Gas Hydrate Research Database and Web Dissemination Channel (Year 2)

Final Technical Report
October 1, 2007 to September 30, 2008

Principle Investigator: Michael Frenkel
Report Prepared By: Kenneth Kroenlein
National Institute of Standards and Technology

October 2008

DE-AI26-06NT42938

K. Kroenlein, V. Diky, R.D. Chirico, A. Kazakov, C.D. Muzny, and M. Frenkel
National Institute of Standards and Technology
Physical and Chemical Properties Division
Thermodynamics Research Center (TRC)
325 Broadway
Boulder, CO 80305-3328, USA

ABSTRACT

To facilitate advances in application of technologies pertaining to gas hydrates, a United States database containing experimentally-derived information about those materials is being developed. This work is being done by the TRC Group at NIST in Boulder, Colorado paralleling a highly-successful database of thermodynamic properties of molecular pure compounds and their mixtures and in association with an international effort on the part of CODATA to aid in international data sharing. Development and population of this database relies on the development of three components of information processing infrastructure: 1) guided data capture (GDC) software designed to convert data and metadata into a well-organized, electronic format, 2) a relational data storage facility to accommodate all types of numerical and metadata within the scope of the project, and 3) a gas hydrate markup language (GHML) developed to standardize data communications between "data producers" and "data users". All project benchmarks for the second year of this project were met. Guided data capture software designed for this project was used to collect data from an extensive collection of source material, including prototype data for all gas-hydrate target data sets. Significant modifications were made to the existing SOURCE database design to accommodate complex materials including gas hydrates. Tables for storage of gas hydrate thermophysical and crystallographic data in SOURCE were designed and are being actively populated. In collaboration with a CODATA task group, a gas hydrate mark-up language was designed to accommodate the broad range of information associated with the study of gas hydrates.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

To facilitate advances in application of technologies pertaining to gas hydrates, a United States database containing experimentally-derived information about those materials is being developed. This work is being performed by the Thermodynamics Research Center (TRC) Group at the National Institute of Standards and Technology (NIST) in Boulder, Colorado paralleling a highly-successful database of thermodynamic properties of pure molecular compounds and their mixtures and in association with an international effort on the part of the Committee on Data for Science and Technology (CODATA) to aid in international data sharing. Development and population of this database relies on the development of three components of information processing infrastructure: 1) guided data capture (GDC) software designed to convert data and metadata into a well-organized, electronic format, 2) a relational data storage facility to accommodate all types of numerical and metadata within the scope of the project, and 3) a gas hydrate markup language (GHML) developed to standardize data communications between "data producers" and "data users". In the second year of this three-year effort, all project benchmarks were met.

Preliminary development of GDC software for a Windows platform was successfully completed. This software is utilized in the data collection process to aid technically competent individuals who are not expert in data collection and measurement technology in capturing all relevant information from a literature source. This information can be used by an expert to produce optimally accurate property and uncertainty predictions.

The literature archive of gas hydrate data publications established in Phase I was bibliographically sorted and extended from approximately 3,500 unique sources to approximately 5,000 unique sources, including approximately 400 containing data determined appropriate for this effort. This feat was accomplished through efforts by the new TRC Gas Hydrate Data Entry Facility, staffed by undergraduate students in relevant technical fields. The organization and approaches taken by this group parallel those used in the TRC Data Entry Facility, a fundamental part of the primary operations of TRC.

Significant modifications were made to the existing SOURCE database design to accommodate complex materials including gas hydrates. Tables for storage of gas hydrate thermophysical and crystallographic data in SOURCE were designed and are being actively populated.

TRC Gas Hydrate Data Entry Facility used the developed GDC software for the capture of gas hydrate data sets from those articles determined to be appropriate for this data collection and dissemination effort. This includes exemplar data sets for all data targets specified during Phase I of this project. These captured datasets are being stored as formatted text files pending final review and upload.

The gas hydrate markup language (GHML) was modified to meet the needs of the broadly-based gas hydrate community while maintaining consistency with the IUPAC-standard ThermoML. This data format is actively being used in international data-sharing development efforts.

A new Sun SPARC Enterprise T5240 server has been obtained and configured to run RDBMS system Oracle 9i and MySQL to serve the data needs of this project. A modified version of SOURCE with appropriate support for the complex relationships existent within a gas hydrate has

been created. After review by a member of the TRC group, the data captured by the Gas Hydrate Data Entry Facility is actively being uploaded to this data archive.

Development of all elements of the data processing software infrastructure for gas hydrates constitutes and active population of the program database represents completion of the research activities outlined in the Statement of Work for Phase I and II of the project and provides a solid foundation for information dissemination over the World Wide Web (Phase III).

REPORT DETAILS

Introduction

The interdisciplinary field of gas hydrate research is undergoing rapid growth. Publication rates in peer-reviewed journals have displayed nearly exponential growth in the century following the discovery of hydrates in the laboratory, culminating in 3010 refereed publications in the 1990's [1]. Much of this recent growth is due to the perceived value of methane clathrate as a non-petroleum-derived large-scale energy resource [2]. Recent estimates of the world's naturally-occurring hydrated methane vary widely, ranging from 2.5×10^{15} m^3 [3] to 1.2×10^{17} m^3 [4] at STP, but the amount of organic carbon in hydrates can be conservatively estimated as a factor of two greater than the total of all remaining petroleum and natural gas reserves [5]. The remote locations where hydrate exists and the dispersed nature of the deposits has prevented prospecting at present, but the perceived potential has encouraged many nations, including Japan, Germany, India, China, Korea, Taiwan, Canada, and the United States to invest heavily in hydrate recovery programs.

Study of natural hydrate occurrences has led to the understanding that they typically exist close to their thermodynamic stability limit [4], so slight changes in ambient temperature or pressure may result in catastrophic release of methane, a potent greenhouse gas, with implications on global climate change [6] and seafloor slope stability [7]. Massive releases of organic carbon to the atmosphere and mass extinction events during the Permian Triassic [8], Late Jurassic [9], Late Paleocene Thermal Maximum [10], and other eras are often connected to the sudden release of hydrated gas.

Gas hydrate publication rates are now such that a diligent researcher could be easily overwhelmed in attempting to maintain a broad understanding of the state of the art. One solution to this difficulty is the centralization of critically evaluated data sets. A database pertaining to clathrate hydrates is being developed to facilitate understanding of naturally occurring hydrate interactions with geophysical processes, aid in the application of hydrate knowledge to technologies involved in resource recovery and storage, and support the gas hydrate research community in general. This database, whose scope includes thermodynamic and structural data, will provide to researchers the ability to submit new datasets and retrieve high quality, critically evaluated data. By establishing the hydrate database at the United States National Institute of Standards and Technology (NIST) in Boulder, Colorado, the viability of this project is secured well into the future. A critically evaluated hydrate database is essential for eliminating data redundancies, highlighting key data gaps, and providing an assurance of data quality to aid research efforts within the broader research community.

A hydrate database center has been established at NIST as part of the Thermodynamics Research Center (TRC) [11]. The existing database at the core of previous TRC activities is the NIST SOURCE Data Archival System (SOURCE) [12-13] which is the largest relational archival experimental data system, currently including more than 120 properties (including chemical structural information) for pure compounds, mixtures, and chemical reactions, with data records numbering in the millions. The gas hydrate database will be a critically evaluated dynamic data set, allowing for continuous updating and reliability analysis. TRC has extensive experience in software development for dynamic critical data evaluation, with particular application to thermophysical properties. The ThermoData Engine software [14-15], developed at TRC, is the first full-scale implementation of the dynamic data evaluation concept [16]. TRC currently has agreements with major publishers in the field of thermophysical properties for implementation of data quality assurance (DQA) procedures at the time of data submission by authors. Data files are provided by authors using guided data capture (GDC) software [17-18] for file generation.

This approach assures that submitted data are in an appropriate format [19-21] and include sufficient supporting information regarding methods and materials to allow for accurate reliability estimates. Application of this proven model is essential to the success of this hydrate database.

The data-transfer approaches associated with this data capture and storage effort are being coordinated with CODATA, which has been developing a markup language called Gas Hydrate Markup Language (GHML) [22-25] for communicating gas hydrate data throughout the research community and an international hydrate portal technology for centralized access to a number of database efforts. It is intended that all database output will be fully consistent with ThermoML and GHML. All data will be collected within an extension to the SOURCE data system operated with a relational data management system and support for internet-based dissemination. The database at NIST will include published structural and thermodynamic hydrate data. The design of the SOURCE data storage facility has been modified to accommodate all hydrate data within the scope of the current efforts. The GDC software architecture has been extended to allow capture of gas-hydrate data, and will include GHML compatibility. Data quality analysis tools in the GDC software have been extended to accommodate property data for gas hydrates.

Literature Archive and Data Capture

A primary task during Phase II of this program was the collection and characterization of a literature archive. The archive, maintained as a collection of electronic PDF files, contains approximately 5,000 unique sources, where approximately 3,500 unique sources were present in the literature archive provided to the effort by Dr. E. Dendy Sloan. Of the total archive, approximately 400 have been determined to contain data usable by this data collection effort. The archive includes peer-reviewed journal articles, technical reports, master's theses and doctoral dissertations in a number of languages and dating as far back as the 18th century.

Even rudimentary review of an archive of this size for bibliographic information is a non-trivial activity. For the task of reviewing this article set and evaluating data content, the TRC Gas Hydrates Data Entry Facility was established in January 2008, paralleling the previously established TRC Data Entry Facility. This new group includes four undergraduates in relevant technical fields from the University of Colorado at Boulder and the Colorado School of Mines who are solely allocated to this data project. Under the direction of Dr. Kenneth Kroenlein of the TRC group, the students reviewed source materials, assembled an in-house citation database in order to track documents and collected information from those files following data collection protocols established for the proven TRC Data Entry Facility.

