GAS HYDRATE MARKUP LANGUAGE: LABORATORY DATA

Tom Smith[1*], John Ripmeester[2], Dendy Sloan[3], and Tsutomu Uchida[4]

*[*1] M.I.T. Systems Inc, USA*
Email: tom.smith@cinaplex.com
[2] The Steacie Institute for Molecular Sciences, National Research Council of Canada, Canada
Email: John.Ripmeester@nrc.ca
[3] Center for Hydrate Research, Colorado School of Mines, USA
Email: esloan@mines.edu
[4] Hokkaido University, Japan
Email: t-uchida@eng.hokudai.ac.jp

ABSTRACT

Laboratory Hydrate Data is one of the three constituent modules comprising the XML based Gas Hydrate Markup Language (a.k.a. GHML) schema, the others being Field Hydrate data by Löwner et al. and Hydrate Modeling by Wang et al. This module describes the characteristics of natural and synthetic gas hydrates as they pertain to data acquired via analysis within a laboratory environment. Such data include the preservation history (i.e.: technique, pressurization gas and pressure, etc), Macroscopic data (i.e.: water-sediment ratio, appearance, P-T behavior, etc) as well as that of the Microscopic realm.

Keywords: Hydrate, Database, Gas Hydrate Markup Language, GHML, XML, Laboratory data, Field data, Modeling, Simulation

1 INTRODUCTION: MOTIVATION FOR THE GAS HYDRATE MARKUP LANGUAGE

The ever increasing volume of data in modern society coupled with the critical need for simplified and efficient sharing of said data is exemplified in almost every sector of society. Nowhere is this perhaps more evident than the scientific community where data located at one research facility could prove invaluable to another. Historically it has been sufficient to merely store such data in isolated and disparate databases for later retrieval and reporting. Data was then subsequently requested and provided to third party entities in a variety of ad-hoc formats leaving the researcher with the unfortunate and daunting task of having to process these various non-standardized extracts, rather than being able to focus on the actual research at hand.

It has now become absolutely essential to facilitate the exchange of data in internationally standardized and accepted formats such as in the Gas Hydrate Markup Language (aka: GHML) pioneered and developed by the CODATA Gas Hydrate Data Task Group. The GHML is an Extensible Markup Language (aka: XML) based implementation and standard, which is readily designed to allow the modeling and subsequent exchange of data pertaining to the more common Gas Hydrate constructs encountered in the research environment. By careful coordination with the Gas Hydrate research community, a workable and viable GHML schema has effectively been realized.

Concurrent with establishing the architecture of the GHML, existing related markup languages and standards were investigated and researched in detail to help ensure compliance with industry accepted standards and practices where applicable. As such, careful concern was given to such items as the integration of preexisting markup language constructs, enumeration, naming conventions, attributes, and abstraction

This paper is meant to serve as an overview to the laboratory data portion of the Gas Hydrate Markup Language. Though fairly descriptive, it is not designed nor meant to serve as or replace the GHML documentation. Where advantageous and where clarity is best served, details have been included herein. For additional details outside the scope of this paper, please refer to the GHML documentation.

2 LABORATORY DATA SCOPE OF COVERAGE

The laboratory data specifically focuses on data gathered within the laboratory setting as well as the preservation history and basic origin information of the sample. These data include certain relevant metadata for the sample, the source and conditions involving, for example, the extraction of a natural hydrate sample, its preservation history, and macroscopic qualities. A section for microscopic and mesoscopic details has been incorporated for anticipated future expansion.

The module does not include field-specific related data such as those pertaining to boreholes, lithography, etc., nor does it contain provisions for numerical modelling. Thus for these one would need to refer to the Field Hydrate data module by Löwner et al (2007) or the Numerical Modeling data module by Wang et al. (2007), respectively, which encompasses the heretofore mentioned data items.

The primary purpose of the schema is to provide a standardized method by which to communicate gas hydrate data amongst potentially disparate and unrelated organizations across the internet. Because it is meant to communicate gas hydrate data, the scope of what can and can not be exchanged has necessarily been limited, and thus the schema itself provides a form of validation check on the XML documents that are to be exchanged.

3 STRUCTURE OF SCHEMA

The laboratory data portion of the Gas Hydrate Markup Language is constructed in a clear modular format which makes it easy to understand and implement as well as add on additional components as needed to its schema. Consisting of five primary blocks, research and requirements gathering were done to ensure that the schema was carefully modeled to reflect the way in which data are currently gathered and recorded in the laboratory environment, as opposed to a somewhat more idealized or abstract method. This method was chosen in order to help facilitate the integration of the GHML with existing databases and data efforts across the research community. Because it is modeled in this fashion, the GHML may necessarily diverge from certain industry standards and/or recommendations in order to meet current data sharing needs. The following figure illustrates the top level tags for the laboratory GHML.

