through capture of all information needed to define the numerical values prior to their capture.

STEP 5 in the process is capture of numerical property values. Singular numerical properties (i.e., those with no state variables) such as critical temperature or properties with a single fixed state variable, such as “normal-boiling temperature”, are entered directly on forms used for capture of metadata. All tables of values are captured in a common data-table form. The data can be pasted directly into this form from text files, many PDF or HTML (files maintained by most online scientific journals), common spreadsheet software, and OCR software translations of scanned images. The data-table form provides many useful operations not commonly supported including column transposition, conversion of a two-dimensional data matrix to one dimension (consisting of single columns for each state variable and the property), conversion of multi-column tables to a single-column layout, change of sign, etc. These operations greatly speed data capture and minimize the addition of typographical errors.

STEP 6 in the data-capture process involves a preliminary assessment of uncertainties for the numerical property values based upon the captured metadata including the experimental method, property precision, precision of independent variables, sample purity, etc. This topic will be discussed in a separate publication.[8]

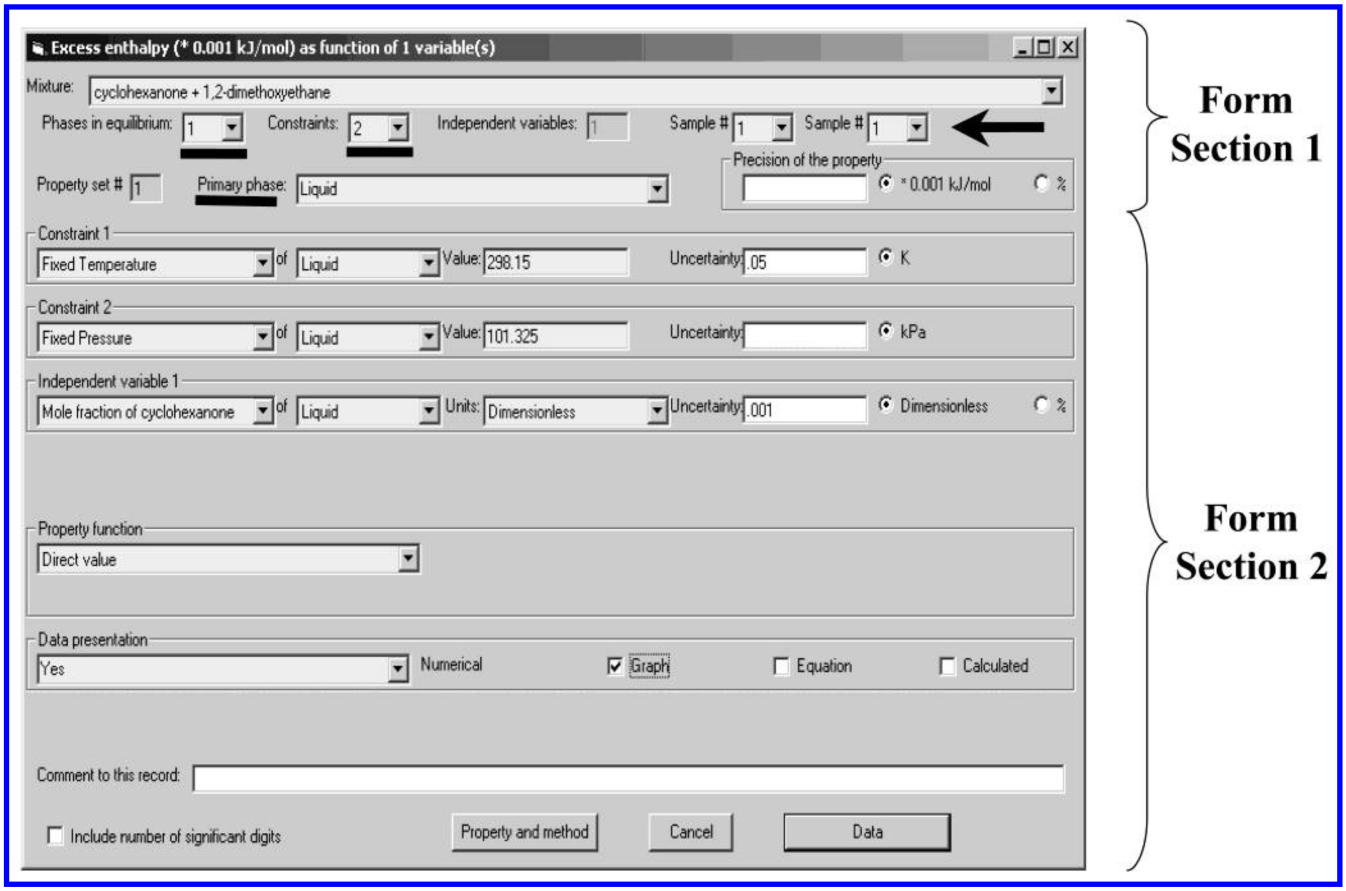
The final step, STEP 7, is graphical data-quality analysis. This step includes automatic plotting of properties involving more than two numerical values and is a key component in the assurance of data quality (Section VI).