Information from original data sources is not entered directly into the NIST SOURCE Data Archival System (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 a central repository. 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 at NIST, 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 SOURCE. To meet these goals, interactive guided data capture (GDC) software, written in Microsoft Visual Basic, was developed. 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 software.

With the development of collaborations with major peer reviewed journals for the capture of experimental data as they are published, an additional role for GDC evolved. In addition to the creation of batch data files for loading into the SOURCE, the GDC software simultaneously creates a separate text document coded in XML [19-21] format for easy access and use by the general scientific and data management community. These formatted text documents are available on the internet together with a full description of the XML definitions and schema.

TRC data-quality-assurance (DQA) policies, as they relate to a database effort such as SOURCE, can be subdivided into six steps: (1) literature collection, (2) information extraction, (3) data-entry preparation, (4) data insertion, (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 are the steps which development of GDC serves to provide expert guidance to novice data compilers. Additional detail of steps 5 and 6 have been discussed previously [27].

GDC functions to guide inexperienced but technically-competent individuals through the process of extracting 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 software 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. Thus GDC 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 GDC are scientists with varying levels of experience but with competence in the fields of chemistry and chemical engineering.

The GDC system 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 original software, without support for gas hydrate property capture, is available for free downloading via the World Wide Web [18]. Comprehensive documentation for the software is included. The GDC software has features 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. Additional information on the development of GDC can be found in the literature [17].

In order to capture experimental data sets pertaining to samples of gas hydrate, the existing GDC software required significant modification. Whereas data normally processed through GDC is either for a pure compound or a mixture of a small number of well-defined compounds in well-defined ratios, a gas hydrate is a non-stoichiometric structure where determination of crystal compositional distribution may not ever be measured but can still yield valuable data. Whereas it might be desirable to simply designate such studies as unreliable, the comparative paucity of data may preclude such a determination. The solution to this conflict was determined to be the creation of an original data structure within the GDC framework which behaves in many ways like a new compound, defined by the combination of its constituents and known thermodynamic properties. With these modifications, the GDC software supports the capture and organization of data pertaining to bulk properties (e.g. mass specific volume, thermal conductivity, heat capacity at constant pressure per unit mass, speed of sound), phase equilibrium with an arbitrary number of components and phases, crystalline structure and enthalpy of hydrate decomposition for gas

hydrates. In particular, the data structure for crystalline structure represents an entirely new development with this software as no previously existing formulation existed within the GDC software. The level of functionality thus attained represents significant progress towards a completed GDC software package for gas-hydrate data; however, experience in this area has shown that even once a piece of data capture software is completed and in use for database population, continued capture of published data sets may motivate minor modifications to aid in future efforts.

The basic tree structure of GDC data organization (Figure 1) is primarily organized around the data source document. Following from that are definitions of major chemical components in the systems presented within the citation and their specific samples with detailed purity information. A gas hydrate system is then defined by a combination of those chemical components (Figure 2) and a gas hydrate sample is defined though the association of the samples of those components as well as the conditions under which the hydrate was formed, if appropriate (Figure 3). It is only once this detailed information regarding purity on constituent compounds is defined that measured properties are entered, allowing for a detailed understanding of the resultant data reliability.

In order to guarantee a well-defined thermodynamic state and to prevent storage of dependent variables as independent, the system is constrained according to the Gibbs Phase Rule; for example, if a three-phase region is being defined in a gas hydrate sample formed from two guest molecules (Figure 4), there exist two degrees of freedom in the system and hence two dependent variables are required. The data for the newly defined system is then recorded in an internal data table (Figure 5). To prevent transcription errors on the part of the data entry technician, data is directly copied from electronic versions of the source, either obtained via electronic distribution or via text recognition software applied to digitized material. Data consistency can then be verified using native graphing capabilities (Figure 6) within the GDC software.

In order to properly characterize enthalpy of decomposition of a gas hydrate, it is necessary to have well defined ratios of host to guest molecules; for this reason, the system for such a decomposition is characterized as a physical reaction and the enthalpy of decomposition is stored as an enthalpy of reaction (Figure 7). This methodology has additional benefits in that, as the comparatively slow kinetics associated with hydrate formation and the dynamics of hydrate decomposition may yield a condition where the ambient pressure and temperature at which a study are performed do not necessarily correspond to the associated equilibrium phase boundary, such data can be accurately stored for future consideration and critical review.

For bulk property measurements, such as mass specific volume, thermal conductivity, heat capacity at constant pressure per unit mass, or speed of sound, experimental measurement techniques do not vary significantly from those implemented in studies of pure compounds. For this reason, a significant amount of parallelism was possible between the newly-developed and previously-existing treatments. In order to have a well characterized bulk measurement, we first must define the type of property being measured and the method of measurement (Figure 8) after which the conditions under which the measurements must be defined (Figure 9). The numerical data is captured and existing TRC internal methods are used to estimate the reliability of the provided data (Figure 10) and, for larger data sets, the previously mentioned native graphing capabilities can be utilized.

Characterizing crystal structure is a wholly novel addition to the GDC intended for gas hydrate data collection; in order to maintain future extensibility as well as collect detailed information

about the cage structure, information is stored regarding the crystallographic space group, all possible unit cell dimensions and both raw and processed information regarding constituent atom distribution (Figure 11). To provide a reasonable guarantee of generality and compatibility with likely crystalline structure data sets, this new data structure was modeled upon the Crystallographic Information File (CIF) data file format used prevalently within the crystallographic community for communication of experimental results [28], which is an International Union of Crystallography (IUCr) standard.

Database Architecture

The stated goal of the TRC group at NIST is the capture from the world's literature of essentially all experimental data available for thermophysical and thermochemical properties of organic chemical compounds in an automatically-interpretable form. The enormous growth of published thermophysical and thermochemical property data makes simply tracking all newly published materials an intensive task. This ostensibly comprehensive collection is to serve as the basis for dynamic data evaluation, a concept implanted by TRC in its ThermoData Engine (TDE) [14-15]. Fundamental to a large-scale data analysis effort such as this is a well-structured database complete with appropriate metadata and uncertainties.

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 NIST SOURCE Data Archival System [12-13] 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. SOURCE now contains nearly three million numerical values for many thousands of pure compounds, binary and ternary mixtures, and reaction systems. In conjunction with the expansion associated with this project, SOURCE is being migrated to a Sun SPARC Enterprise T5240 server running RDBMS system Oracle 9i and MySQL. The eventual Gas Hydrate data dissemination (scheduled to be accomplished during the third year of the project) will be established via replication of the relevant portion of SOURCE on the external server running outside of the NIST firewall, as per NIST network security policy, with free and open access. Presently, there are several Dell servers running Red Hat Enterprise Linux available for this purpose.

In designing data structures to accommodate the gas hydrate data sets, limitations associated with how complex materials were defined in the system were highlighted. In order to support these relationships as well as those of ionic liquids, stereo-isomeric mixtures and other complex samples, a new table structure has been designed. The relationships for the total gas hydrate system are shown broadly in Figure 12 and the specific details of defining a complex are shown

in Figure 13. All gas-hydrate-specific tables are denoted by the "GH" prefix. To define a chemical complex (table CMPLXID), a series of well-defined compounds (table CMPID) are associated with compositional information if appropriate through a pivot table (table CMPLXCMP). Each complex is assigned a TRC-specific identifier which is unique between both the CMPID and CMPLXID tables, allowing property data to be defined equivalently independent of the type of system it is associated with. A gas hydrate complex entry is then tied together with the literature source of its data through the unmodified, previously-existing literature reference tables in the GHSYSREF table. As purity information of feedstock is relevant to the ultimate properties of a crystal sample, that information is tied to the system for each component present in the system through a gas-hydrate specific GHSAMPLE table.

If the given study was crystallographic in nature, a corresponding table entry is made in table GHSTRUCT (Figure 14). This contains information on rudimentary crystallographic data (space group, lattice parameters) in addition to experimental conditions (system temperature, system pressure, uncertainty methodology). If additional information is provided on either interatomic spacing or Cartesian atomic distribution within the unit cell, such information is stored in tables GHSTRCUTRAW or GHSTRCUTPROC, respectively. This data structure follows the organization of a Crystallographic Information File (CIF), a standard within the crystallography community [28].

Characterizing the complex phase equilibria associated with gas hydrates and necessary to properly specify the conditions of a thermophysical measurement required significant extension from the previously existing SOURCE data storage format. Given that gas hydrate systems may contain a wildly-varying number of chemical components, the fixed table width approach previously utilized to guarantee proper system constraint becomes untenable with a gas-hydrate system. This is easily demonstrated with the application of the Gibbs phase rule to a hydrate-forming natural gas system in equilibrium with sea-water. The Gibbs phase rule states the number free parameters that must be specified in a system in order for that system to be thermodynamically well-defined is equal to two more than the difference of number of chemical components and the number of phases. For an eight-component representation of sea water and of natural gas, this three-phase condition would require fifteen data for proper constraint. Designing a single table to accommodate this set as well data sets containing two compounds and four phases would be very inefficient and would be inherently limited if a more complex set were encountered in the future.

The solution adopted by TRC in this scenario is shown in Figure 15. Each data series from a given study, defined to be a set of measurements performed by equivalent methods on a system with a prescribed set of phases present, is uniquely defined in the GHDATASETS table. Observed phases for this data set are stored in the GHPHASELST table. As nearly all data points have temperature and pressure data associated with them, the primary key for a given data point is specified in the GHTP table. Any additional compositional information for that point is stored in the GHCOMPOSITION table, which stores not only the data and uncertainty, but the compound for which and phase in which the measurement was made. Data integrity for that composition data is checked by enforcing that the referenced phases and compounds are present in the system and GHSAMPLE data provided. Property data sets, such as speeds of sound or specific heats, are stored in the GHPROP table with similar constraints. Through simple arithmetic, the thermodynamic completeness of a set can be determined.

