Proceedings of the 6th International Conference on Gas Hydrates (ICGH 2008),
Vancouver, British Columbia, CANADA, July 6-10, 2008.

STANDARDIZATION AND SOFTWARE INFRASTRUCTURE FOR GAS HYDRATE DATA COMMUNICATIONS[*]

K. Kroenlein[a†], R. Löwner[b], W. Wang[c], V. Diky[a], T. Smith[d], C. D. Muzny[a], R. D. Chirico[a], A. Kazakov[a], E. D. Sloan[e], and M. Frenkel[a]

[a]Physical and Chemical Properties Division, National Institute of Standards and Technology, 325 Broadway, Boulder, Colorado 80305-3328, USA
[b]GeoForschungsZentrum Potsdam, Telegrafenberg, Potsdam, D-14473, Germany
[c]Computer Network Information Center, Chinese Academy of Science, 4, 4th South Street Zhong guan cun, P. O. Box, 349, Beijing 10080 China
[d]MIT Systems, 565 Plandome Road #294, Flushing, NY 11367-1597, USA
[e]Center for Hydrate Research, Chemical Engineering Department, Colorado School of Mines, Golden, Colorado 80401, USA

ABSTRACT

Gas Hydrates Markup Language (GHML) has been under development since 2003 by the CODATA Task Group "Data for Natural Gas Hydrates" as an international standard for data storage and transfer in the gas hydrates community. We describe the development of this evolving communication protocol and show examples of its implementation. In describing this protocol, we concentrate on the most recent updates that have enabled us to include ThermoML, the widely used IUPAC XML communication standard for thermodynamic data, into the GHML schema for the representation of all gas hydrate thermodynamic data. In addition, a new GHML element for the description of crystal structures is described. We then demonstrate a new tool - Guided Data Capture for Gas Hydrates - for the rapid capture of large amounts of data into GHML format. This tool is freely available and publicly licensed for use by any gas hydrate data producer or collector interested in using the GHML format. An effort will be made to achieve a consensus between scientific journals publishing thermophysical and structural data for gas hydrates to recommend their authors use this new software tool in order to generate GHML data files at the time of the submission of scientific articles. Finally, we will demonstrate how this format can be used to advantage when accessing data from a web-based resource by showing on-line access to GHML files for gas hydrates through a web service.

Keywords: data communication, data capture, database, gas hydrates, GHML, XML

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 alone [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 x 10^{15} m^3 [3] to 1.2 x 10^{17} m^3 [4] at STP, but the amount of organic carbon in

[*] Contribution of the National Institute of Standards and Technology, not subject to copyright in the U.S.
[†] Corresponding author: E-mail:kennethk@boulder.nist.gov

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, in association with an international effort on the part of the Committee on Data for Science and Technology (CODATA), 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 Thermodynamic Research Center (TRC) [11]. The existing database at the core of previous TRC Group activities is 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. 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 [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 have been extended to accommodate property data for gas hydrates.

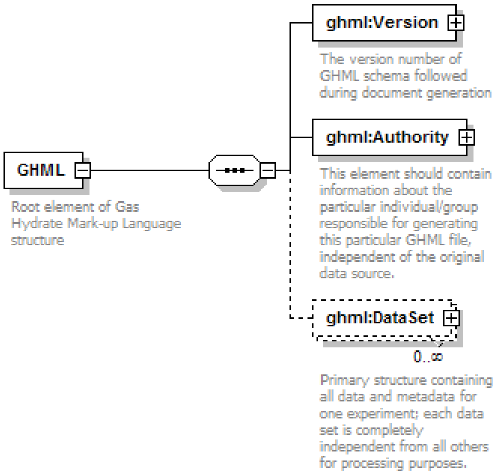

GHML
Root element of Gas Hydrate Mark-up Language structure
ghml:Version
The version number of GHML schema followed during document generation
ghml:Authority
This element should contain information about the particular individual/group responsible for generating this particular GHML file, independent of the original data source.
ghml:DataSet
0..∞
Primary structure containing all data and metadata for one experiment; each data set is completely independent from all others for processing purposes.