IV. APPLICATION OF THE GIBBS PHASE RULE IN GUIDED DATA CAPTURE

The structure of the SOURCE data system[2] is based upon a fundamental relationship of thermodynamics: the Gibbs phase rule. The phase rule is a simple algebraic relationship

$$\nu = n + 2 - p - c \qquad (1)$$

which defines the number of state variables ν based upon the number of chemical components n, the number of phases



Excess enthalpy (* 0.001 kJ/mol) as function of 1 variable(s)

Mixture: cyclohexanone + 1,2-dimethoxyethane

Phases in equilibrium: 1 Constraints: 2 Independent variables: 1 Sample # 1 Sample # 1

Property set # 1 Primary phase: Liquid

Precision of the property: ● * 0.001 kJ/mol ○ %

Constraint 1: Fixed Temperature of Liquid Value: 298.15 Uncertainty: .05 ● K

Constraint 2: Fixed Pressure of Liquid Value: 101.325 Uncertainty: ● kPa

Independent variable 1: Mole fraction of cyclohexanone of Liquid Units: Dimensionless Uncertainty: .001 ● Dimensionless ○ %

Property function: Direct value

Data presentation: Yes Numerical ☑ Graph □ Equation □ Calculated

Comment to this record:

□ Include number of significant digits

Figure 4. The Mixture Data Header Form. Section 2 does not appear before completion of Section 1. Items in Section 1, which must be completed or confirmed are indicated by the three bold lines and arrow.

in equilibrium p, and the number of constraints c, such as fixed temperature, pressure, or composition.

For properties of individual chemicals (i.e., $n = 1$), application of the Gibbs phase rule is often treated implicitly in the GDC through property definitions, which specify experimental constraints such as vapor saturation or constant pressure. Because identification of phases and constraints is much more complex for multicomponent chemical systems, the Gibbs phase rule is treated explicitly in these cases.

A sample GDC form demonstrating explicit application of the Gibbs phase rule is shown in Figure 4. The form is used to capture all metadata for a data set involving a mixture. The example shown in Figure 4 is for the property “enthalpy of mixing” (i.e., the “excess enthalpy”) for a two-component mixture. The same form is used for any two- or three-component mixture and does not depend on the property. This form is reached after specification of the components (including sample descriptions), the mixture, and the property. The steps necessary to capture required metadata prior to entry of numerical mixture data are shown in Figure 5 in a flow diagram.

Section 1 of the form shown in Figure 4 is used in the explicit capture of information required for application of the Gibbs phase rule. There are five fields in Section 1 that must be completed or confirmed, but for nearly all practical cases, this number is reduced to three. The number of components in the mixture defines n in Equation 1 and is captured automatically by the software. (The example shows a two-component mixture: $n = 2$.) With the number of components specified, choices for number of phases p and constraints c are restricted based on the phase rule (Equation 1). (For example, if $n = 2$, $p + c$ cannot exceed 4, and would equal 4 only for a system under invariant conditions, such as a quadruple point.) This allows integer p and c values to be chosen from short pre-defined lists of integers. Furthermore, selection of a value for p restricts choices for c to those which ensure compliance with the Gibbs phase rule and *vice versa*.

The third item that must be defined in Section 1 of Figure 4 is the phase associated with the property values. In the example, the association is apparent, but for multiphase compositional properties, this must be specified carefully. Fields for sample selection for each component are indicated by a bold arrow in the figure. Only rarely are properties reported for multiple samples in a given source document; consequently, these rarely require attention.