At present, the GDC batch output files generated by undergraduate operators in the Gas Hydrates Data Entry Facility are being uploaded into this archive after each file is checked for consistency with the original source material by Dr. Kenneth Kroenlein.

Gas Hydrate Markup Language

Thermodynamic property data represent a key foundation for development and improvement of all chemical process technologies. However, rapid growth in the number of custom-designed software tools for engineering applications has created an interoperability problem between the formats and structures of thermodynamic data files and required input/output structures for the software applications. Establishment of efficient means for thermodynamic data communications is absolutely critical for provision of solutions to such technological challenges as elimination of data processing redundancies and data collection process duplication, creation of comprehensive data storage facilities, and rapid data propagation from measurement to data management system and from data management system to engineering application. Taking into account the diversity of thermodynamic data and numerous methods of their reporting and presentation, standardization of thermodynamic data communications is very complex.

A component of the work performed in the previous project year consisted of reconciling the then-existent GHML schema [22-25] and ThermoML [19-21]. These efforts resulted in a series of parallel (i.e., non-intersecting) sections by essentially independent authors that described various types of property data ("field" [23], "laboratory" [24], and "modeling" [25]), where consistency between ThermoML was generated primarily via modification of the laboratory section and the addition of citation information. The structure of ThermoML is based on rational storage of property data with the origin of the data as a major component of the organization construct. By recreating this approach within the large and varied types of data sets associated with each subcategory of information identified by the original GHML development team, consistent approaches can be used to refer to corresponding elements within the larger tree structure. Achieving consistency with ThermoML in design philosophy and content, internal consistency within the separate branches in style and nomenclature and sufficient flexibility so as to allow storage and transfer of both clathrate hydrates of natural gas and of more exotic guest molecules which are of interest for both basic research and energy storage technologies required significant restructuring of the GHML schema. This eliminates redundant information storage and excessive complexity while maintaining a well constrained system with good flexibility for future research interests. As GHML is an international effort under the auspices of the International Council for Science's Committee on Data for Science and Technology (CODATA), any proposed schema had to be approved by that body, and this happened at a meeting of the CODATA Hydrate Database Steering Committee on October 27, 2007, with the note that additional unification across the disparate GHML branches was desirable.

Examination of data sets across the range of disciplines associated with the field of "gas hydrate" studies revealed a range of data which was both unsupported and not clearly representable within the prescribed framework. In response, a modification of GHML was formulated which combined the FieldData, LabData and ModelData elements into a single DataSet element (Figure 16). Rather than specifying the structure of datasets to be encoded within the XML Schema Definition (XSD), this DataSet element specifies the encoding for metadata common to many different datasets, in the broad categories literature citation (Figure 17), investigation details (Figure 18), chemical compound information (Figure 19), and sample history (Figure 20), and then specifies the data organization of a formatted data-tuple (doubly-delimited list) through the inclusion of data labels that include appropriate data attributes to maintain data relationships

(Figure 21); for example, a mole fraction data series includes relational information to specify a compound being measured and the phase in which it was measured (Figure 22). This development has been discussed previously [29].

GHML as described is currently being used as the basis for the current web-dissemination technology development efforts underway by the CODATA gas hydrate task group.

Conclusions

Development of a database for gas hydrate physical property information has proceeded according to the timeline for the project. All milestones within Phase II of the Statement of Work were met or exceeded.

The literature archive of gas hydrate data publications established in Phase I has been extended and processed by the new-established TRC Gas Hydrate Data Entry Facility. This includes application and minor modification of the GDC software developed during Phase I, including data capture for all target data categories.

New computing hardware has been obtained and configured to serve the data needs of this project. A relational data storage facility capable of accommodating all types of numerical and metadata within the scope of the project was developed based upon extensions of the previously existing SOURCE data system has been created and is being actively populated by data collected by the TRC Gas Hydrate Data Entry Facility.

The gas hydrate markup language (GHML) was modified to meet the needs of the broadly-based gas hydrate community while maintaining consistency with the IUPAC-standard ThermoML. This data format is actively being used in international data-sharing development efforts.

Development of all elements of the data processing software infrastructure for gas hydrates constitutes and active population of the program database represents completion of the research activities outlined in the Statement of Work for Phase I and II of the project and provides a solid foundation for information dissemination over the World Wide Web (Phase III).

INDEX OF GRAPHICAL MATERIALS

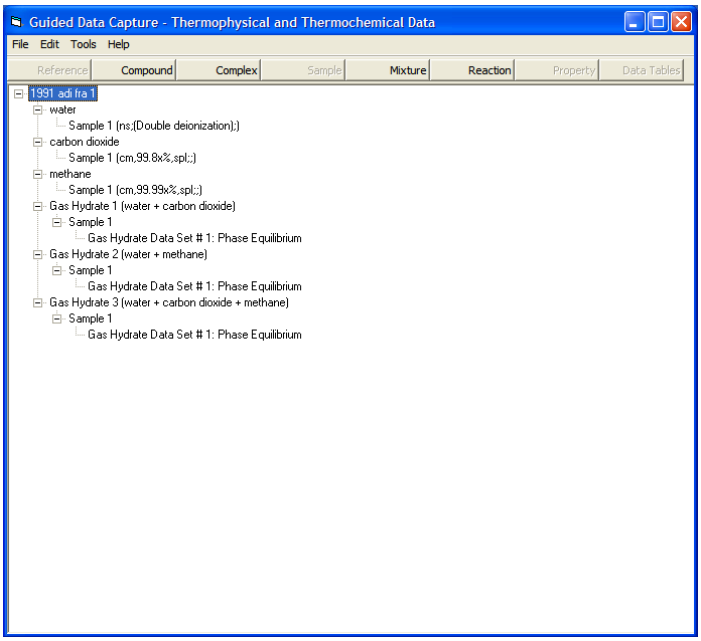

Figure 1. Screen capture of tree structure for a gas hydrate sample characterization within GDC

