Gas Hydrate Research Database and Web Dissemination Channel (Year 1)

Final Technical Report
October 1, 2006 to September 30, 2007

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

October 2007

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 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 gas hydrate markup language (GHML) developed to standardize data communications between “data producers” and “data users”, and 3) a relational data storage facility to accommodate all types of numerical and metadata within the scope of the project. All project benchmarks for this initial year of the project were met. A functional version of GDC with appropriate modifications for the capture of gas hydrates was developed. The existing CODATA GHML was modified to be compatible with the IUPAC standard for thermodynamic data communications, ThermoML. The Gas Hydrate Database has been designed to serve as a primary data storage facility for thermophysical and structural property data for gas hydrates.

TABLE OF CONTENTS

ABSTRACT 2
TABLE OF CONTENTS 3
EXECUTIVE SUMMARY 4
Introduction 6
Development of Project-Specific Support Resources 7
Guided Data Capture Software 8
Database Architecture 11
Gas Hydrate Markup Language 13
Conclusions 14
REFERENCES 36
LIST OF ACRONYMS AND ABBREVIATIONS 37

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 done by the Thermodynamic 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 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 gas hydrate markup language (GHML) developed to standardize data communications between "data producers" and "data users", and 3) a relational data storage facility to accommodate all types of numerical and metadata within the scope of the project. In the first year of this three-year effort, all project benchmarks were met. A functional version of GDC with appropriate modifications for the capture of gas hydrates was developed. The existing CODATA GHML was modified to be compatible with the IUPAC standard for thermodynamic data communications, ThermoML. The Gas Hydrate Database has been designed to serve as a primary data storage facility for thermophysical and structural property data for gas hydrates.

As listed in the milestones for Phase I of the project, it was necessary to establish a project Advisory Committee and to select initial sources of data for capture and storage. Accordingly, a project Advisory Committee consisting of 5 internationally recognized experts in the fields of gas hydrates research and chemical informatics was established in October 2006.

Establishing a literature archive at NIST was given an enormous head start through the generous contribution of the Hydrate Literature Database electronic documents (maintained at the Center for Hydrate Research, Colorado School of Mines, in Golden, Colorado, USA). This near-comprehensive archive of documents related to gas hydrates includes approximately 3700 documents in electronic (pdf) format together with full citation information in a structured database. Relevant articles are being migrated from the Hydrate Literature Database to the NIST SOURCE Data Archival System (SOURCE). SOURCE is an Oracle-based relational database that currently contains 90,000 bibliographic sources and approximately 3 million experimental thermochemical and thermophysical property values for pure, binary, and ternary chemical systems. The hardcopy archive of the gas hydrate literature was established as a subset of the NIST/TRC Data Entry Facility archive on the NIST campus in Boulder.

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 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. In order to capture experimental data sets pertaining to samples of gas hydrates, the existing version of GDC has been significantly modified to accommodate the possible non-stoichiometric composition, complex phase equilibria and crystal structure of gas hydrates. The basic tree structure of GDC data organization is 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. 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. New principal elements of GDC were developed and implemented to collect information related to the crystalline structures of gas hydrates. These elements were modeled in accordance to the Crystallographic Information File (CIF) format used widely in the crystallographic community for communication of experimental results. CIF is a standard of the International Union of Crystallography (IUCr).

The GHML language was developed previously under the auspices of CODATA as a series of parallel (*i.e.*, non-intersecting) sections by near-independent research groups. The sections described various types of property data ("laboratory", "field", and "modeling"). As a result of efforts within the current project, a new version of GHML was developed enforcing consistencies between related data blocks, as well as establishing full compatibility with the IUPAC standard for thermodynamic data communications, ThermoML. These modifications were considered and approved by the CODATA Hydrate Task Group during its meeting in Vancouver, British Columbia in October 2007.

The Gas Hydrate Database was developed through modifications of the NIST SOURCE Data Archival System to accommodate all meta- and numerical data elements designed within GDC and GHML for gas hydrates. A principal expansion of SOURCE was related to the establishment of the data fields necessary to store information pertaining to crystalline structures of gas hydrates.

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

REPORT DETAILS

Introduction

To facilitate advances in application of technologies pertaining to clathrate hydrates of natural gas, otherwise known as “gas hydrates”, a United States database pertaining to those materials is being developed. This database, whose scope includes thermodynamic and structural data, will provide researchers in the hydrate community the ability to submit new datasets and the ability to retrieve high-quality, critically-evaluated data sets. By establishing the hydrate database at the United States National Institute of Standards and Technology (NIST) in Boulder, Colorado, in association with an international effort on the part of the Committee on Data for Science and Technology (CODATA), the viability of this project well into the future will be secured. A critically-evaluated hydrate database is essential for eliminating data redundancies, highlighting key data gaps, and providing an assurance of data quality to aid in pertinent research efforts within the United States.

Estimates of the world’s naturally occurring hydrated methane vary greatly, ranging from 5×10^{11} kg [1] to 4.1×10^{15} kg [2]. The vast supply of methane in hydrate is frequently projected to provide for future energy demands [3]. The amount of organic carbon in hydrates can be conservatively estimated as twice the total of all fossil fuels. The remote locations where hydrate exists has prevented current prospecting, but many nations have invested heavily in hydrate recovery programs, including Japan, Germany, India, Canada, and the United States. In addition naturally occurring hydrate typically exists at the brink of thermodynamic stability, so slight changes in ambient temperature or pressure can result in catastrophic release of methane, a potent greenhouse gas, with implications on global climate change [4] and seafloor slope stability [5]. Massive releases of organic carbon to the atmosphere and mass extinction events during the Permian Triassic [6], Late Jurassic [7], Late Paleocene Thermal Maximum [8], and other eras are often connected to the sudden release of hydrated gas.

An overwhelming amount of hydrate data is being (and has been) generated, so an orderly means of accessing them is required. As only one illustration of this knowledge explosion, consider the number of hydrate refereed publications as an indication of the increase in data generation; in the first decades of the last century (1900-1910) there were only two (2) refereed hydrate publications and any interested worker could access both archival journals in which they were published. However, in the last decade of the last century (1990-2000) there were 3010 refereed hydrate publications, extrapolating to over 8,000 additional hydrate-related publications in this decade (2000-2010).

Two examples are provided for how the database would affect recent US Department of Energy (DOE) and Minerals Management Service (MMS) projects.

1. Currently BP and ConocoPhillips have a DOE field project to produce hydrated gas from a well in Alaska. In preparation for the project, extensive modeling is being done (by labs in government, industry, and academia) to determine the sensitivity of various aspects of the well testing. Assessment of these sensitivity parameters depends on availability of a large dataset, from geoscience (geology, geochemistry, and geophysics), thermo-physical properties, etc. - all of which will be readily available through the proposed database.
2. The MMS has been commissioned to perform a multi-year study on the amount of hydrated gas in the waters and permafrost of the USA. This study involves accessing the phase equilibrium conditions, geothermal gradients, geoscience data (such as in the above example) etc. which will require substantial effort by a number of people and

organizations. The proposed database would enable ready access to the appropriate data in a very efficient manner.

Using the existing national databases from other countries as starting points, the database in-development will capitalize on internet access. Submitted data will be subject to analysis of quality, to be published alongside the previously-available hydrate property data. The hydrate database will be structured so that it

1. is easily comprehended and accessible on the Internet
2. produces fast and efficient results
3. allows for data submission
4. evaluates data integrity
5. contains source information, including full bibliographic information

A hydrate database center will be established at the NIST facility in Boulder, Colorado by the Thermodynamic Research Center (TRC) [14]. The existing database at the core of the previous TRC Group activities is SOURCE [18, 19]: the largest relational archival experimental data system, which currently includes more than 120 properties (including structural information) for pure compounds, mixtures, and chemical reactions with data records numbering in the millions. The database will be a critically evaluated dynamic data set, allowing for continuous updating and reliability analysis. The NIST TRC Group has extensive experience in software development for dynamic critical data evaluation with particular application to thermophysical properties. The ThermoData Engine software [9, 10], developed at TRC, is the first full-scale implementation of the dynamic data evaluation concept [11]. The TRC Group 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 [12, 13] for file generation. This approach assures that submitted data are in an appropriate format [15-17] and include sufficient supporting information regarding methods and materials to allow for accurate reliability estimates; application of this proven and stable model is key to the success of the USA Hydrate Database.