Figure 1 Structure of the root element of GHML

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 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 primary focus of development for this data dissemination effort has been the reformulation of the Gas Hydrate Markup Language, or “GHML”, [22-25] to take advantage of the data transfer experience gained through development and use of the IUPAC-standard ThermoML [19-21]. Significant restructuring of the schema was required in order to make GHML consistent with the design philosophies that have contributed to the success of previous data collection efforts. 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 to allow storage and transfer of a broad range of data 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 expansion.

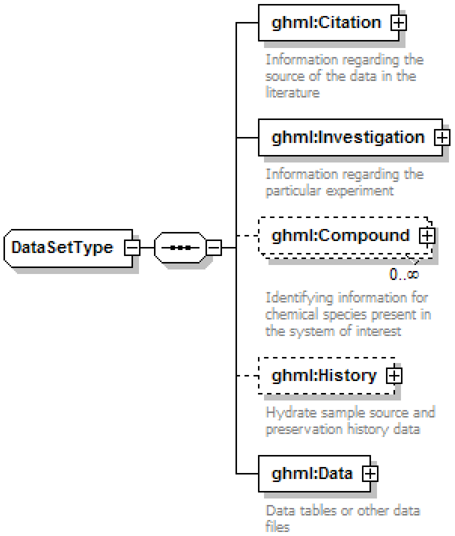

DataSetType
ghml:Citation
Information regarding the source of the data in the literature
ghml:Investigation
Information regarding the particular experiment
ghml:Compound
0..∞
Identifying information for chemical species present in the system of interest
ghml:History
Hydrate sample source and preservation history data
ghml:Data
Data tables or other data files

Figure 2 Structure of the DataSet element in GHML

The root element of this newly generated GHML (Figure 1) includes metadata structures “Version” and “Authority” for designating the individual or group responsible for the XML representation and which version library was used, facilitating very long-term support for groups participating in this data sharing effort. An arbitrary number of individual data records and their associated metadata can be reported in the remaining root element structure, DataSet.

A key element in data sharing is maintaining the link between the numerical data sets and the information on how they were generated. These metadata are often significantly more voluminous than the numerical data associated with them and significantly more difficult to characterize, but are essential both for giving credit to the experimentalists whose efforts yielded this fruit and in understanding strengths and weaknesses of a given data set. Information of this type is stored in the DataSet element (Figure 2). Of particular note is the proven “Citation” element (Figure 3) taken from the ThermoML structure of the same name and new to GHML, which allows data sources to be fully described. Note also the inclusion of a “History” element; as experimental studies of gas hydrate samples from the field are frequently transported a significant distance for *ex situ* analysis, this structure allows full description of the samples source and conditions in transit.

Another important feature of GHML is significant structural flexibility. The fields of study that fall under the purview of the term “gas hydrates” are quite varied in the types of data they report and in the level of detail a single data source may present. The flexibility is expressed in two primary ways. First, rather than attempting to generate a tree structure that explicitly contains all possible data of interest to the field, a series of data labels have been distributed among a set of dictionary files. Each of these files can then be maintained by experts in that category of data types, such as chemical properties or petrology, and any individual interested in communicating via the software need only be concerned with those dictionary files pertinent to the field of interest. Second, as the number of chemical compounds that may be of interest for a given piece of work is very large, chemical species are not explicitly listed within the structure. Rather, chemical distributions such as mole fraction use an identifying key to create a relation with a compound record, which in turn uses the IUPAC International Chemical Identifier (“InChI”) [26] as a unique and freely available chemical identifier, capable of including complex structural and isotopic information while maintaining human readability for simple compounds.