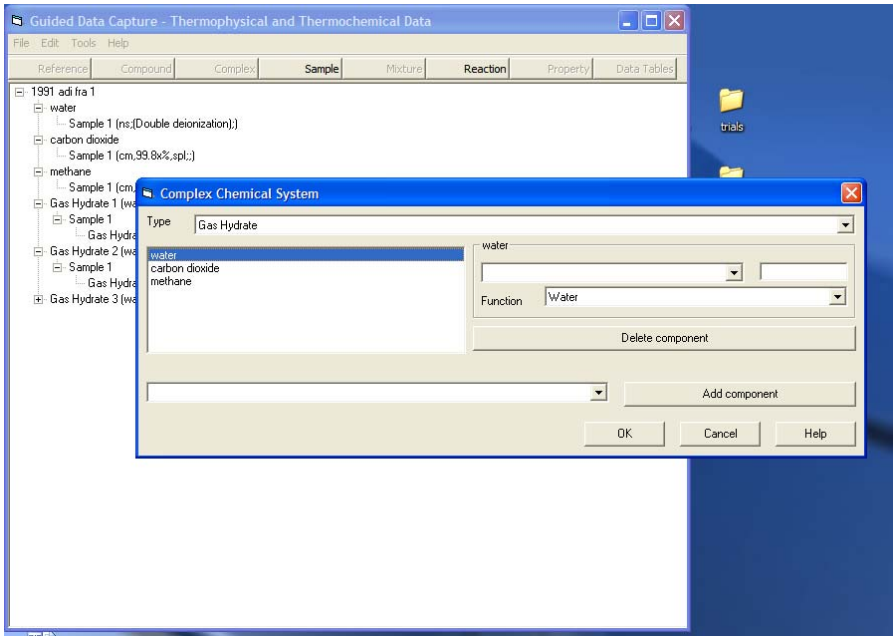

Figure 2. Screen capture of GDC dialog for definition of a gas hydrate system

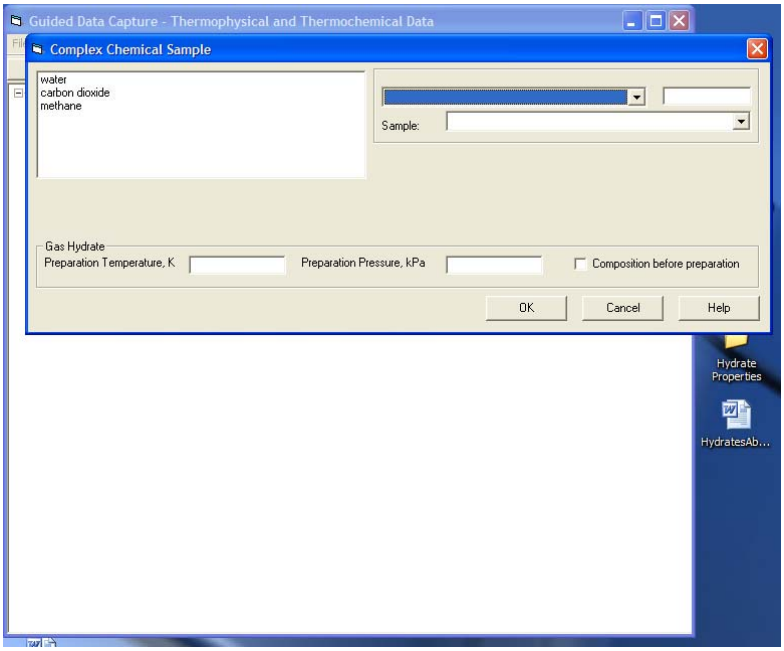

Figure 3. Screen capture of GDC dialog for definition of a gas hydrate sample

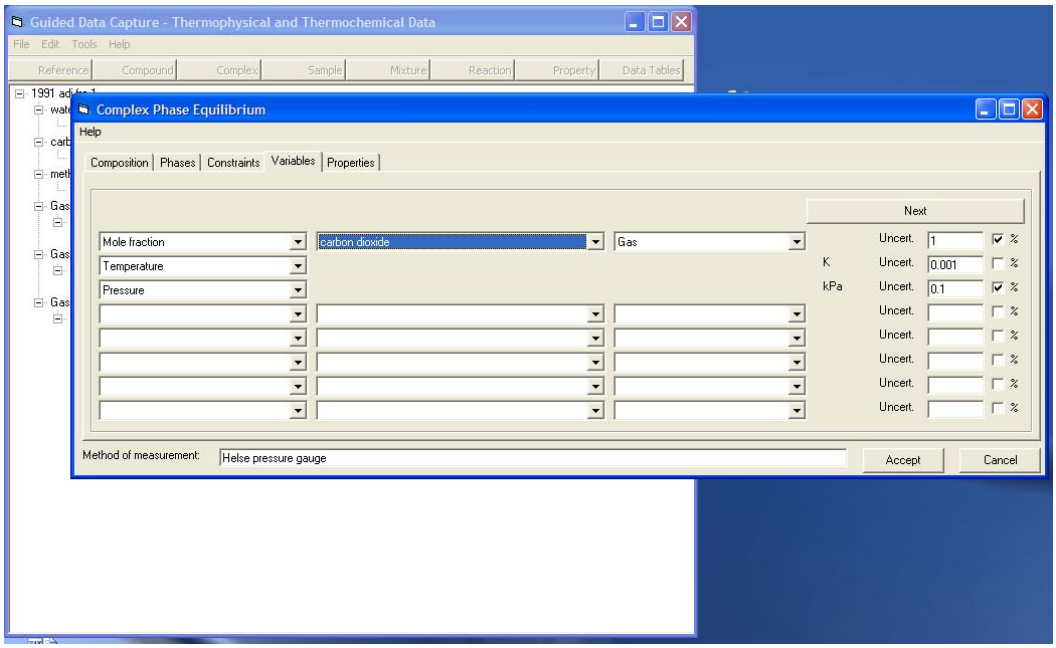

Figure 4. Screen capture of GDC dialog for defining phase equilibrium constraints and variables on a given set of phase equilibrium data

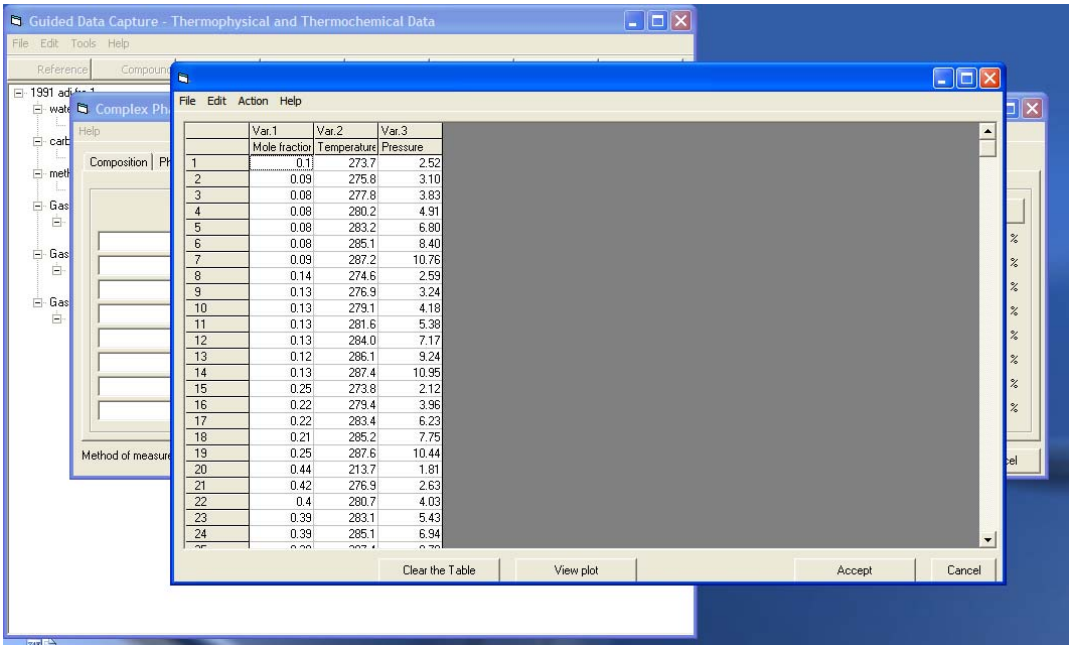

Figure 5. Screen capture of GDC dialog for entering tabulated data associated with a given set of phase equilibrium data

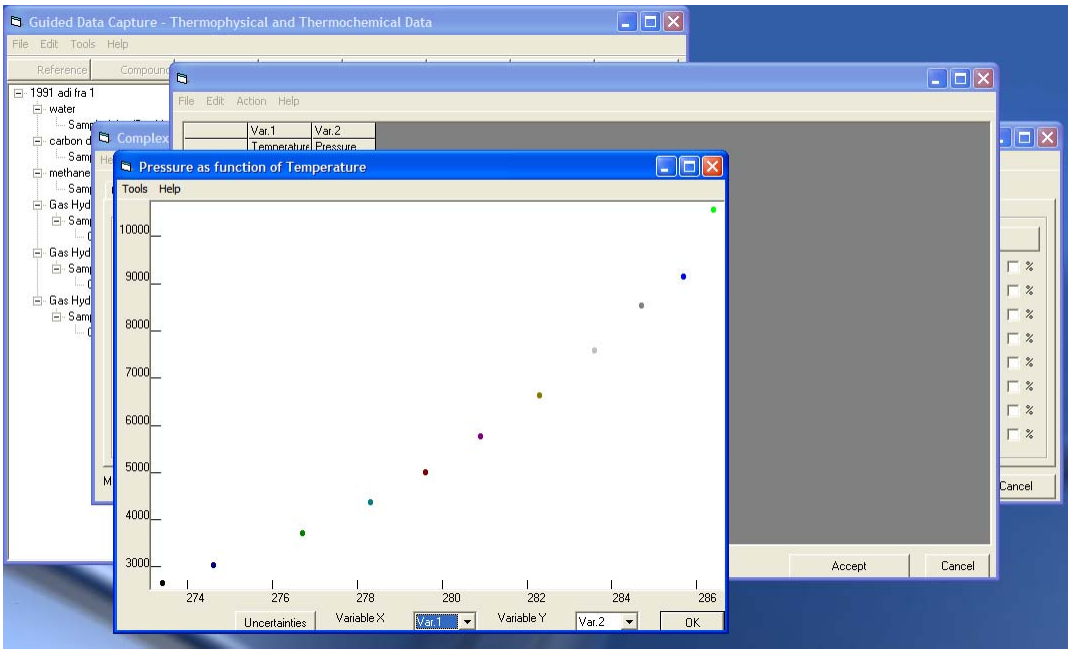

Figure 6. Screen capture of natively-generated graph of data entered into GDC tabulated data dialog

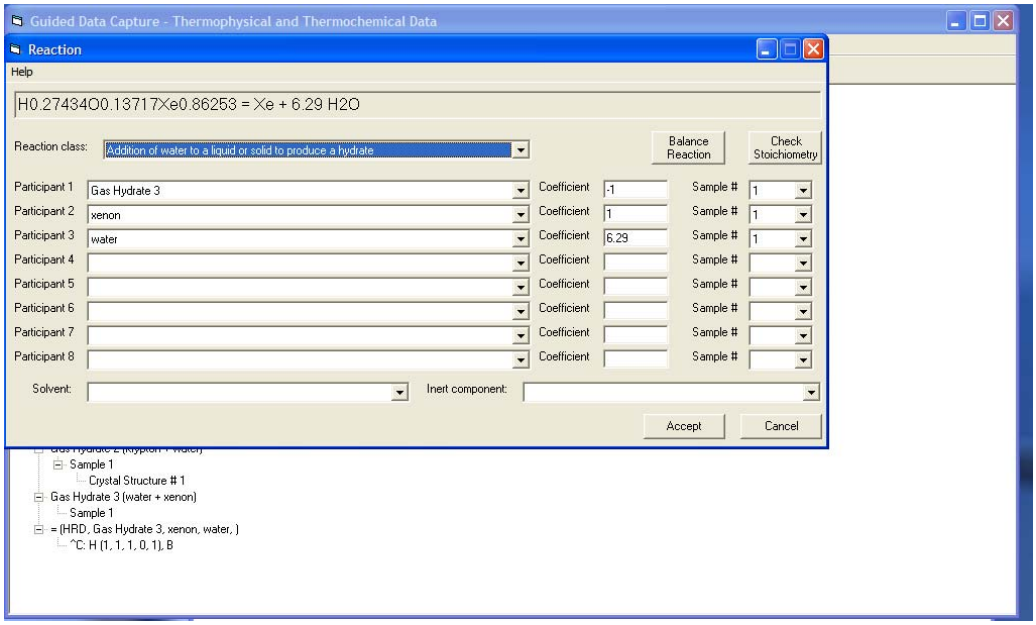

Figure 7. Screen capture of GDC dialog for defining the physical reaction associated with gas hydrate decomposition