Upon completion of Section 1, the number of state variables υ is calculated automatically with the Gibbs phase rule, and Section 2 of the form is displayed with the necessary input elements. (This is shown as STEP 6 in the flowchart of Figure 5.) The elements include p phase-selection menus, c constraint specifications, and υ state variable specifications. Constraint specification requires an identifier (selected from a pre-defined list), the phase to which it is related, and its value in specified units. State variable specification requires an identifier (also selected from a menu), the phase to which it is related, and its units. Numerical values for state variables are captured in a subsequent tabular form. Control of the entire process is based on restrictions imposed by the Gibbs phase rule.

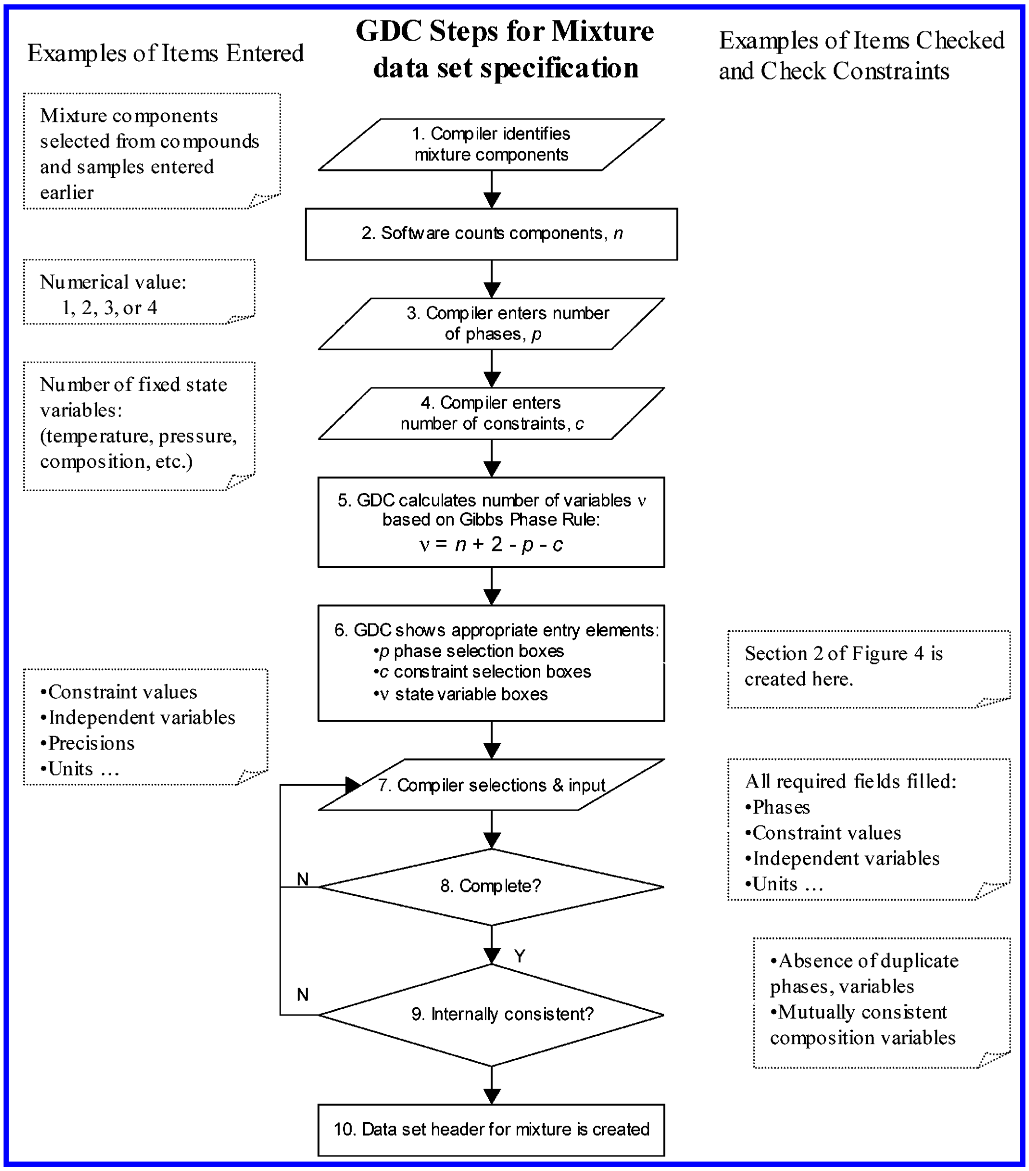

Examples of Items Entered
GDC Steps for Mixture data set specification
Examples of Items Checked and Check Constraints
Mixture components selected from compounds and samples entered earlier
1. Compiler identifies mixture components
2. Software counts components, n
Numerical value: 1, 2, 3, or 4
3. Compiler enters number of phases, p
Number of fixed state variables: (temperature, pressure, composition, etc.)
4. Compiler enters number of constraints, c
5. GDC calculates number of variables v based on Gibbs Phase Rule:
v = n + 2 - p - c
6. GDC shows appropriate entry elements:
•p phase selection boxes
•c constraint selection boxes
•v state variable boxes
Section 2 of Figure 4 is created here.
•Constraint values
•Independent variables
•Precisions
•Units …
7. Compiler selections & input
All required fields filled:
•Phases
•Constraint values
•Independent variables
•Units …
N
8. Complete?
Y
N
9. Internally consistent?
•Absence of duplicate phases, variables
•Mutually consistent composition variables
10. Data set header for mixture is created

Figure 5. Steps in the creation of a data set header for a mixture within the GDC. The data set header includes all information needed to give meaning to the numerical property values.

Application of this principle ensures data integrity and greatly eases the data-capture process through flexible but clearly defined requirements.

Names of chemicals identified in the mixture specification are inserted automatically into compositional variables in the “independent variable” and “constraint” lists (shown in Section 2 of Figure 4). This allows identification through readily understood phrases such as, “mole fraction of cyclohexanone”, rather than, “mole fraction of component 1”, thus minimizing common errors caused by confusion of components. This and other features such as flexibility in compound name displays and flexibility in units for property values demonstrate the underlying design concept of close correspondence between the original source and the captured information. All conversions to uniform naming conventions, standardized units, etc., are transparent to the data compiler.