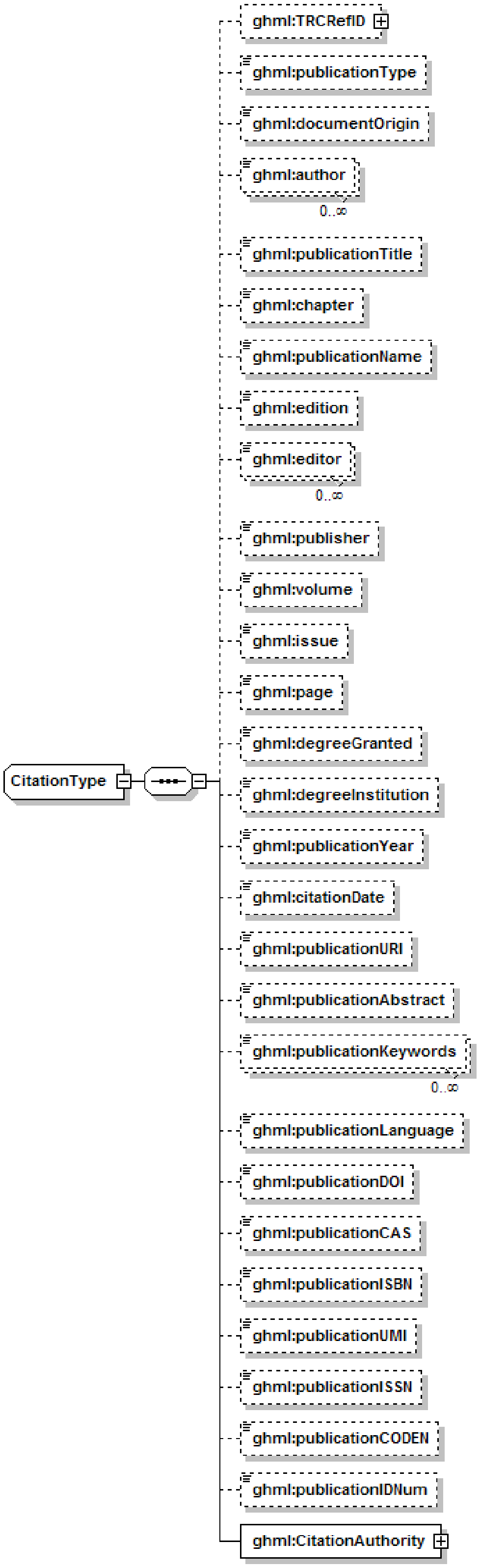


Figure 3 Structure of the Citation element in GHML

GUIDED DATA CAPTURE SOFTWARE

Thermodynamic data from original data sources are not entered directly into SOURCE but are captured or “compiled” in the form of batch data files (coded ASCII text) under normal operation at TRC. This allows application of extensive completeness and consistency checks during the capture process before the data are loaded into the central repository. Extensive expertise has traditionally been required for data compilation due to the complexity of the properties and chemical systems involved. 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) 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.

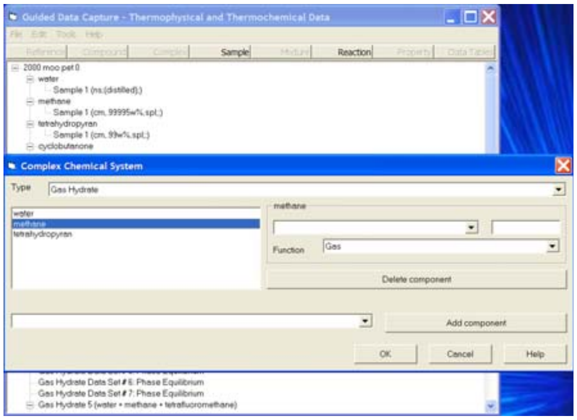
With the development of collaborations with major peer-reviewed journals for the capture of experimental data as they are published, an additional role for GDC evolved. In addition to the creation of batch data files for loading into the SOURCE archive, GDC simultaneously creates a separate text document 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.

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 by 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 [27].

GDC guides 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. It will be shown that 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, allowable codes, etc. The users of GDC are typically 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 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. GDC has features that can readily detect some 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 is in the literature [17].