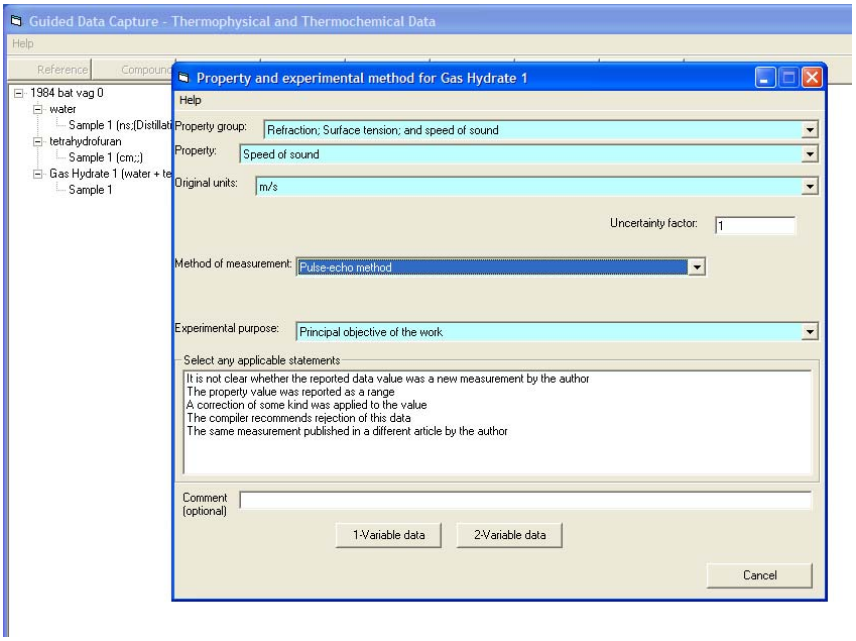

Figure 8. Screen capture of GDC dialog for defining a type of bulk measurement and the associated measurement methodology

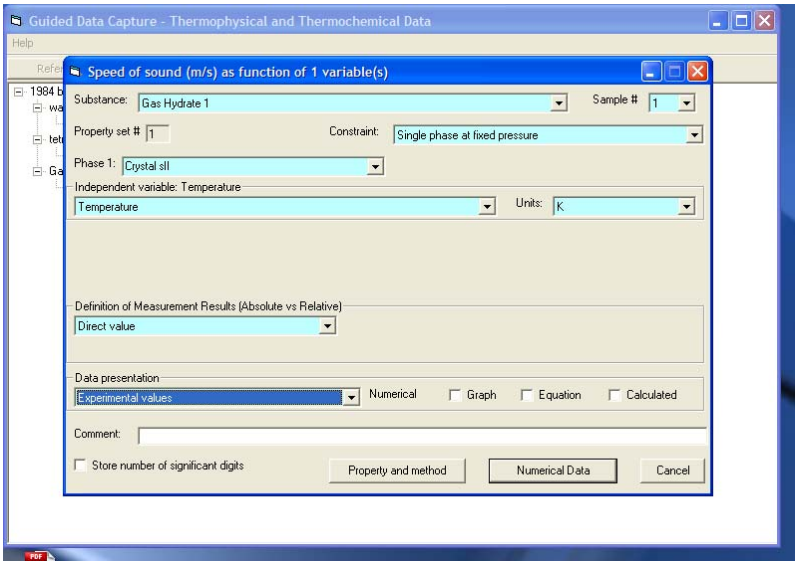

Figure 9. Screen capture of GDC dialog for defining the thermodynamic conditions under which a bulk measurement was performed

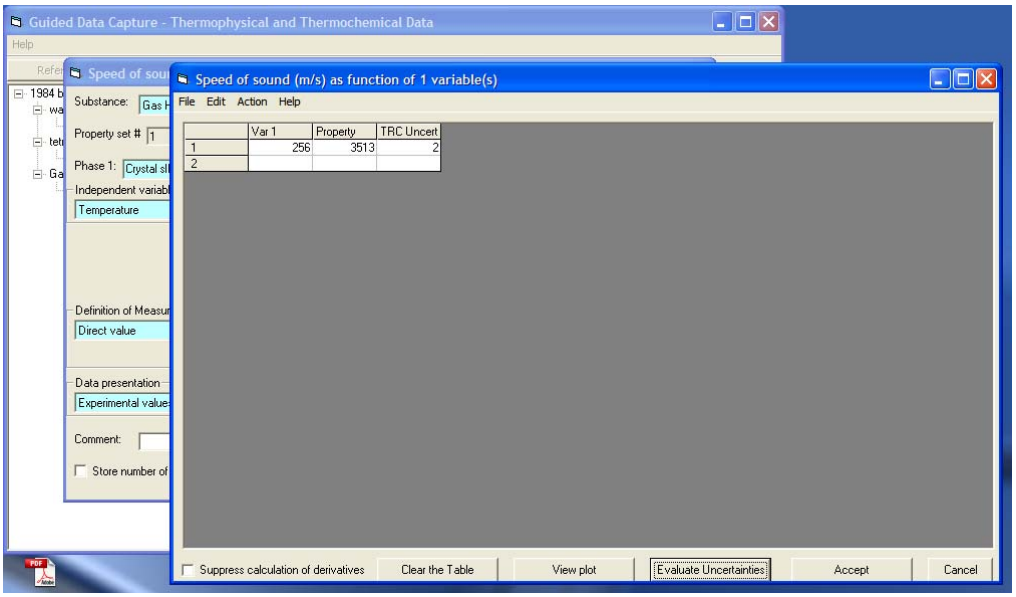

Figure 10. Screen capture of GDC dialog for entering tabulated data associated with a given set of bulk property data with automated reliability estimate

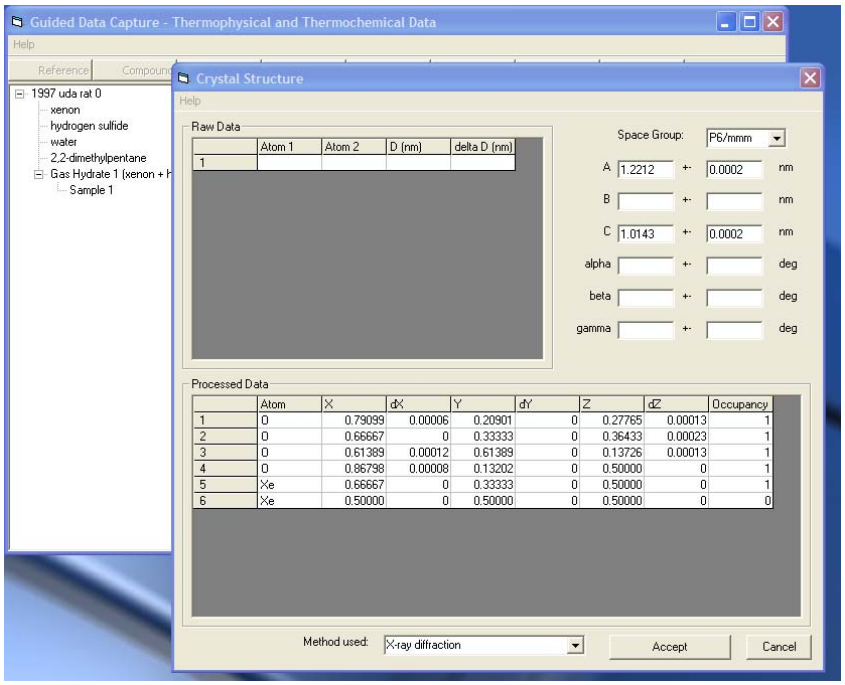

Figure 11. Screen capture of GDC dialog for storing crystallographic data, including space group, unit cell parameters and atom distribution