Internal consistency checks are made by the GDC prior to acceptance of the mixture-property form (Figure 4). Checks are made for (a) a match between the number of phases listed in Section 1 and the number of distinct phases listed in Section 2; (b) the erroneous selection of the same state variable twice, if two or more state variables are required; and (c) the selection of different composition types (e.g., mole fraction, mass fraction, molarity) in the same data set for a ternary mixture. Condition (c) is allowed for special cases only with confirmation. If any of the listed anomalies is detected, the data-capture process is returned to the inconsistent element for confirmation or correction.

V. GUIDED DATA CAPTURE FOR VAPOR−LIQUID EQUILIBRIUM (VLE) EXPERIMENTS

Because of their critical importance in modeling many chemical separation and purification processes, determinations of vapor−liquid equilibria for two and three-component mixtures represent some of the most commonly reported experiments in the field of chemical engineering. Capture of this information poses a particular challenge because the number of state properties measured often exceeds the number required by the Gibbs phase rule for specification of the chemical system. Consequently, the system is overdetermined and must be divided into several data sets for logical storage. An example follows.

For a binary mixture in vapor−liquid equilibrium with no constraints (i.e., $n = 2$, $p = 2$, and, $c = 0$), two independent variables must be specified as per the Gibbs phase rule (Equation 1). However, experimental results are commonly reported in terms of four quantities (temperature, pressure, liquid composition, and vapor composition). Only three of these are necessary to fully define the chemical system. Furthermore, designation of two of these quantities as “variables” and the third as a “property” is an arbitrary distinction.

The GDC simplifies VLE data capture through use of a special data table, which allows capture of VLE data tables in overdetermined sets, as they are commonly reported. The complete data table from the literature source is pasted into the special data-capture table, and the property, unit, and phase for each column is specified through pre-defined menus. The software analyzes the information and automatically determines how many independent data sets should be created based on the Gibbs phase rule. The software also automatically assigns “variables” and “properties”. Generally, one data set is created for each phase, and composition is taken as an independent variable. If temperature T or pressure P is not given in the table from the literature source, it is assumed that it is constant and must be entered as a constraint. When both T and P are present in the table, the GDC detects which has more distinct values, and assigns it to be a “property”, and the other is assigned to be an “independent variable”. The purpose of these designations is to optimize graphical presentation for data-quality assessment described in Section VII.