The data-transfer approaches associated with this data capture and storage effort are being coordinated with the International Council for Science's Committee on Data for Science and Technology (CODATA), which has been developing a markup language with the purpose of communicating gas hydrate data throughout the research community called Gas Hydrate Markup Language (GHML) [20-22] and an international hydrate portal for centralized access to a number of database efforts. All database output will be fully consistent with ThermoML and GHML. All data will be collected within the NIST SOURCE data system operated with the Oracle data management system. The new hydrate database will be a relational Oracle-based data system with Web-Oracle infrastructure for internet-based dissemination. The initial stage database will include published structural, thermodynamic, and geoscience hydrate data. The design of the NIST SOURCE data storage facility has been modified to accommodate all hydrate data within the scope of the current proposal. The NIST GDC software architecture has been extended to allow capture of gas-hydrate data, and will include GHML compatibility. Data quality analysis tools in the NIST GDC have been extended to accommodate property data for gas hydrates.

Development of Project-Specific Support Resources

As it is outlined in the list of the milestones for the Phase I of the project, it was necessary to establish project Advisory Committee and to select initial sources of data for capture and storage. Accordingly, the project Advisory Committee was established in October of 2006. The members of the Advisory Committee are:

Dr. Dendy Sloan (Committee Chair)
Center for Hydrate Research
Colorado School of Mines
Golden, CO 80401
esloan@mines.edu (303-273-3723)

Dr. Timothy S. Collett
USGS Denver Federal Center MS 940
Denver, Colorado 80225
tcollett@usgs.gov (303-236-5731)

Dr. George Claypool
Former Chief Scientist / Chevron JIP Project
8910 W. 2nd Avenue
Lakewood, CO 80226
geclaypool@aol.com (303-237-8273)

Tom Smith, Data and Security Architect, Office of Converging Technologies
Queens College, City University of New York
Flushing, NY 11367-1597
Tom.smith@cinaplex.com (718-997-5935)

Dr. Michael Frenkel, NIST, Director / Thermodynamics Research Center
325 Broadway, Mailstop 838.01
Boulder, CO 80305-3328
Frenkel@boulder.nist.gov (303-497-3952)

An initial meeting of the Advisory Committee was held October 12, 2006 at NIST in Boulder. A second meeting was held May 9, 2007 in conjunction with a meeting involving the developers of the beta version of GHML. All members were in attendance at both meetings and each meeting included the staff of TRC. As expert members of the gas-hydrate community in the areas of property research (Dr. Sloan), energy resource characterization (Dr. Collett), and field formation and assessment (Dr. Claypool), the members provided guidance in establishing the scope of the database collection to be housed at NIST.

Establishing a literature archive at NIST was given an enormous head start through the generous contribution to these efforts of the Hydrate Literature Database electronic documents (maintained by Dr. Dendy Sloan at the Center for Hydrate Research, Colorado School of Mines). This near-comprehensive archive of documents related to gas hydrates includes approximately 3700 documents in electronic (pdf) format together with full citation information in an EndNote database. Relevant articles are being migrated from the Hydrate Literature Database to the NIST SOURCE Data Archival System (SOURCE). SOURCE is an Oracle-based relational database that currently contains 90,000 bibliographic sources and nearly 3 million data experimental thermochemical and thermophysical property values for pure, binary, and ternary chemical systems. The hardcopy archive of the gas hydrate literature is established as a subset of the NIST/TRC Data Entry Facility archive on the NIST campus in Boulder.

Guided Data Capture Software

Information from original data sources is not entered directly into 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 software (GDC), 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.

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 [15-17] 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. Aspects of steps 5 and 6 were discussed in an earlier paper [23].

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 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.

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 original software, without support for gas hydrate property capture, is available for free downloading via the World Wide Web [13]. Comprehensive documentation for the software is included. GDC 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 [12].

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 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 [24], which is an International Union of Crystallography (IUCr) standard.

Database Architecture

TRC rejoined the 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 [11] as implemented in TDE [9, 10].

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 [18, 19] 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. The NIST SOURCE Data System resides and is maintained on an enterprise-level Sun Microsystems SunFire 280R 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 data system on the external server running outside of the NIST firewall. Presently, there are several Dell servers running Red Hat Enterprise Linux available for this purpose.

Due to the conflicts similar to those discussed pertaining to GDC development, extensions to SOURCE were required to accommodate thermophysical properties, phase equilibria, and crystal structure data for gas hydrates with the same level of critical review as for the existing system. New tables that have been added can be divided into four groups: gas hydrate composition

description (table GHCOMP), gas hydrate sample description (tables GHSAMPLE and GHSAMPLECOMP), gas hydrate crystal structure data, and gas hydrate complex phase equilibrium data. Detailed descriptions of new data structures are provided below, where it should be understood that any information that is not included in these groups is stored within the previously available SOURCE data structures (inclusive of gas hydrate bulk property data and gas hydrate physical reaction data).

A schematic representation of the TRC SOURCE database with necessary modifications to accommodate the storage of gas hydrate data can be found in Figure 12. For ease of reference, the structure of the database prior to modifications is enclosed by a red dashed line and all newly-implemented, gas-hydrate specific tables have been given the prefix “GH”. In general, structural additions follow the patterns established in development of the GDC software. The composition of a gas hydrate sample, as stored in the new GHCOMP table (Figure 13), contains fields for storing a gas hydrate registry number (field GHRN) which uniquely define each combination of hydrate formers, and fields to store the relevant information regarding which compounds are present in what roles: component registry numbers (field CMPRN), the composition metric and associated numerical amounts (fields CMPTYPE and CONTENT) and whether this compound functions as a host or guest (field FUNCT).

Information regarding a particular sample of gas hydrate, as associated with a particular set of property measurements, is stored in the new tables GHSAMPLE and GHSAMPLECOMP (Figure 14). Linked by the unique identifier GHSAMPLEID, these two tables contain information regarding the state of the specific sample as a whole (PREP_T – preparation temperature, PREP_P – preparation pressure, COMPFLAG – flag indicating if the composition reported before preparation (may not correspond to actual composition)) and references to the individual component samples (field CMPSAMPLEID) and the particulars of their molar distribution (fields CMPTYPE and CONTENT).

Once both the source document (using previously existing SOURCE data structures) and the state and purity of a sample (using previously existing SOURCE data structures in conjunction with the gas-hydrate-specific structures) have been defined appropriately, information regarding physical properties can be stored with confidence in future accountability. Data sets can be divided into three categories: bulk physical properties, crystallographic structure information and complex phase equilibria. Bulk physical properties such as mass specific volume, thermal conductivity, heat capacity at constant pressure per unit mass, or speed of sound, SOURCE has proper previously-existing data structures and thus no further modifications are necessary for these data. There has been no prior attempt to include crystallographic data within SOURCE and so wholly original data tables must be added to accommodate them. These tables, labeled GHSTRUCT, GHSTRUCTRAW and GHSRUCTPROC (Figure 15) generally replicate the entries from the GDC crystal structure form and thus trace their structure back to the CIF file common in crystallographic data distribution [24].

Characterizing the complex phase equilibria requires significant extension of the existing format due to the large increase in potential numbers of combinations of phases and components present in some potentially-interesting gas hydrate data sets. The gas hydrate complex equilibrium data block contains four different tables (Figure 16) to record measurement method (table GHEQDSET), the phases present and compounds present therein (table GHEQPHASES), the types and values of contraints known to exist on the system (table GHEQVARCONSTR) and lastly the particular data points associated with the measurement (table GHEQDATA), all linked by a unique data set ID (field GHDSETID). The GHEQPHASES table contains a set of fields labeled CMPSAMPLEID*n*, where each contains the sample identifiers for the *n*th component

present in the listed phase; presently up to 8 components are allowed. A second unique identified is generated for each phase within the system allowing for appropriate reference to system properties that are specific to a particular phase, such as molar composition. Care is taken to as well store information pertaining to the reported uncertainty of the measurement (field UNCTYPE – type of reported uncertainty (relative or absolute); field UNCERT – value of uncertainty according to UNCTYPE).

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.

An extensive meeting/workshop was held (May 1-4, 2007 at NIST in Boulder, CO) that included all NIST project personnel, the Advisory Committee for this project (May 2nd only), and the complete development team of the beta version of GHML. The development team consisted of Dr. Dendy Sloan and Tom Smith (of the Advisory Committee), Weihua (Willa) Wang (Computer Information Center, Chinese Academy of Sciences, Beijing, China), and Ralf Löwner (IT Project Manager, GeoForschungsZentrum, Potsdam, Germany). The primary focus of the week's meetings was reconciliation of the then-existent GHML schema [20-22] with ThermoML [15-17]. Numerous implementation details were discussed and a roadmap was established for completion of this milestone. Long-term issues related to establishment of a Web portal for data dissemination were discussed also. A time line was set for transfer of the GHML schema to NIST for full ThermoML integration and finalization of the GHML schema as scheduled.