The existing GDC software required significant modification in order to capture experimental data sets pertaining to gas hydrates. Data processed through GDC have been historically for either a pure compound or a mixture of a small number of well-defined compounds in well-defined ratios. A gas hydrate is a nonstoichiometric structure where studies that may not ever measure crystalline composition can still yield valuable information. Whereas it might be desirable simply to designate such studies as unreliable, the comparative paucity of explicitly constrained data precludes such a determination. The solution to this conflict was determined to be the creation of an original data structure within the GDC framework that 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 equilibria 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


Guided Data Capture - Thermophysical and Thermochemical Data
File Edit Tools Help
Reference Compound Complex Sample Mixture Reaction Property Data Tables
Complex Chemical System
OK
Cancel
Help
Delete component
Add component

2000 moo pet 0
water
Sample 1 (ns:(distilled);)
methane
Sample 1 (cm, 99995w%,spl;)
tetrahydropyran
Sample 1 (cm, 99w%,spl;)
cyclobutanone

Type: Gas Hydrate

water
methane
tetrahydropyran

methane
Function: Gas

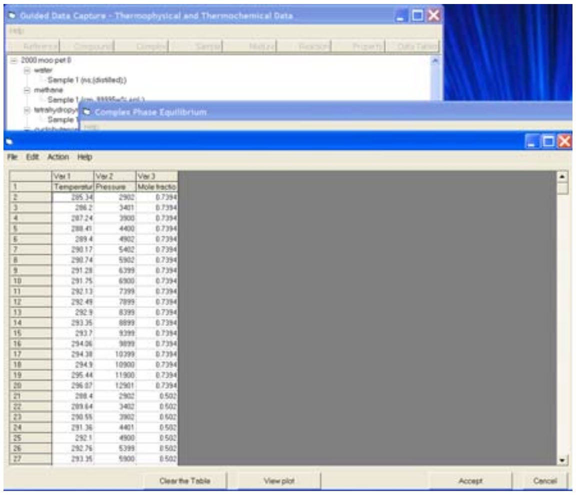
Gas Hydrate Data Set # 6: Phase Equilibrium
Gas Hydrate Data Set # 7: Phase Equilibrium
Gas Hydrate 5 (water + methane + tetrafluoromethane)

Figure 4 Screen capture of GDC dialog for definition of a gas hydrate system.

an entirely new development for GDC, as no previously existing formulation existed within the 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 after 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 is organized around the source document. Following from that are definitions of major chemical components in the systems studied and their specific samples with detailed purity information. A gas hydrate system is then defined by a combination of those chemical components (Figure 4), and a gas hydrate sample is defined through the association of the samples of those components, as well as the conditions under which the hydrate was formed, if appropriate. It is only when 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.

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


Guided Data Capture - Thermophysical and Thermochemical Data
Complex Phase Equilibrium
File Edit Action Help
Clear the Table
View plot
Accept
Cancel

	Var 1	Var 2	Var 3
1	Temperatur	Pressure	Mole fractio
2	295.34	2902	0.7394
3	286.2	3401	0.7394
4	287.24	3900	0.7394
5	288.41	4400	0.7394
6	289.4	4902	0.7394
7	290.17	5402	0.7394
8	290.74	5902	0.7394
9	291.28	6399	0.7394
10	291.75	6900	0.7394
11	292.13	7399	0.7394
12	292.49	7899	0.7394
13	292.9	8399	0.7394
14	293.35	8899	0.7394
15	293.7	9399	0.7394
16	294.06	9899	0.7394
17	294.38	10399	0.7394
18	294.9	10900	0.7394
19	295.44	11900	0.7394
20	296.07	12901	0.7394
21	288.4	2902	0.502
22	289.64	3402	0.502
23	290.55	3902	0.502
24	291.36	4401	0.502
25	292.1	4900	0.502
26	292.76	5399	0.502
27	293.35	5900	0.502