VI. GUIDED DATA CAPTURE FOR CHEMICAL REACTIONS AND EQUILIBRIA

The GDC captures experimental property information for two types of reactions. The first type involves a change in the chemical composition of the system between the initial and final states (e.g., combustion in oxygen). Typical properties captured are energies and enthalpies of reaction. The second type involves systems in which the reactants and products are both present at equilibrium concentrations. Equilibrium constants, which are convertible directly to Gibbs energies of reaction, are captured typically for these systems.

Specification of chemical reactions within the GDC is similar to that for mixtures with all reaction components selected from pre-defined lists. Enthalpies of combustion of pure compounds with oxygen form a special class of reactions because of their importance as the primary basis for enthalpy of formation derivations and their key role in the calculation of reaction enthalpies and Gibbs energies. For these key reactions, the reactants and products are often well defined based upon the elemental formula of the reacting compound. The GDC automatically completes the reaction, including the complete stoichiometry, once the compound and reaction class have been specified. This minimizes typographical errors and allows an independent check of the stoichiometry reported in the original source. To eliminate this common data-capture problem, the stoichiometry of all reactions is checked by the software before acceptance.

VII. DATA QUALITY ASSURANCE FEATURES OF THE GDC

As discussed in a recent article from our research group,[7] the term “data quality”, when applied to experimental-property databases, has two distinct attributes. The first relates to “uncertainties” assigned to numerical property values by experimentalists or professional evaluators. The second attribute refers to “data integrity” and describes the degree to which the content of the database adheres to that in the original sources and conforms to the database rules. This attribute is assessed by the extent to which stored data are complete and fully in accord with database record definitions. The following provides examples of how the GDC addresses these issues.

Enforcement of Data System Scope. This enforcement is achieved through rejection of invalid document types, data types (experimental “observed” data only), and properties. All of these items are selected from pre-defined menus, which automatically enforces scope requirements. *Experimental* methods only are associated with all numerical values, which precludes capture of values resulting from estimations, correlations, ab initio calculations, *etc.*

Enforcement of Completeness. This is achieved through use of mandatory records defined by means of standardized forms. For example, in Figure 3, it is seen that complete metadata specification must be entered, checked, and accepted prior to capture of any numerical property values. Similarly, in Figure 5, metadata for mixtures are fully captured and checked before numerical values.

As noted, a navigation tree (Figure 2) is used for visual management of information including deletion, addition, and editing. “Enforcement of completeness” is implemented through the management procedures because dependent information cannot appear beyond its scope. This means, for example, that each property value must be associated with a sample description, and that each sample description must have an associated compound, etc. Deletion of an item with dependent information, such as a sample description with associated method and property data, is not possible without confirmation. If confirmed, the selected item and all dependent information are deleted, thus ensuring the hierarchical structure of the batch data file. Also, multiple data sets for a given property are assigned unique identifiers (set numbers) automatically to avoid duplication.

Capture of “Hidden” Properties: An Automated GDC Procedure To Aid Completeness of Data Abstraction. “Hidden” properties are defined here as properties of pure compounds incorporated in data sets for mixtures, and properties of binary mixtures incorporated in data sets for ternary mixtures. (For example, properties listed in an original source for a mixture in which the mole fraction of a component equals one.) Detection of such properties without software assistance is extremely cumbersome and error prone and would require careful analysis of extensive tables of numerical values.

The procedure for detection of “hidden” properties is activated after all apparent experimental property values have been extracted from an original source. The procedure analyzes each data point in the created batch data file for all mixtures in search of conditions that would allow association of that data point with a chemical system of reduced component number. If such a condition is detected, the corresponding "reduced" system (pure component or binary mixture) is searched in the batch data file for the presence of the same data point. If not found, a new data point is added to an existing data set, or a new data set or even new binary mixture identification is created.

In addition to aiding complete capture of information, the detection of "hidden" properties has proven valuable in detection of errors related to compound identification in original sources. Authors reporting properties for mixtures often study a series of mixtures involving a common component. If the order of the component identities is erroneously inverted, this is detected readily during review of the "hidden" properties captured for the common component.