Establishment of the time line for transfer of the GHML schema to NIST for full ThermoML integration and finalization of the GHML schema was a key outcome of the May 2007 meeting/workshop. July 1 was set as a final date for changes to the schema by the non-NIST GHML developers and for transfer of the beta version of GHML to NIST for integration with the IUPAC standard ThermoML markup language. The GHML language had been developed as a series of parallel (i.e., non-intersecting) sections by essentially independent authors that described various types of property data ("laboratory" [21], "field" [20], and "modeling" [22]).

In order to make GHML consistent with the design philosophies which have contributed to the success of the TRC group's previous data collection efforts significant restructuring of the schema was required. 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.

The root element of the new proposed GHML includes moving the citation element from being distributed throughout the schema to a centralized record off of the root element, as well as including version information for validation in case of future revision of the schema (Figure 17). This creates a parallel construct to the root element of ThermoML. The remaining information for characterizing a laboratory sample in ThermoML is then shifted to one branch further into the tree. As the schema section for characterization of data measured under laboratory conditions is intended for both synthetic hydrates and for naturally-occurring hydrate measured *ex situ*, it is essential that there be information storage capability allowing for the preservation history in transporting the sample to the final location for analysis (Figure 18). This marks a significant departure from previous TRC efforts in that those specifically excluded any property which depended on the preparation history of the sample.

The citation element (Figure 19) has been reorganized to provide only those fields relevant for the type of source document being characterized; to maintain one-to-one compatibility with existing ThermoML documents and resources, all fields in the ThermoML formulation have a corresponding element in GHML for any data which could be relevant to a given source document type; the "Unspecified" category demonstrated herein contains all possible fields.

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 must be approved by that body. A meeting of the CODATA Hydrate Database Steering Committee was held on October 27, 2007 to coincide with a meeting of the Organizing Committee for the Sixth International Conference on Gas Hydrates in Vancouver, British Columbia. At this meeting, the proposed modifications for GHML have been presented and accepted.

Conclusions

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

A project Advisory Committee, consisting of 5 internationally recognized experts in the fields of gas hydrates research and chemical informatics, was established in October 2006.

A literature archive for gas hydrate data publications was established at NIST in Boulder. The archive was given an enormous head start through the generous contribution of the Hydrate Literature Database electronic documents compiled at the Center for Hydrate Research, Colorado School of Mines, Golden, Colorado.

Guided data capture (GDC) software was designed and implemented. GDC is a software tool for conversion of data and metadata from literature format into a well-organized electronic format appropriate for electronic analysis and database integration. It is expected that the software will continue to evolve during the electronic migration of the literature archive as novel data sets are encountered.

The gas hydrate markup language (GHML) was modified to ensure consistency with the IUPAC-standard ThermoML. Development of this format will expedite communications between "data producers" and "data users" in the gas-hydrate community worldwide.

A relational data storage facility capable of accommodating all types of numerical and metadata within the scope of the project was developed based upon the previously existing SOURCE data system. Use of an existing, highly successful data structure as the basis for the new gas-hydrate data organization and dissemination project is important in ensuring its ultimate success.

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

INDEX OF GRAPHICAL MATERIALS

Figure 1. Screen capture of tree structure for a gas hydrate sample characterization within GDC... 17
Figure 2. Screen capture of GDC dialog for definition of a gas hydrate system..... 18
Figure 3. Screen capture of GDC dialog for definition of a gas hydrate sample 19
Figure 4. Screen capture of GDC dialog for defining phase equilibrium constraints and variables on a given set of phase equilibrium data 20
Figure 5. Screen capture of GDC dialog for entering tabulated data associated with a given set of phase equilibrium data 21
Figure 6. Screen capture of natively-generated graph of data entered into GDC tabulated data dialog 22
Figure 7. Screen capture of GDC dialog for defining the physical reaction associated with gas hydrate decomposition..... 23
Figure 8. Screen capture of GDC dialog for defining a type of bulk measurement and the associated measurement methodology..... 24
Figure 9. Screen capture of GDC dialog for defining the thermodynamic conditions under which a bulk measurement was performed 25
Figure 10. Screen capture of GDC dialog for entering tabulated data associated with a given set of bulk property data with automated reliability estimate 26
Figure 11. Screen capture of GDC dialog for storing crystallographic data, including space group, unit cell parameters and atom distribution..... 27
Figure 12. Schematic representation of TRC SOURCE database (inside red dashed line) with modifications required to accommodate storage of gas hydrate (GH) data..... 28
Figure 13. Structure of new GHCOMP table describing gas hydrate composition..... 29
Figure 14. Structure of new GHSAMPLE and GHSAMPLECOMP tables describing specific gas hydrate samples..... 30
Figure 15. Structure of new GHSTRUCT, GHSTRUCTRAW and GHSTRUCTPROC tables for storage of crystal structure data 31
Figure 16. Structure of new GHEQDSET, GHEQVARCONSTR, GHEQDATA and GHEQPHASES tables for storage of phase equilibrium data..... 32
Figure 17. Root element of proposed modified GHML for consistency with ThermoML 33
Figure 18. LabData element of proposed modified GHML for consistency with ThermoML 34
Figure 19. Citation element of proposed modified GHML for consistency with ThermoML 35

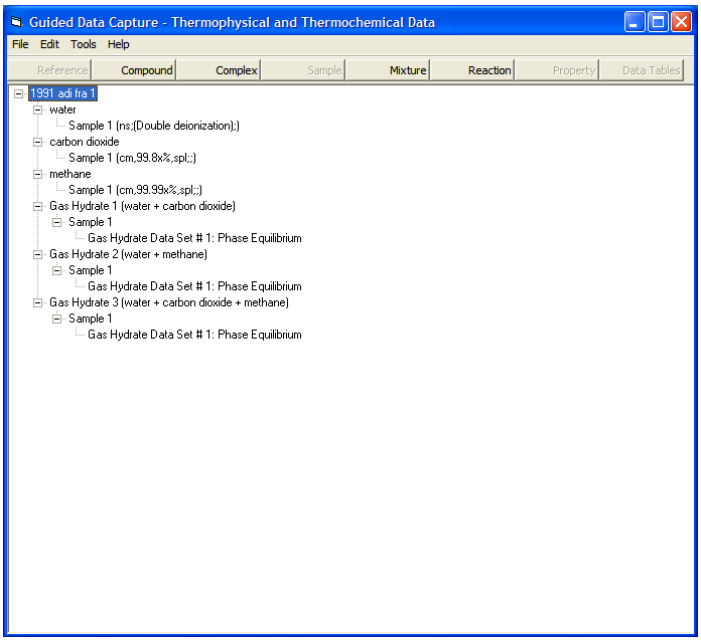

Figure 1. Screen capture of tree structure for a gas hydrate sample characterization within GDC