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

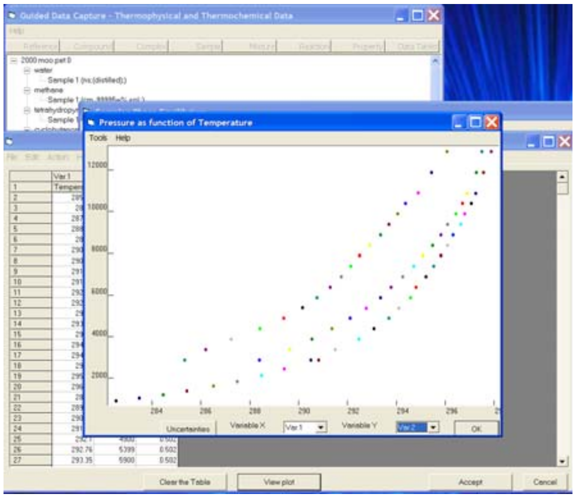
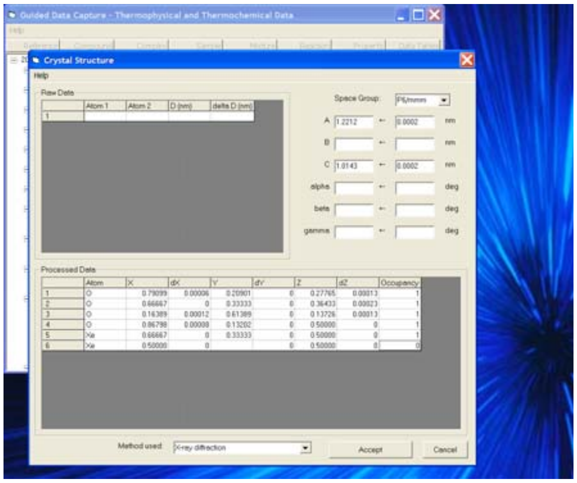


Figure 6 Screen capture of natively generated graph of data entered into GDC tabulated data dialog

Enthalpy of decomposition of a gas hydrate is a function of the crystal occupancy, thus proper characterization requires a well-defined ratio of host to guest molecules. GDC treats such processes as physical reactions, with the enthalpies of decomposition stored as enthalpies of reaction. This methodology has additional benefits in that, as the comparatively slow kinetics associated with hydrate formation and the dynamics of hydrate



	Atom 1	Atom 2	D (nm)	delta D (nm)
1				

Space Group: P6/mmm

A	1.2212	←	0.0002	nm
B		←		nm
C	1.0143	←	0.0002	nm
alpha		←		deg
beta		←		deg
gamma		←		deg

	Atom	X	dX	Y	dY	Z	dZ	Occupancy
1	O	0.79099	0.00006	0.20901	0	0.27765	0.00013	1
2	O	0.66667	0	0.33333	0	0.36433	0.00023	1
3	O	0.16389	0.00012	0.61389	0	0.13726	0.00013	1
4	O	0.86798	0.00008	0.13202	0	0.50000	0	1
5	Xe	0.66667	0	0.33333	0	0.50000	0	1
6	Xe	0.50000	0		0	0.50000	0	0

Method used: X-ray diffraction

Figure 7 Screen capture of GDC dialog for storing crystallographic data, including space group, unit cell parameters and atom distribution

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.

Experimental measurement techniques for bulk property measurements, such as mass specific volume, thermal conductivity, heat capacity at constant pressure per unit mass, or speed of sound, 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, after which the conditions under which the measurements must be defined. The numerical data are captured and existing TRC internal methods are used to estimate the reliability of the provided data 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 to 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 7). Modeling this new data structure upon the Crystallographic Information File (CIF) format, an International Union of Crystallography (IUCr) standard used prevalently within the crystallographic community for communication of experimental results [28], provides a reasonable guarantee of generality and compatibility with likely crystalline structure data sets.

DATABASE ARCHITECTURE

NIST is prsuing the 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 [16] as implemented in TDE [14-15].