Minimization of Manual Input through Use of Pre-Defined Lists, Button Selections, Check Boxes, and Other Graphical Means, While Ensuring Input Validity. Keyboard input to the GDC is provided exclusively for entry of isolated numerical values, general comments, document titles, and new chemical and author names. Most numerical values are captured through electronic means (PDF files, spreadsheets, etc.) and rarely require manual input. All other input is accomplished through pre-defined menus, check boxes, or other controlled selection processes.

Data Validation and Range Checking. Captured data that are correctly formatted may still include erroneous numerical values. Sources of error include experimental errors, malfunction of equipment, typographical errors in publications, and errors made in data capture. Once captured, these various error types are indistinguishable. There are three kinds of numerical data checking implemented in the GDC: (1) validation, (2) range checking, and (3) visual analysis for data sets consisting of more than two property values.

Validation involves testing for compliance of property values with ranges allowed strictly by the property definition. For example, absolute temperature and pressure cannot be negative, refractive index less than 1, or mole fraction less than 0 or greater than 1. If an invalid value is detected, the data-capture form is not accepted by the software until a correction is made.

Range checking involves detection of property values that are outside of the expected range for organic compounds. Simple examples include refractive indexes greater than two, molarities greater than ten, or critical temperatures greater than 1000 K. Implementation of such checking necessarily involves a balance between effectiveness and software complexity. In development of the GDC, it was decided that anomaly detection would be based upon the captured information only. For example, correlations based upon simple compound characteristics (i.e., molecular weight and elemental formula) would be used, but correlations requiring secondary property values from external sources would not (e.g., critical temperature as a function of normal boiling point). Range checking involving compound family comparisons (i.e., alkanes, alcohols, etc.) or inter-property relationships based upon thermodynamic principles are outside the scope of the GDC. The purpose of the range checking described here is detection of clear errors involving units, compound identity, typographical errors, *etc.*

Additional checks including single-property and multiple-property consistency tests are completed at regular intervals for all numerical property values in the SOURCE data system, as described previously.[7] These extensive checks play a key role in "anomaly detection" in the SOURCE system and are outside the scope of the GDC.

The capture of experimental-method information allows for additional range checks based upon typical operating conditions for common apparatus. For example, a glass ebulliometer for the measurement of vapor pressure would be used rarely for pressures less than 1 kPa or greater than 300 kPa; a differential-scanning calorimeter (DSC) would seldom be used for temperatures less than 100 K or greater than 1000 K. Any values outside of these ranges are likely to be due to errors in unit identification such as "bar" instead of "kPa" for pressure, or degrees Celsius "°C" instead of absolute temperature "K". The quality of range checking is improved continuously in the GDC as new limiters are recognized and implemented.

Visual Analysis of Captured Data Using Graphical Representation*s*. Plotting of the captured property values is an effective tool for detection of anomalous values. Authors of original sources rarely provide plots for all experimental results. Furthermore, typographical errors, which would not be included in the authors' plots, are often introduced during the document preparation and publication process.

Visual data analysis is based upon two-dimensional plots generated automatically by the software. Simple cases involving property values with one independent variable are plotted automatically with the individual values connected by lines. Data sets with more than one independent variable can be plotted as a function of any state variable. Values for state variables that are within the defined "tolerance" range define subsets. The tolerance level (default 2.5% of the value of the independent variable associated with the subset) can be adjusted to optimize the graphical display.

Property values within each subset are connected automatically by lines in the sequence of increasing plot-variable value. Each subset is represented by a different color line in the plot. This algorithm is effective if the experimental values form a near-rectangular frame in the space of the independent variables. It is not effective in two cases: (1) when all independent variables change synchronously in each experimental series within the data set, and (2) when an independent variable has an ambiguous relation to the property (e.g., two compositions of a mixture corresponding to a same boiling temperature).

Figure 6 shows density values for a two-component liquid mixture as a function of composition. The three nearly parallel curves represent subsets for different temperatures. The plot was generated from a simple three-column table of temperature, composition, and density. A typical anomalous value, which may have resulted from experimental or typographical error, is indicated in the figure. For some properties, numerical data sets can be very large and identification of a single anomalous value can be difficult. After selection of an anomalous point in the graph, the GDC automatically returns focus to the source table and highlights the value in question. This feature greatly eases error identification and correction. Any portion of the plot can be magnified readily, which is a valuable feature in the analysis of extensive numerical data sets.