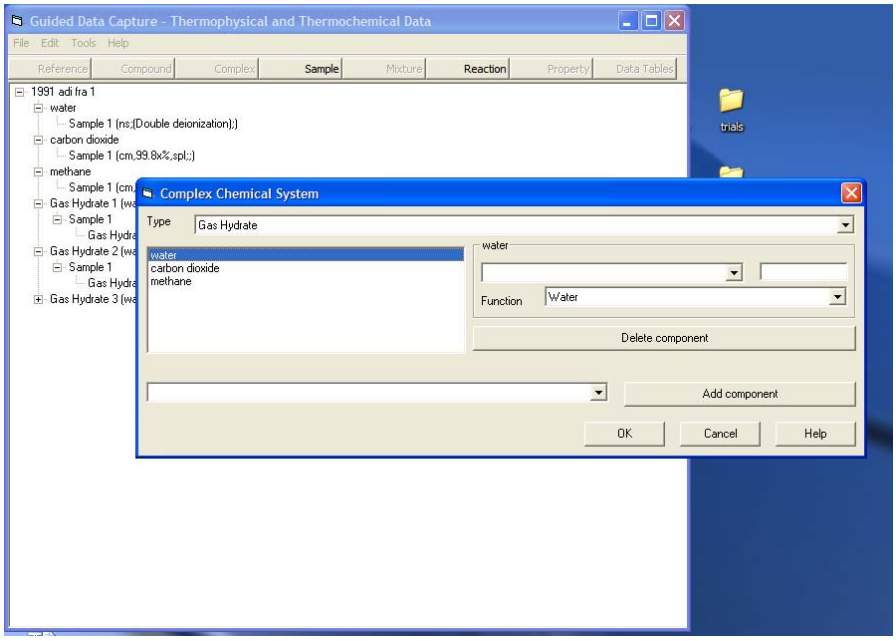

Figure 2. Screen capture of GDC dialog for definition of a gas hydrate system

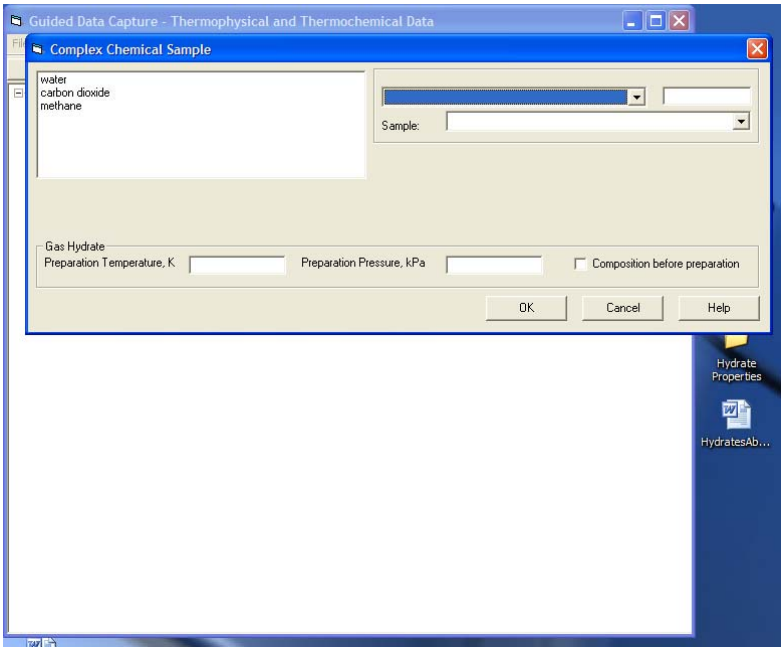

Figure 3. Screen capture of GDC dialog for definition of a gas hydrate sample

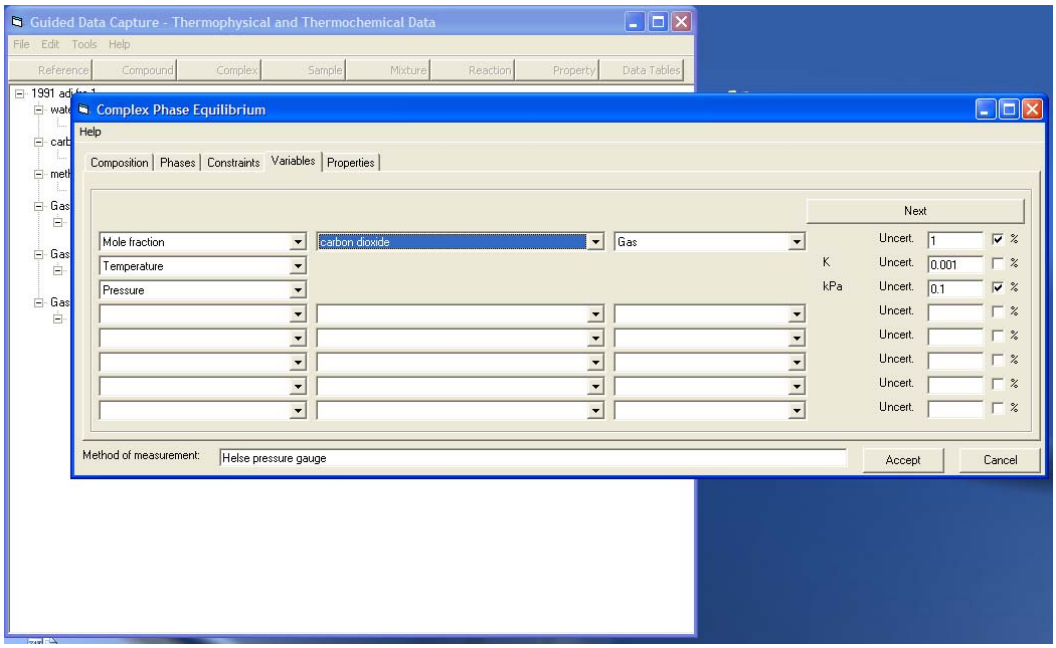

Figure 4. Screen capture of GDC dialog for defining phase equilibrium constraints and variables on a given set of phase equilibrium data

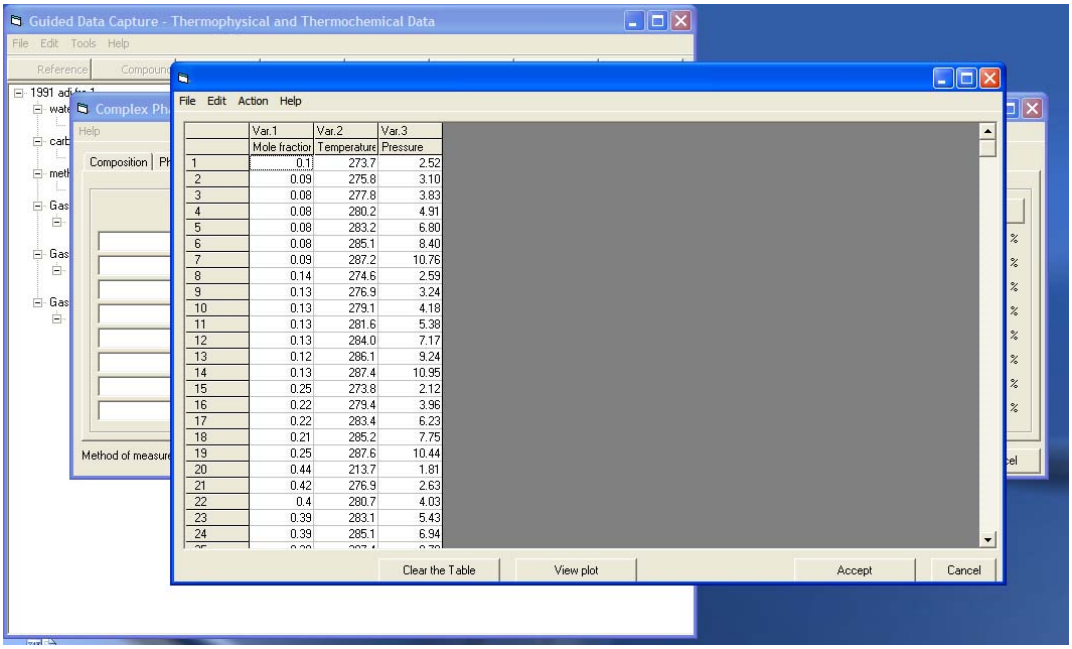

Figure 5. Screen capture of GDC dialog for entering tabulated data associated with a given set of phase equilibrium data

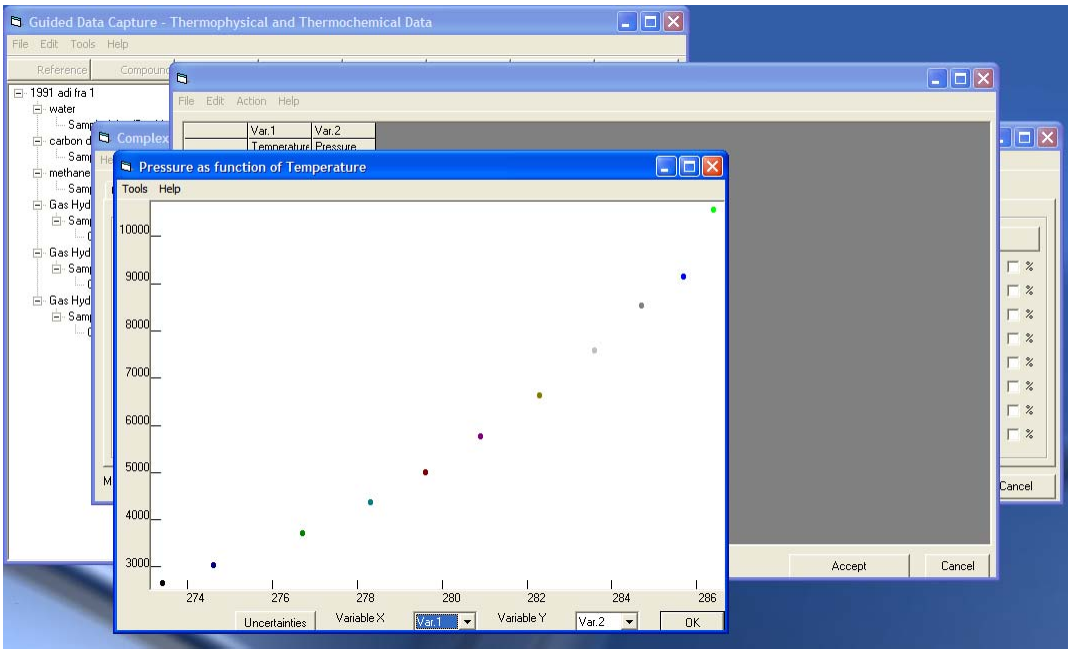

Figure 6. Screen capture of natively-generated graph of data entered into GDC tabulated data dialog