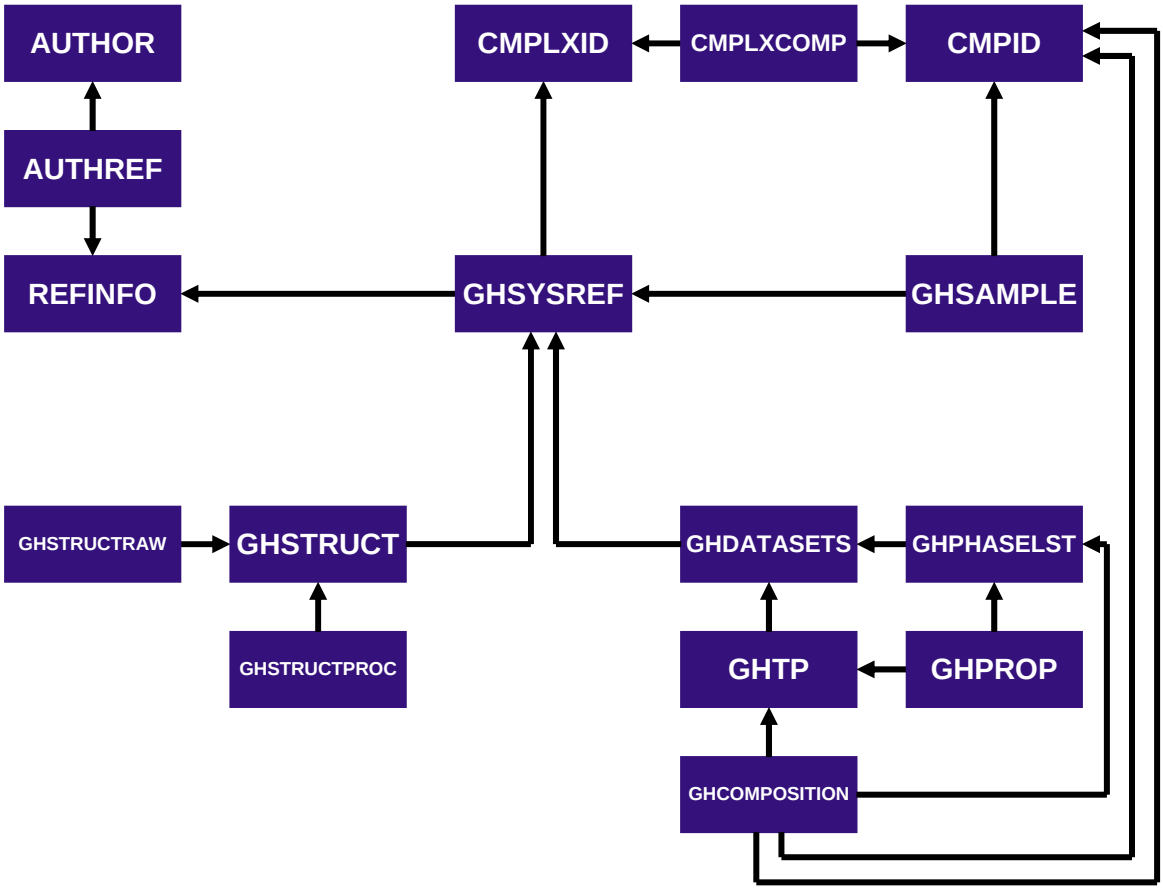

Figure 12. Schematic representation of new SOURCE table structure and gas-hydrate-relevant table substructure

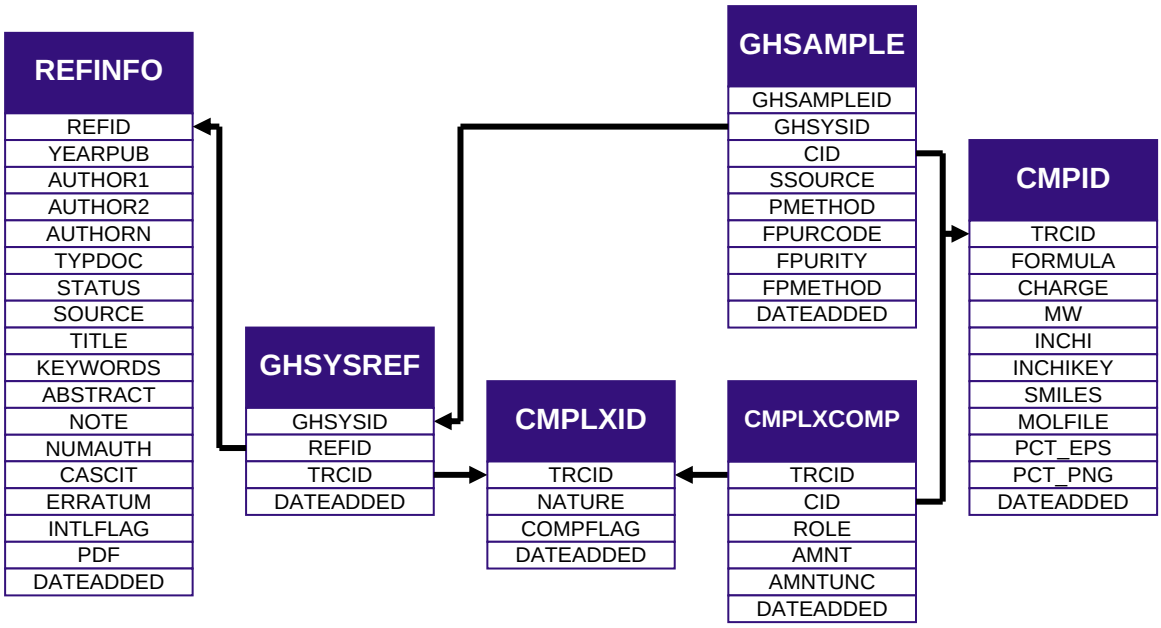

Figure 13. SOURCE tables relevant to defining a specific gas hydrate sample, dependant upon the literature source of the data, the chemical compounds present and the compositional purity of the feed materials

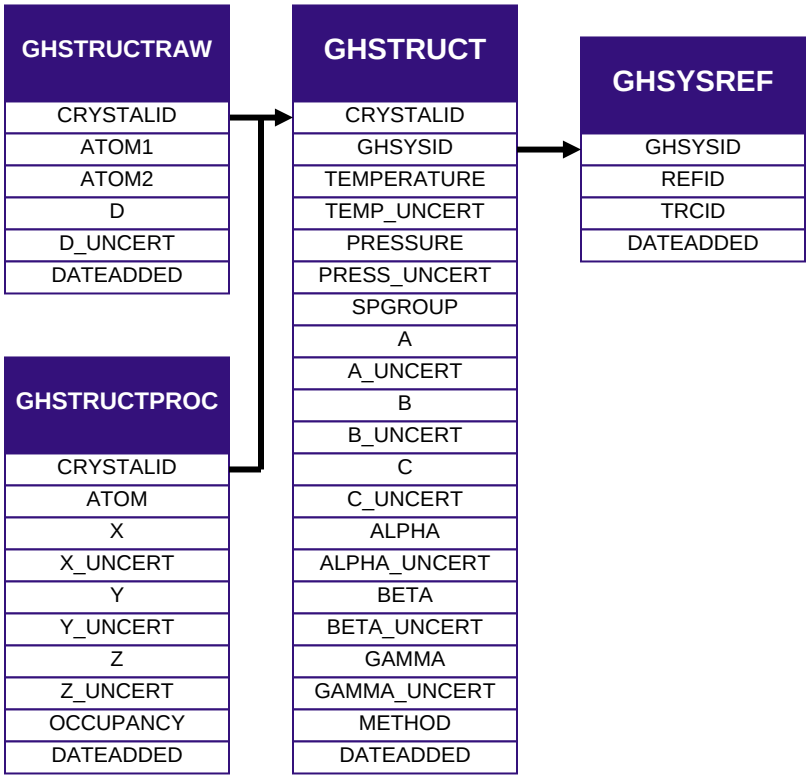

Figure 14. SOURCE tables relevant to defining data from crystallographic studies, including atomic distribution if reported

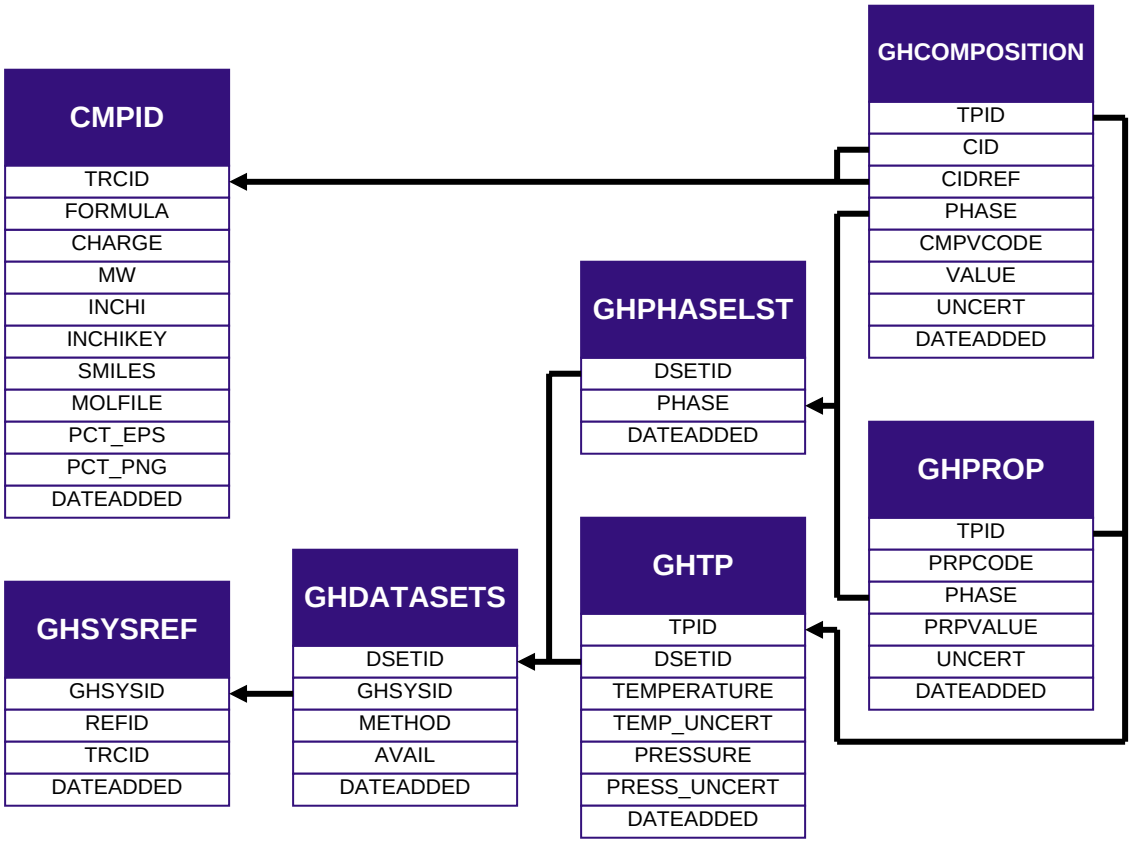

Figure 15. SOURCE tables relevant for defining thermodynamic state and property data, including temperature, pressure and compositional information