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 [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 more than three million numerical values for many thousands of pure compounds, binary and ternary mixtures, and reaction systems.

Conflicts similar to those pertaining to GDC development required extensions to SOURCE to

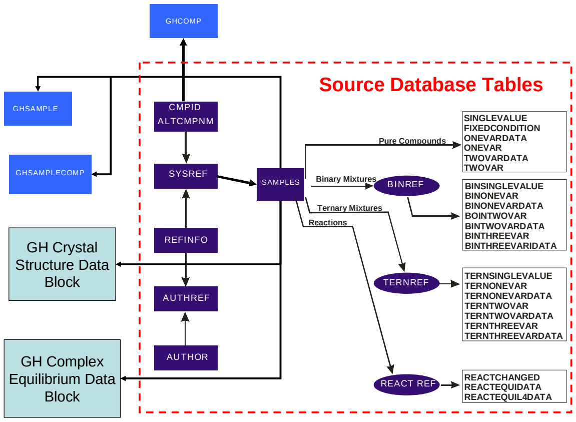


Figure 8 Schematic representation of TRC SOURCE database (inside red dashed line) with modifications required to accommodate storage of gas hydrate (GH) data

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 8. 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, contains fields for storing a gas hydrate registry number and constituent compound information. The registry number uniquely identifies each combination of hydrate formers. Stored compound information includes component registry numbers, whether this compound functions as a host or guest, and the composition metric, with associated numerical values.

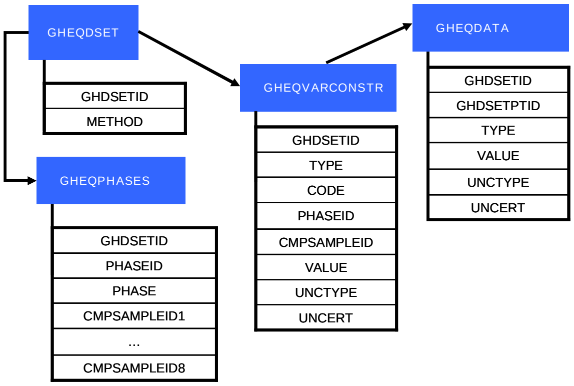


Figure 9 Structure of new GHEQDSET, GHEQVARCONSTR, GHEQDATA and GHEQPHASES tables for storage of phase equilibrium data

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. Linked by a sample-specific unique identifier, these two tables contain information regarding the state of the specific sample as a whole and references to the individual component samples and the particulars of their molar distribution.

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 for future accountability. Data sets are divided into three categories: bulk physical properties, crystallographic structure information and complex phase equilibria. For bulk physical properties such as mass specific volume, thermal conductivity, heat capacity at constant pressure per unit mass, or speed of sound, SOURCE has previously existing data structures that remain valid for hydrates, so no further modifications were 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 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 [28].

Characterizing 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 9) to record measurement method (table GHEQDSET), the phases present and compounds present therein (table GHEQPHASES), the types and values of constraints 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 eight components are allowed. A second unique identifier 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. As per previous implementations of SOURCE, care is taken as well to 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).

CONCLUSIONS

An effort toward developing an internationally available hydrate data dissemination channel is underway in the U.S. National Institute of Standards and Technology. This effort will include critical evaluation and comparison of gas hydrate data sets. Information dissemination will be accomplished through the gas hydrate markup language (GHML), which 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. Data Quality Assurance is facilitated though guided data capture (GDC), 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. 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.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support provided by the U.S. Department of Energy through the National Energy Technologies Laboratory via Project Number DE-AI26-06NT42938. We also greatly appreciate the significant effort by the gas hydrate community in general toward the development of the Gas Hydrate Markup Language.

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 14th March 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.* 2006. [online]. [Accessed 14th March 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 14th March 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 14th March 2008]. http://www.iupac.org/projects/2002/2002-055-3-024.html.

[21] *ThermoML*. [online]. [Accessed 14th March 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 14th March 2008]. 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.