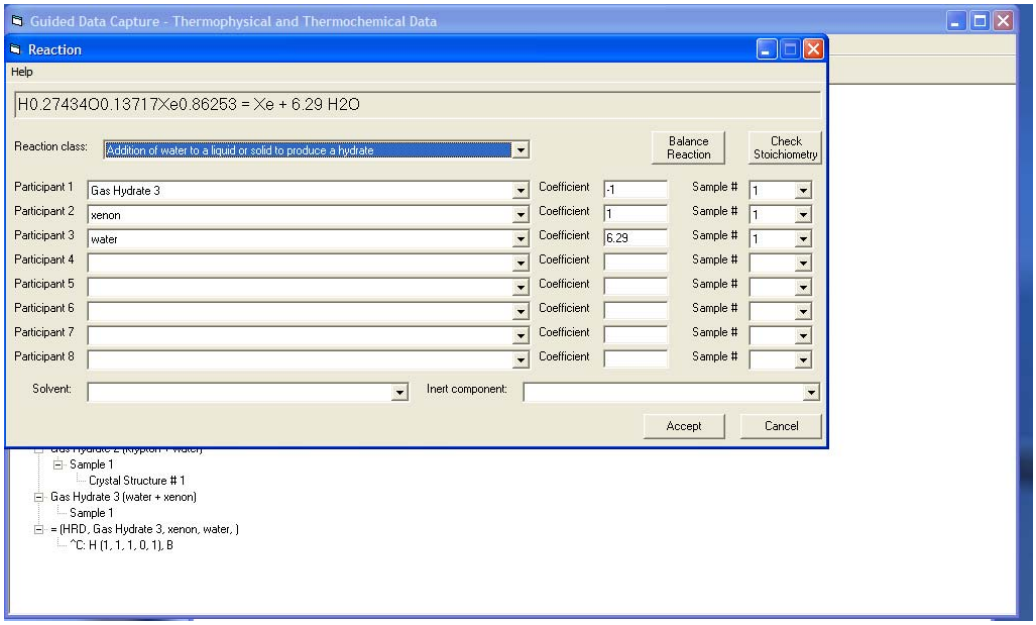

Figure 7. Screen capture of GDC dialog for defining the physical reaction associated with gas hydrate decomposition

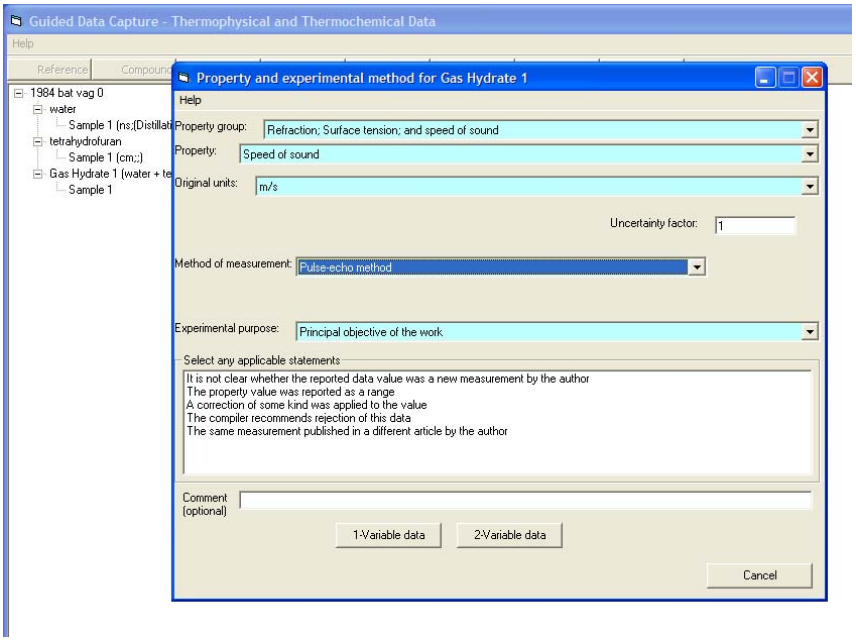

Figure 8. Screen capture of GDC dialog for defining a type of bulk measurement and the associated measurement methodology

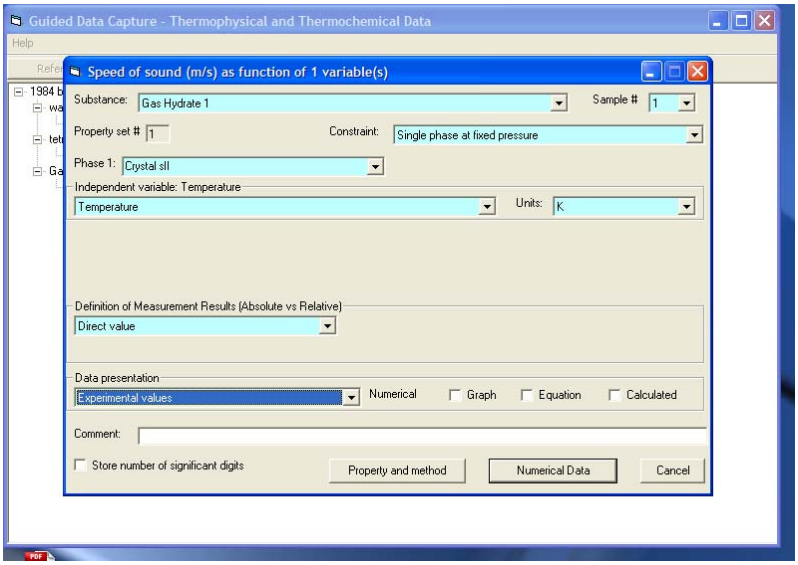

Figure 9. Screen capture of GDC dialog for defining the thermodynamic conditions under which a bulk measurement was performed

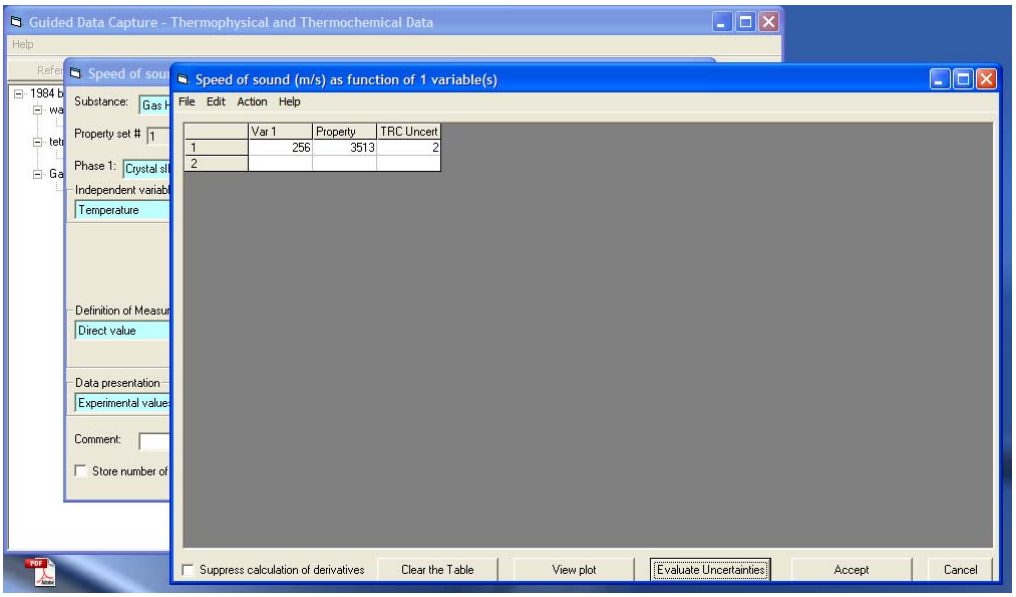

Figure 10. Screen capture of GDC dialog for entering tabulated data associated with a given set of bulk property data with automated reliability estimate