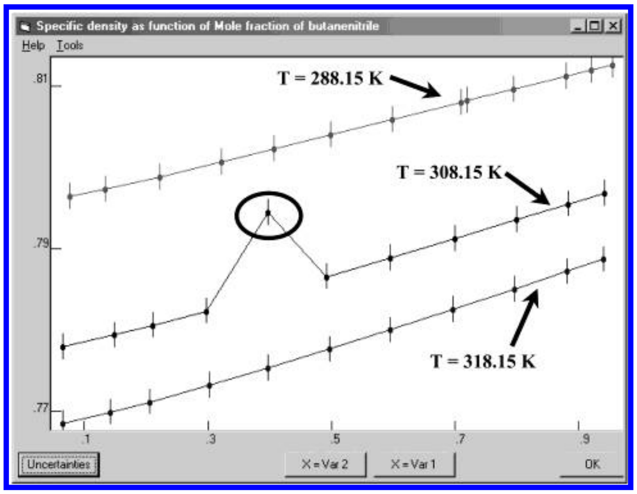


Figure 6. Graphical analysis with the GDC. A sample plot showing densities of a two-component liquid mixture at three temperatures and error bars used in assessment of uncertainty. A potential erroneous value is encircled on the figure.

A distinguishing feature of the SOURCE data system is the inclusion of estimated uncertainties[8] for all of the archived experimental property values. The complete basis for these uncertainty estimates is captured with the GDC and includes the sample purity, experimental method, and precision of the property determinations and state variables. Uncertainty assessments are completed as part of the review and checking process within the TRC Data Entry Facility at NIST (see Figure 1). The estimated error bars are shown for each data point on the software generated plots (see Figure 6) and are invaluable in confirming author estimates of precision and accuracy and as an aid in judging the validity of assessed uncertainties. To preserve traceability of the reported information, observed data inconsistencies similar to that shown in Figure 6 are recorded during the database rectification step of the implementation of the TRC data-quality assurance procedures.[7]

VIII. CONCLUSIONS

Successful mass-scale capture of thermophysical and thermochemical property data from the literature involves careful specification and validation of information from a variety of original sources describing chemical systems of various compositions with numerous properties involving extensive metadata, complex experimental conditions, and extensive numerical tabulations. The Guided Data Capture software described here makes this task not only tractable but was shown to include many features that can improve the quality and integrity of the information as it is captured from the original sources. Full implementation of this program, including its use by authors as their articles are published,[5] will provide an important service to the scientific community through generation of a massive data archive of high quality and integrity and will provide the foundation for the implementation of the Dynamic Data Evaluation concept.

REFERENCES AND NOTES

(1) Marsh, K. N.; Wilhoit, R. C. *Int. J. Thermophys.* **1999**, *20(1)*, 247−255.

(2) Frenkel, M.; Dong, Q.; Wilhoit, R. C.; Hall K. R. *Int. J. Thermophys.* **2001**, *22(1)*, 215−226.

(3) Yan, X.; Dong, Q.; Frenkel, M.; Hall, K. R. *Int. J. Thermophys.* **2001**, *22(1)*, 227−241.

(4) TRC Thermodynamic Tables − Hydrocarbons, Thermodynamics Research Center: National Institute of Standards and Technology, Boulder, CO, 1942−2002; TRC−Thermodynamic Tables − Non-Hydrocarbons, Thermodynamics Research Center: National Institute of Standards and Technology, Boulder, CO, 1942-2002.

(5) Marsh, K. N. *J. Chem. Eng. Data* **2001**, *46(1)*, 1.

(6) Murray-Rust, P.; Rzepa, H. S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 928.

(7) Dong, Q.; Yan, X.; Wilhoit, R. C.; Hong, X.; Chirico, R. D.; Diky, V. V.; Frenkel, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42(3)*, 473−480.

(8) Diky, V. V.; Chirico, R. D.; Wilhoit, R. C.; Frenkel, M. To be submitted.

CI025534T