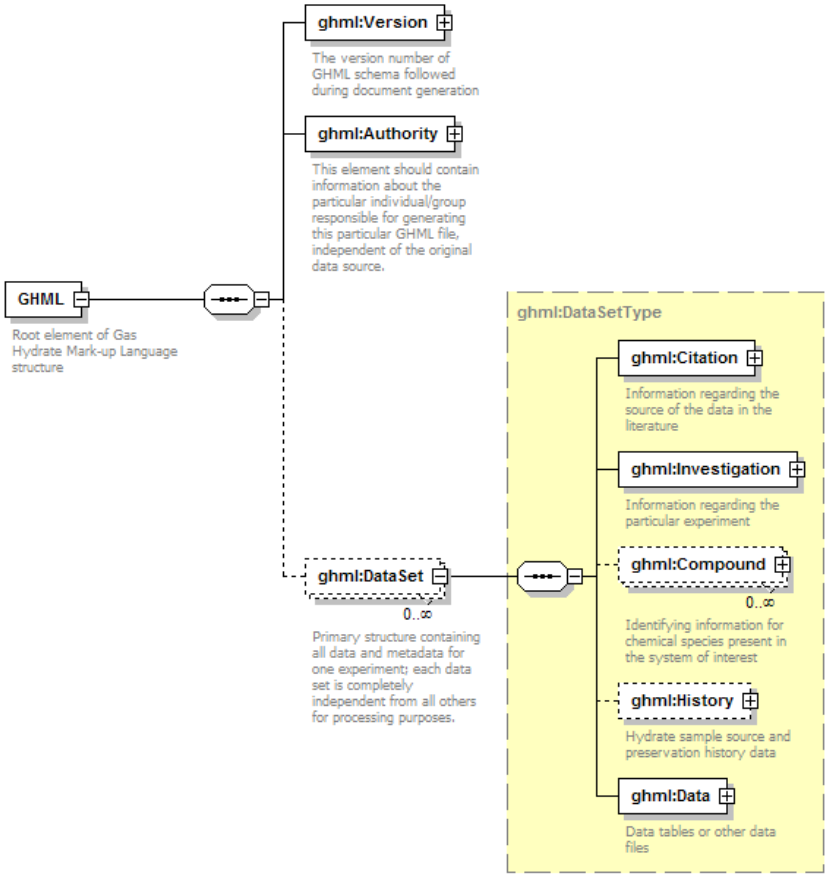

Figure 16. Root element of GHML

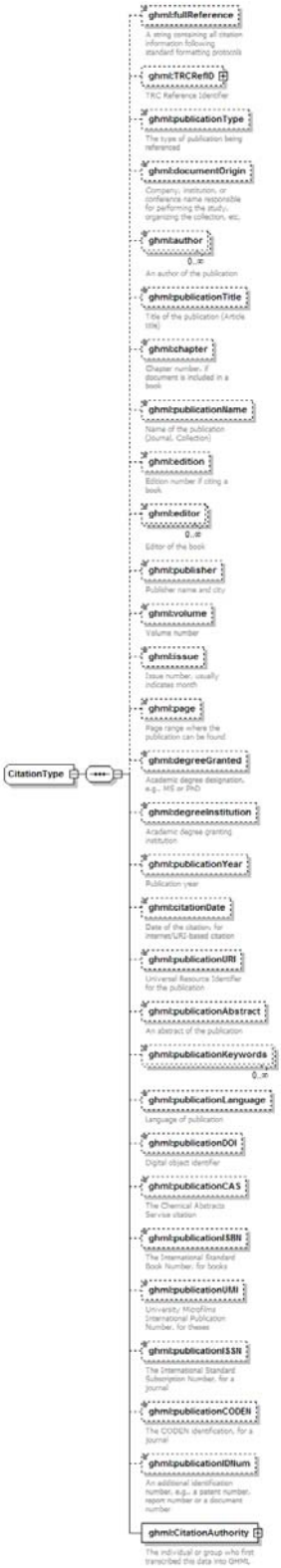

Figure 17. GHML citation element, consistent with ThermoML

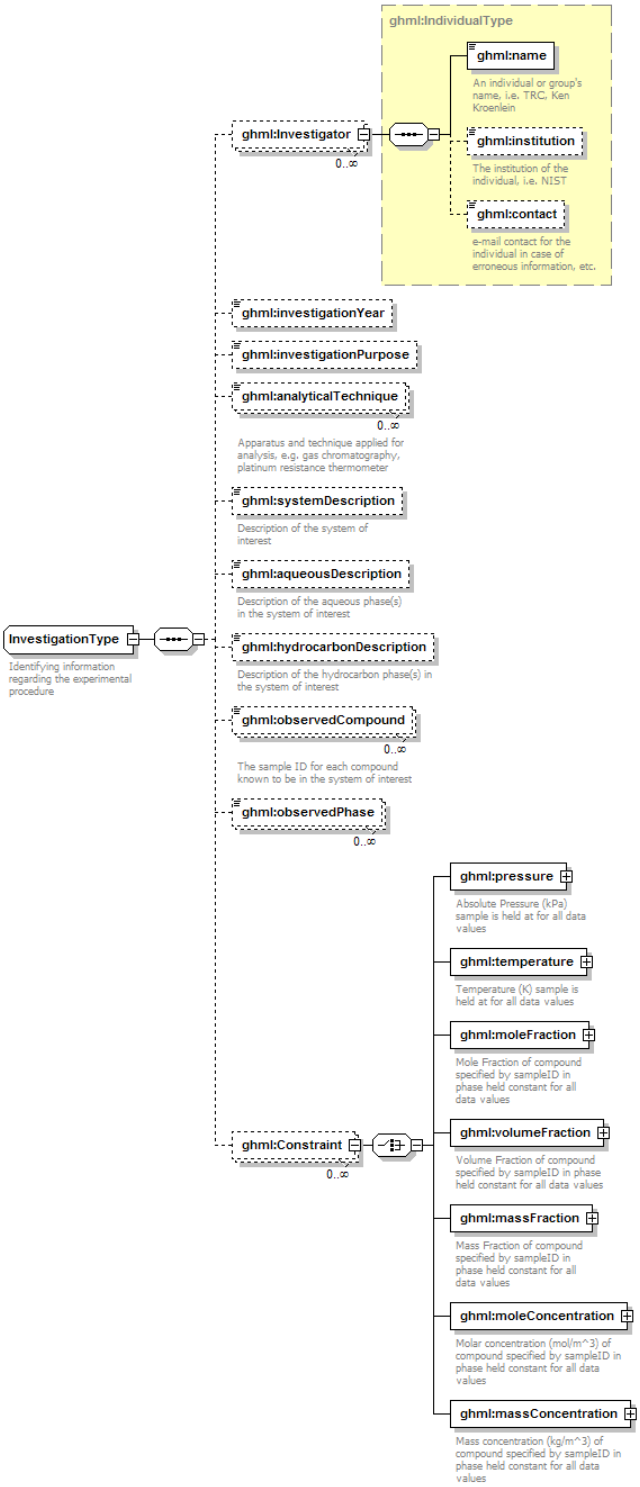

Figure 18. GHML investigation element

Figure 19. GHML compound element

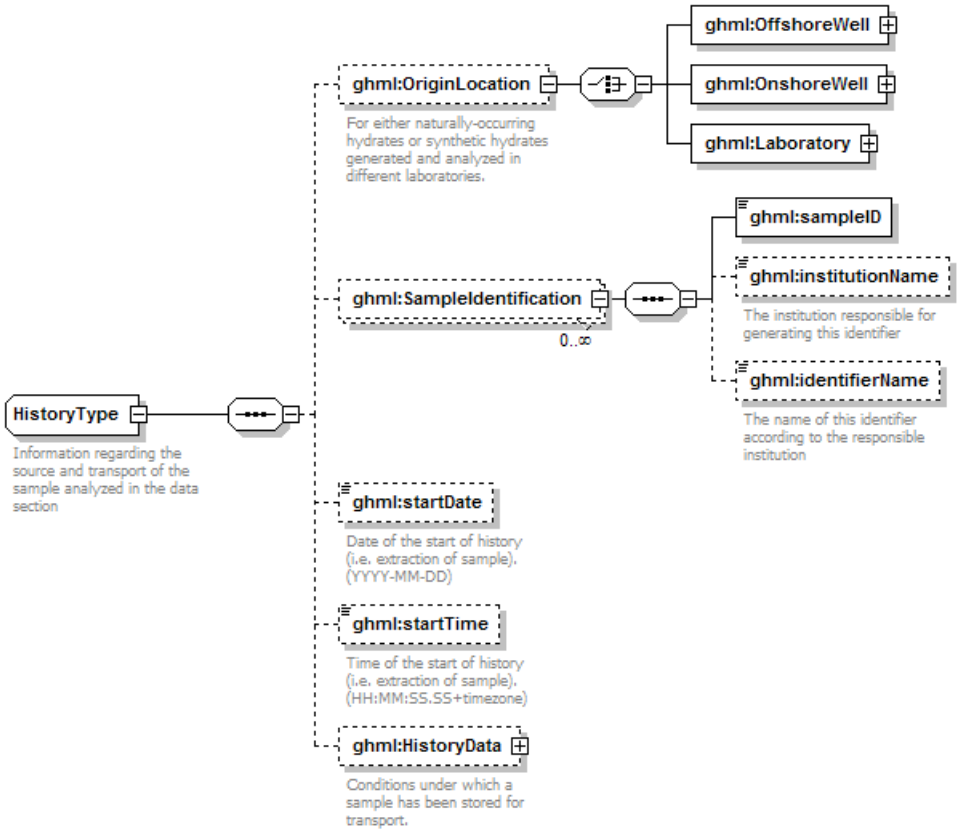

Figure 20. GHML history element

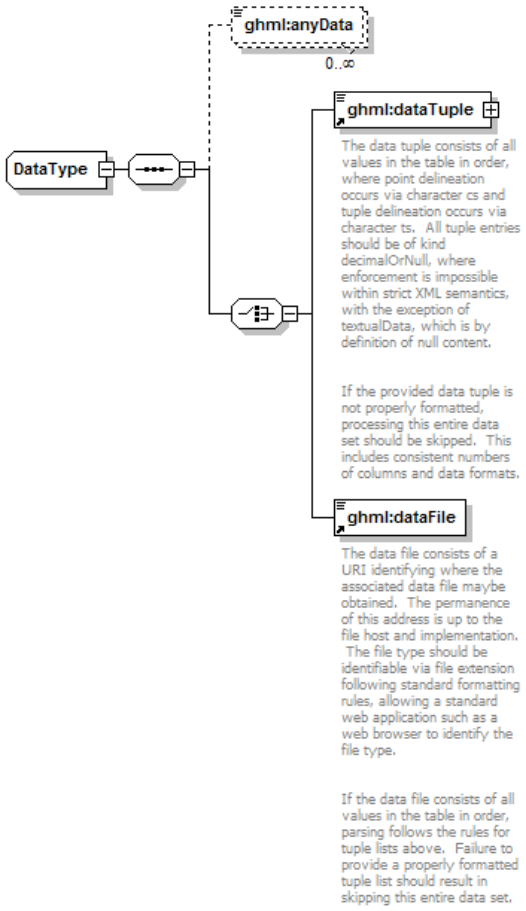

Figure 21. GHML data element

```
<element name="chemicalData">
  <annotation>
    <documentation>Broad Categories describing chemical properties of a sample.  See
  </annotation>
  <complexType>
    <simpleContent>
      <extension base="ghml:chemicalDataType">
        <attribute name="columnNumber" type="integer" use="required"/>
        <attribute name="descriptor" type="string" use="optional"/>
        <attribute name="sampleID" type="integer" use="optional"/>
        <attribute name="phaseID" type="ghml:phaseEnumeration" use="required"/>
      </extension>
    </simpleContent>
  </complexType>
</element>
<simpleType name="chemicalDataType">
  <restriction base="string">
    <enumeration value="Mole Fraction"/>
    <enumeration value="Mass Fraction"/>
    <enumeration value="Volume Fraction"/>
    <enumeration value="Mass Concentration (kg/m^3)"/>
    <enumeration value="Molar Concentration (mol/m^3)"/>
    <enumeration value="Salinity (ppt)"/>
    <enumeration value="Salinity Class"/>
  </restriction>
</simpleType>
```

Figure 22. Exemplar data category from GHML, specifically the ChemicalData subtype

REFERENCES

[1] Sloan ED. *Introductory overview: Hydrate knowledge development*. American Mineralogist 2004;89:1155-1161.

[2] *Methane Hydrate Research and Development Act of 2000* 30 United States Code 1902 note; Public Law 106-193, 2000.

[3] Milkov AV. *Global estimates of hydrate-bound gas in marine sediments: how much is really out there?* Earth-Science Reviews 2004;66(3-4):183-197.

[4] Klauda JB and Sandler SI. *Global Distribution of Methane Hydrate in Ocean Sediment*. Energy & Fuels 2005; 19:459-470.

[5] Radler M. *World crude and natural gas reserves rebound in 2000*. Oil & Gas Journal 2000;98(51):121-123.

[6] Reagan MT and Moridis GJ. *Oceanic gas hydrate instability and dissociation under climate change scenarios.* Geophysical Research Letters 2007;34:L22709.

[7] Kvenvolden KA. *Potential Effects of Gas Hydrate on Human Welfare*. Proceedings of the National Academy of Sciences of the United States of America 1999;96(7):3420-3426.

[8] Berner RA. *Examination of Hypotheses for the Permo-Triassic Boundary Extinction by Carbon Cycle Modeling*. Proceedings of the National Academy of Sciences of the United States of America 2002;99(7):4172-4177.

[9] Padden M, Weissert H and de Rafelis M. *Evidence for Late Jurassic release of methane from gas hydrate*. Geology 2001;29(3):223-226.

[10] Dickens GR. *Carbon addition and removal during the Late Palaeocene Thermal Maximum: basic theory with a preliminary treatment of the isotope record at ODP Site 1051, Blake Nose*. Geological Society, London, Special Publications 2001;183:293-305.

[11] *TRC Group – NIST, Physical and Chemical Properties Division.* 2007. [online]. [Accessed 30th September 2008]. http://www.trc.nist.gov/.

[12] Frenkel M, Dong Q, Wilhoit RC and Hall KR. *TRC SOURCE Database: A Unique Tool for Automatic Production of Data Compilations*. International Journal of Thermophysics 2001;22(1):215-226.

[13] Yan X, Dong Q, Frenkel M and Hall KR. *Window-Based Applications of TRC Databases: Structure and Internet Distribution*. International Journal of Thermophysics 2001;22(1):227-241.

[14] Frenkel M, Chirico RD, Diky V, Yan X, Dong Q and Muzny C. *ThermoData Engine (TDE): Software Implementation of the Dynamic Data Evaluation Concept*. Journal of Chemical Information and Modeling 2005;45(4):816-838.

[15] *ThermoData Engine.* 2008. [online]. [Accessed 30th September 2008]. http://www.trc.nist.gov/tde.html.