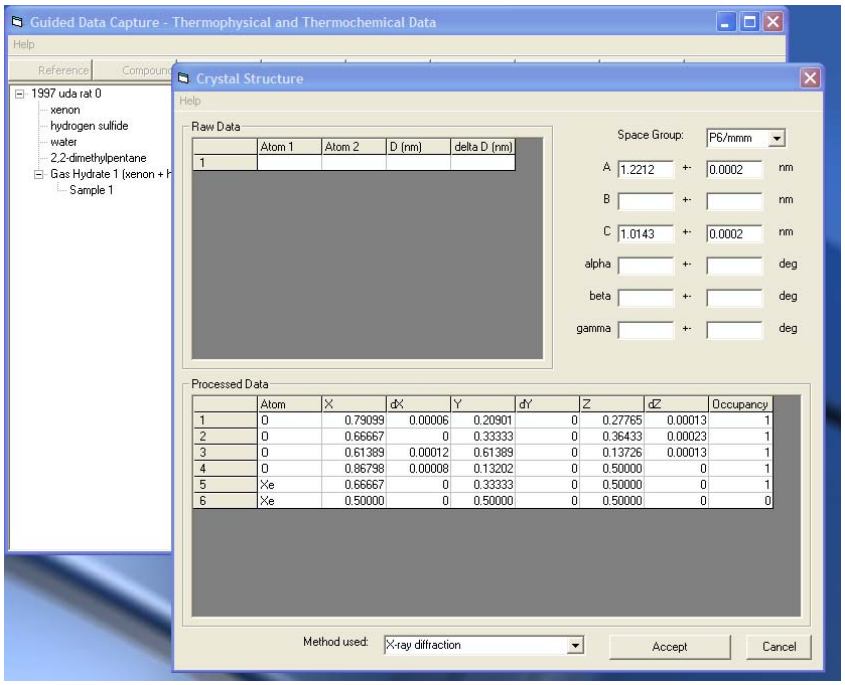

Figure 11. Screen capture of GDC dialog for storing crystallographic data, including space group, unit cell parameters and atom distribution

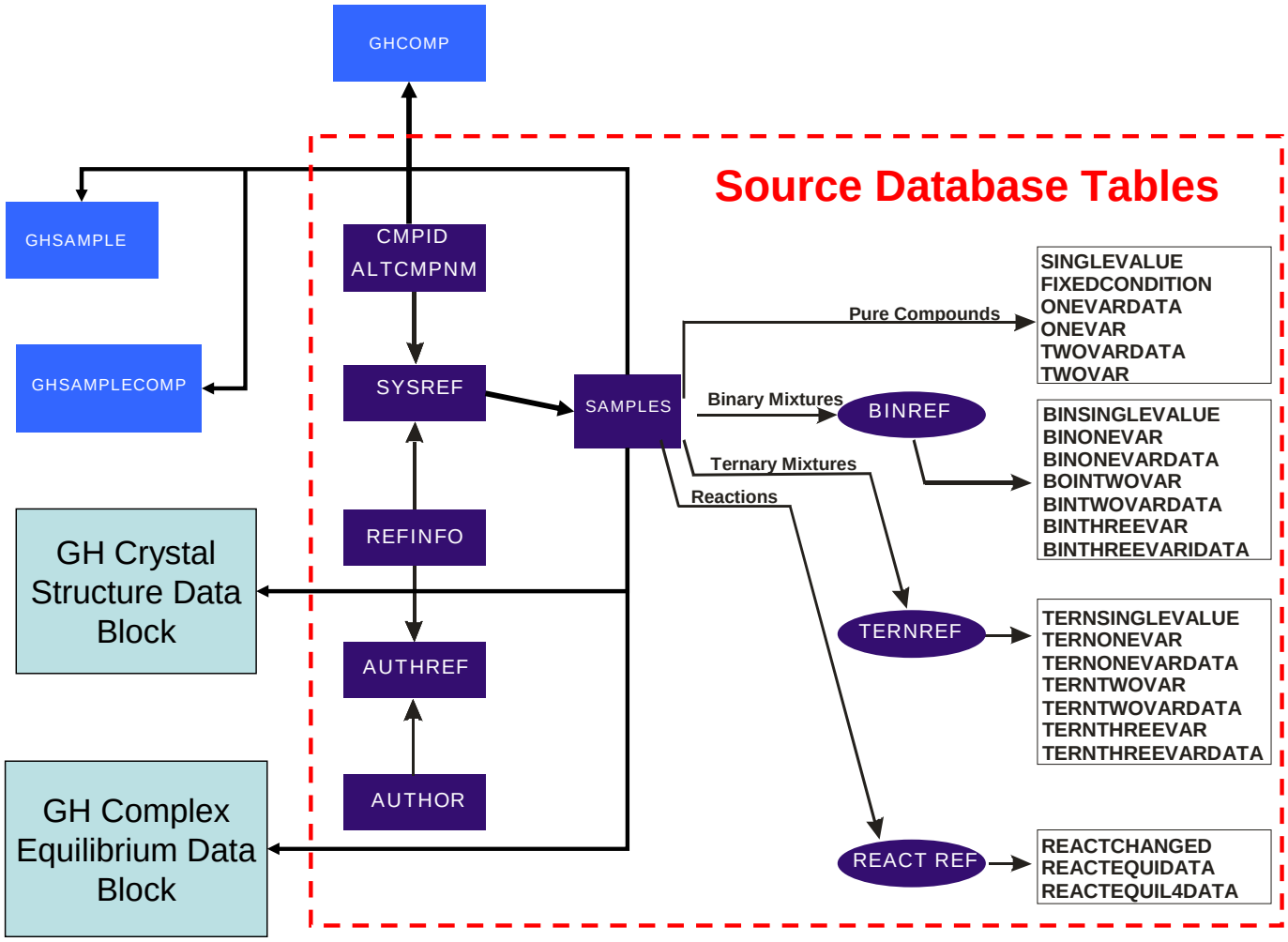

Figure 12. Schematic representation of TRC SOURCE database (inside red dashed line) with modifications required to accommodate storage of gas hydrate (GH) data

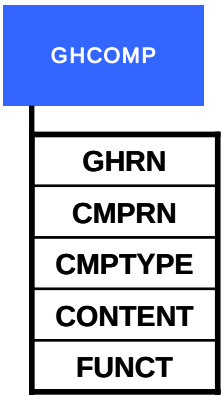

Figure 13. Structure of new GHCOMP table describing gas hydrate composition

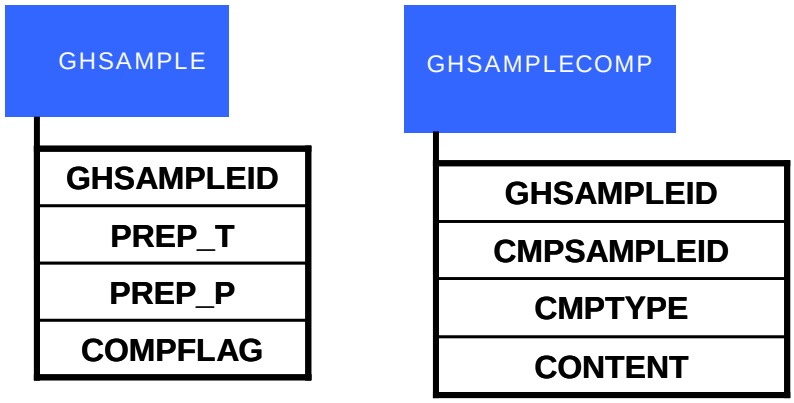

Figure 14. Structure of new GHSAMPLE and GHSAMPLECOMP tables describing specific gas hydrate samples

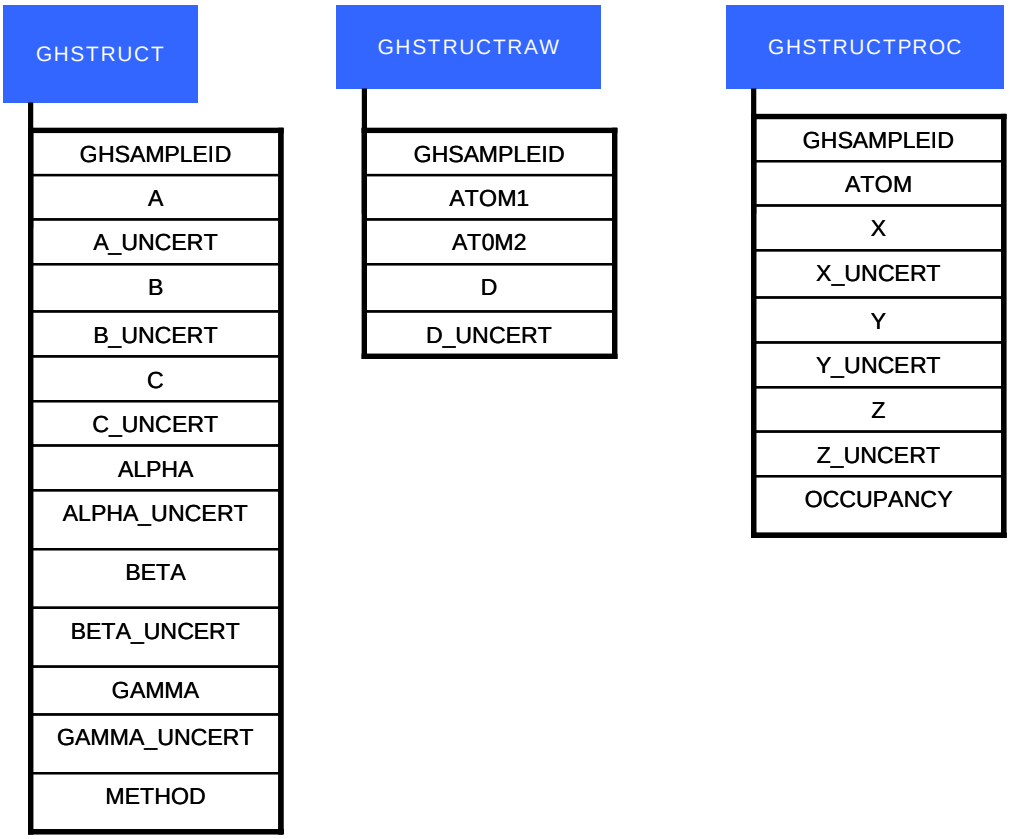

Figure 15. Structure of new GHSTRUCT, GHSTRUCTRAW and GHSTRUCTPROC tables for storage of crystal structure data

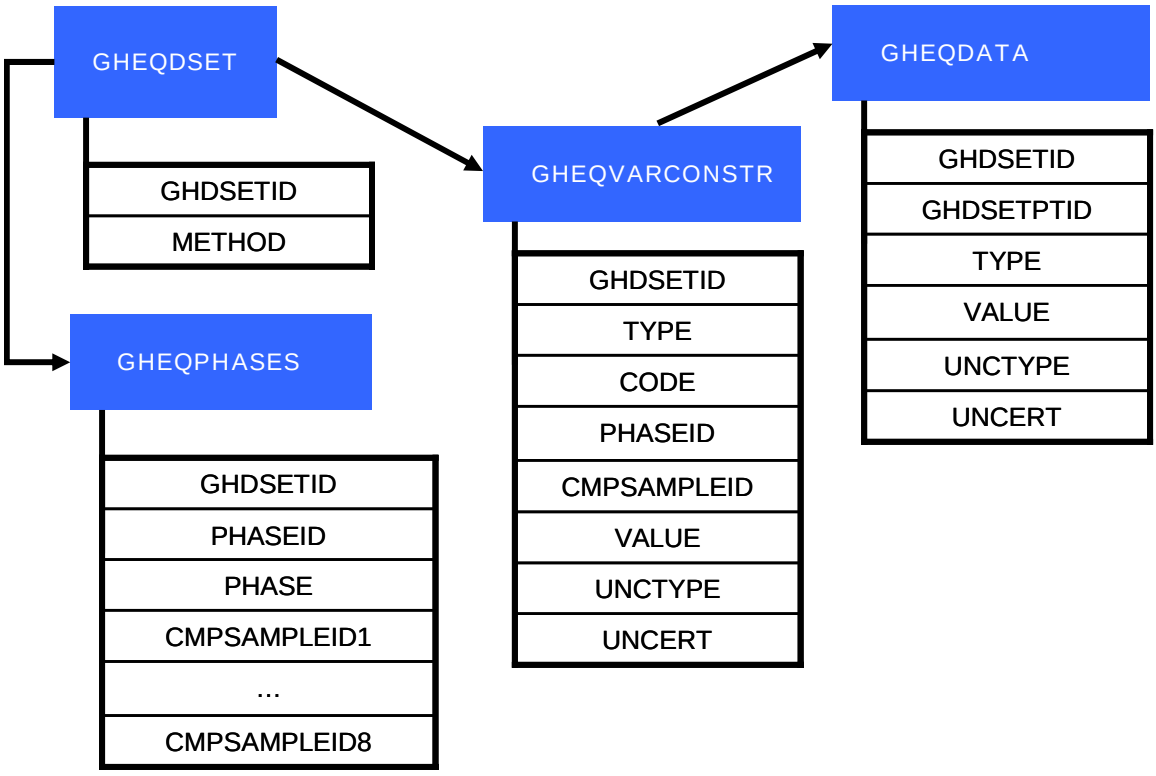

Figure 16. Structure of new GHEQDSET, GHEQVARCONSTR, GHEQDATA and GHEQPHASES tables for storage of phase equilibrium data