[16] Wilhoit RC and Marsh KN. *Future Directions for Data Compilations*. International Journal of Thermophysics 1999;20(1);247-255.

[17] Diky VV, Chirico RD, Wilhoit RC, Dong Q and Frenkel M. *Windows-Based Guided Data Capture Software for Mass-Scale Thermophysical and Thermochemical Property Data Collection*. Journal of Chemical Information and Computer Science 2003;43:15-24.

[18] *Guided Data Capture.* 2007. [online]. [Accessed 30th September 2008]. http://www.trc.nist.gov/GDC.html.

[19] Frenkel M, Chirico RD, Diky VV, Dong Q, Marsh KN, Dymond JH, Wakeham WA, Stein SE, Königsberger E and Goodwin ARH. *XML-Based IUPAC Standard for Experimental, Predicted, and Critically Evaluated Thermodynamic Property Data Storage and Capture (ThermoML)*. Pure and Applied Chemistry 2006;78(3);541–612.

[20] International Union of Pure and Applied Chemistry. 2006. *International Union of Pure and Applied Chemistry* [online]. [Accessed 30th September 2008]. http://www.iupac.org/projects/2002/2002-055-3-024.html.

[21] *ThermoML*. [online]. [Accessed 30th September 2008].

http://www.trc.nist.gov/ThermoML.html.

[22] Sloan D, Kuznetsov F, Lal K, Loewner R, Makogon Y, Moridis G, Ripmeester J, Royer J, Smith T, Tohidi B, Uchida T, Wang J, Wang W and Xiao Y. *A Hydrate Database: Vital to the Technical Community*. Data Science Journal 2007;6(Gas Hydrate Issue):GH1-GH5.

[23] Löwner R, Cherkashov G, Pecher I and Makogon YF. *Field Data and the Gas Hydrate Markup Language*. Data Science Journal 2007;6(Gas Hydrate Issue):GH6-GH17.

[24] Smith T, Ripmeester J, Sloan D and Uchida T. *Gas Hydrate Markup Language: Laboratory Data*. Data Science Journal 2007;6(Gas Hydrate Issue):GH18-GH24.

[25] Wang W, Moridis G, Wang R, Xiao Y and Li J. *Modeling Hydrates and the Gas Hydrate Markup Language*. Data Science Journal 2007;6(Gas Hydrate Issue):GH25-GH36.

[26] International Union of Pure and Applied Chemistry. 2007. *International Union of Pure and Applied Chemistry*. [online]. [Accessed 30th September]. http://old.iupac.org/inchi/.

[27] Dong Q, Yan X, Wilhoit RC, Hong X, Chirico RD, Diky VV and Frenkel MJ. *Data Quality Assurance for Thermophysical Property Databases - Applications to the TRC SOURCE Data System*. Journal of Chemical Information and Computer Science 2002;42(3):473-480.

[28] Hall SR, Allen FH and Brown ID. *The Crystallographic Information File (CIF) - A New Standard Archive File for Crystallography*. Acta Crystallographica Section A 1991;47(6):655-685.

[29] Kroenlein K, Löwner R, Wang W, Diky V, Smith T, Muzny CD, Chirico RD, Kazakov A, Sloan ED and Frenkel M. *Standardization and Software Infrastructure fir Gas Hydrate Data Communications*. Proceedings of the 6th International Conference on Gas Hydrates (IGH 2008), Vancouver, British Columbia, CANADA, July 6-10, 2008.

LIST OF ACRONYMS AND ABBREVIATIONS

ASCII	American Standard Code for Information Interchange, an American National Standards Institute and de facto international standard for digital file character mapping
CIF	Crystallographic Information File, an IUCr standard
CODATA	Committee on Data for Science and Technology, International Council for Science
DQA	Data quality assurance
GDC	Guided data capture, a process for accurate collection of data and associated metadata from literature sources
GHML	Gas Hydrate Markup Language
IUPAC	International Union of Pure and Applied Chemistry
IUCr	International Union of Crystallography, International Council for Science
NIST	National Institute of Standards and Technology, an Agency of the United States Department of Commerce
SOURCE	NIST SOURCE Data Archival System
TDE	ThermoData Engine, NIST Standard Reference Database 103
ThermoML	An XML-based approach for storage and exchange of experimental and critically evaluated thermophysical and thermochemical property data and an IUPAC standard
TRC	Thermodynamics Research Center, Physical and Chemical Properties Division (838) at NIST
XML	Extensible Markup Language
XSD	XML Schema Definition, a file used to specify XML file structure