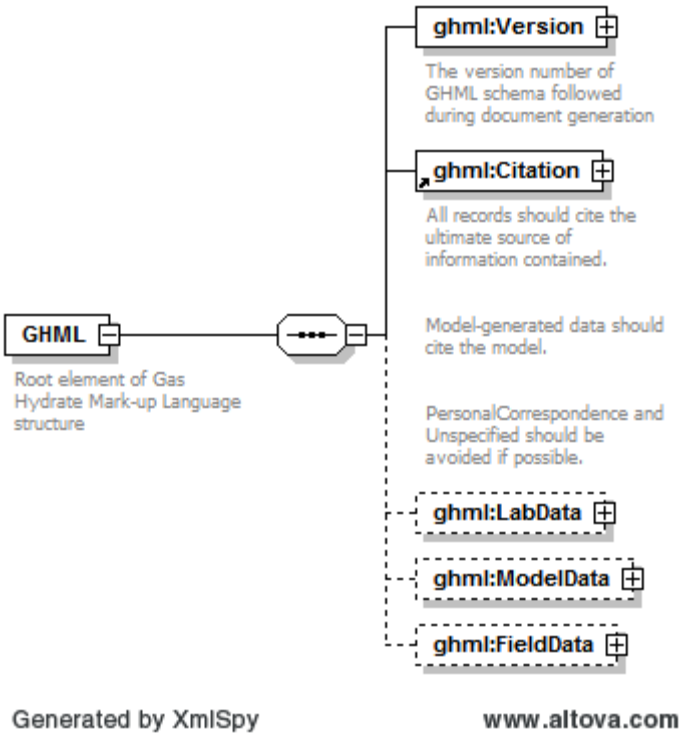

Figure 17. Root element of proposed modified GHML for consistency with ThermoML

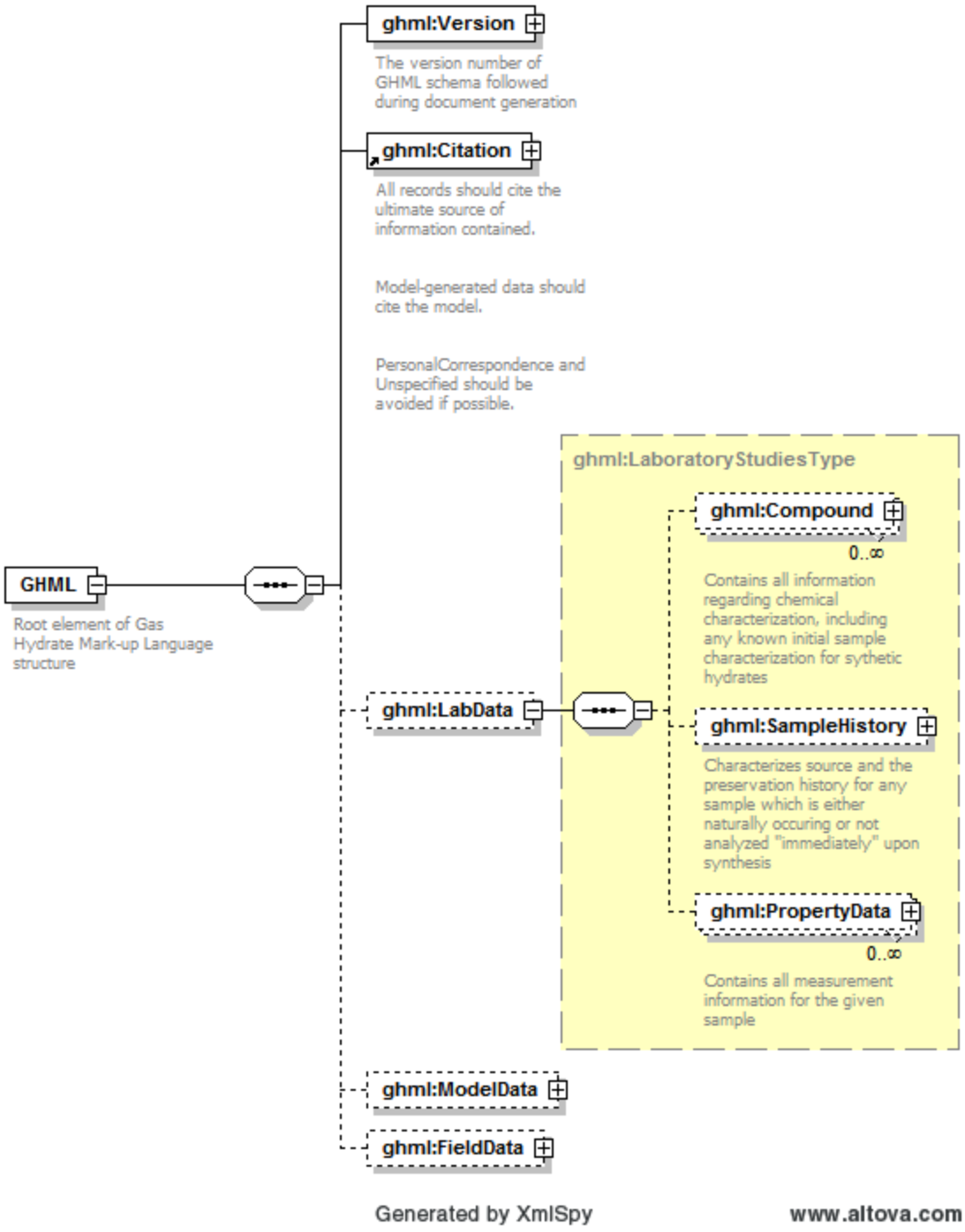

Figure 18. LabData element of proposed modified GHML for consistency with ThermoML

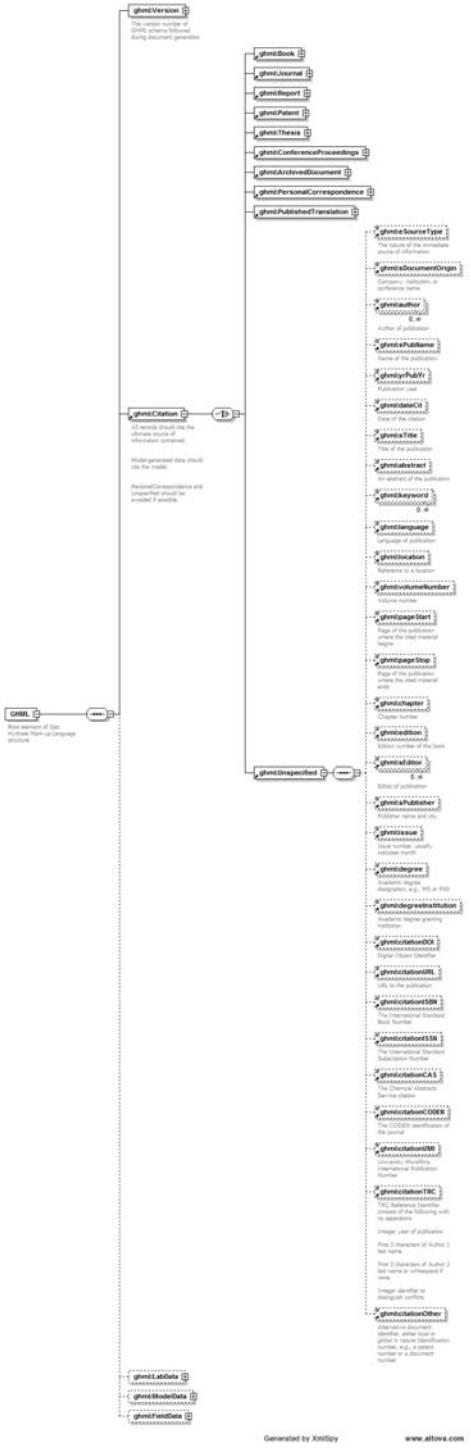

Figure 19. Citation element of proposed modified GHML for consistency with ThermoML

REFERENCES

[1] Ginsburg, G. D. and Soloviev, V. A., (1995) *Offshore Technology Conference* 7693, **1**, 513-518.

[2] Dobrynin, V. M., Korotajev, Y. P. and Plyuschev, D. V., (1981) in Long-term Energy Resources, Pitman, Boston, 727-729.

[3] Dobrynin, V. M., Korotajev, Y. P. and Plyuschev, D. V., (1979) *UNITAR Conf. on Long Term Energy Resources*, 4.

[4] Dickens, G. R., (1999) *Chem. Geol.* **157**, 335-336.

[5] Hovland, M. and Gudmestad, O. T., (2001) in Natural Gas Hydrates: Occurrence, Distribution, and Detection, AGU Monograph 124, 307-315.

[6] Berner, R. A., (2002) *Proc. Nat. Acad. Sciences U. S. of America*, **99** (7), 4172-4177.

[7] Padden, M., Weissert, H. and de Rafelis, M., (2001) *Geol.*, **29** (3), 223-226.

[8] Dickens, G. R., (2001) *Geological Society Special Publication*, **183**, 293-305.

[9] Frenkel, M., Chirico, R. D., Diky, V., Yan, X., Dong, Q., Muzny, C., (2005) *J. Chem. Inf. Model.*, **45**(4), 816-838.

[10] *http://www.trc.nist.gov/tde.html*

[11] Marsh, K. N., Wilhoit, R. C., (1999) *Int. J. Thermophys.* **20**(1), 247-255.

[12] Diky, V. V., Chirico, R. D., Wilhoit, R. C., Dong, Q., and Frenkel, M., (2003) *J. Chem. Inf. Comput. Sci.*, **43**, 15-24.

[13] *http://www.trc.nist.gov/GDC.html*

[14] *http://www.trc.nist.gov/*

[15] Frenkel, M., Chirico, R. D., Diky, V. V., Dong, Q., Marsh, K. N., Dymond, J. H., Wakeham, W. A., Stein, S. E., Königsberger, E., Goodwin, A. R. H., (2006) *Pure Appl. Chem.*, **78**(3), 541–612.

[16] *http://www.iupac.org/projects/2002/2002-055-3-024.html* (Web link to the IUPAC project)

[17] *http://www.trc.nist.gov/ThermoML.html* (ThermoML)

[18] Frenkel, M., Dong, Q., Wilhoit, R. C., Hall, K. R., (2001) *Int. J. Thermophys.*, **22**, 215-226.

[19] Yan, X., Dong, Q., Frenkel, M., Hall, K. R., (2001) *Int. J. Thermophys.* **22**(1), 227-241.

[20] Löwner, R., Cherkashov, G., Pecher, I., and Makogon, Y. F., (2007) *Data Science Journal*, **6**, GH6-GH17.

[21] Smith, T., Ripmeester, J., Sloan, D., and Uchida, T., (2007) *Data Science Journal*, **6**, GH18-GH24.

[22] Wang, W., Moridis, G., Wang, R., Xiao, Y., and Li, J., (2007) *Data Science Journal*, **6**, GH25-GH36.

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

[24] Hall S. R., Allen F. H., Brown I. D., (1991) *Acta Cryst* A47: 655-685.

LIST OF ACRONYMS AND ABBREVIATIONS

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
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, NISI 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	Thermodynamic Research Center, Physical and Chemical Properties Division (838) at NIST
XML	Extensible Markup Language