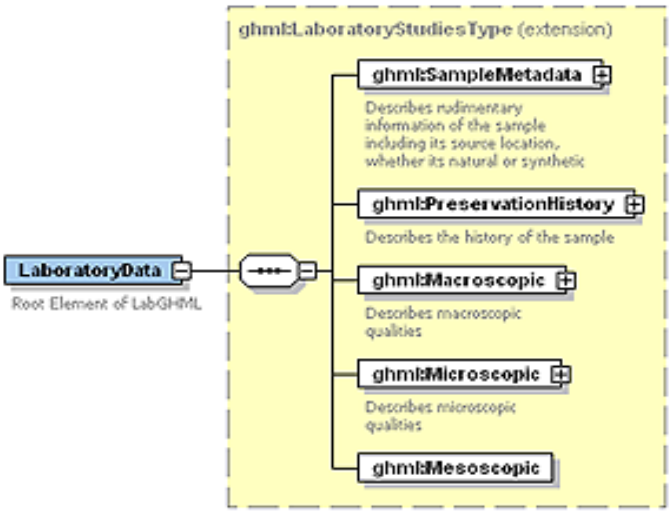

Figure 1. Top level tags for the laboratory GHML

In the laboratory GHML, the schema elements/types are directly analogous to fundamental scientific constructs within the Gas Hydrate community. This method (as opposed to the equally valid choice of implementing a layer of abstraction) was chosen in order to help reinforce the understanding and meaning of the data being exchanged thereby leaving little to no room for ambiguity, which is vital for the success of the GHML, especially when considered in an international venue.

3.1 Namespaces

In designing the GHML, there was much discussion on the topic of 'namespaces' and in particular which methodology was best and should be adopted. Essentially there are three possibilities when one considers this:

a) Do not use a default namespace
b) Set the default namespace to XMLSchema
c) Set the default namespace to the target namespace (GHML chooses this)

Each of the above has its associated advantages and disadvantages. Thus the decision needed to be based not on which was the 'best' overall practice (because arguably there really is not one) but rather which was most advantageous for the future of GHML. Because incorporation of other markup languages may indeed occur at some point in the future, namespace qualifying the various schema components was deemed a requirement and thus the consensus was for option 'c'.

The target namespace is: **ghml**

3.2 Naming Conventions

Careful consideration and thought went into the decision regarding the naming of the elements and data types used throughout. As such when one views the schema, it is relatively obvious as to what data are actually being 'marked' up. Not only has considerable thought been given to the descriptive names of the various 'tags', but equal concern was given to the 'case' of said tags. With that in mind, the GHML adopts a combination of Pascal case, Camel case, and Uppercase for the various tags used throughout as follows:

- Pascal Case: First letter of each concatenated word being capitalized – e.g.: PascalCase
- Camel case : First letter is lowercase and first letter of each subsequent concatenated word is capitalized – e.g.: camelCase
- Upper case : Every letter of the tag is capitalized – e.g.: UPPERCASE

Given below is the outline of the standard to which GHML adheres; however, depending on the meaning, context, and/or source (i.e.: another applicable ML) of the tag, we may decide to diverge from the below standard:

- Complex Types : PascalCase + the word 'Type'
- Simple Types : camelCase + the word 'Type'
- Elements : PascalCase
- Attributes : camelCase, PascalCase, Uppercase

3.3 Enumeration, Units, and Attributes

Where feasible the realm of possible data values has been restricted to either an enumerated list or some range of permitted values. Enumerated lists follow their proper case and thus, for example, the chemical representation of methane would be represented as CH4.

Another example of this would be the enumerated element entitled: *Appearance*, which is limited to having the following possible values:

- Massive
- Nodular
- Pore Hydrate

Attributes are the exception rather than the norm in the GHML. Thus they have been implemented sparingly and where it was decided that their use was a clear advantage; otherwise elements have been opted for in their stead. An example of a clear advantage is in the specification of units. As such, where applicable, the attribute 'uom' has been incorporated to clearly denote the units of the respective data in question.

In general standardization on SI-type units, or some multiple of, has been chosen throughout the GHML. Some exceptions to this choice do exist so as to ease and facilitate the integration of data that may exist in other units. Concern was specifically given to circumstances where conversion of such data might prove infeasible or difficult at best given available resources.

One such example is that of *pressure*. Pressure typically should be recorded in the unit of *Pascal*, but data do exist in pounds per square inch. Thus *psi* has been incorporated to address this current reality in the research sector.

3.4 Synopsis of Laboratory Data GHML

Consisting of five primary blocks the Laboratory data GHML is outlined as follows:

- *SampleMetadata* – Describes details of the source of the sample.
- *PreservationHistory* – Describes the history of how the sample has been preserved including its Pressure/Temperature behaviour as a function of time.
- *Macroscopi*c – Describes various macroscopic qualities of the sample such as its appearance, colour, water/sediment ratio and so on.
- *Microscopic* – Though not fully implemented in this Beta version, the anticipated purpose of this block will be to describe any desired microscopic qualities.
- *Mesoscopic* – That which does not clearly fall into the realm of Macroscopic or Microscopic, will be placed within this block. Currently this is stub block which does not as of yet have any elements/types within the current Beta version.

In the following subsections, each of the five heretofore mentioned blocks will be overviewed in summary. For a detailed explanation please see the GHML documentation.

3.4.1 Sample Metadata Block

This block describes metadata information related to the origin of the sample. The following outlines the associated top level tags for this block:

Source	Indicates whether the sample was sourced from Onshore, Offshore or is Synthetic
SampleID	Indicates the unique identifier for the particular sample at hand. The identifier could be an alphanumeric string.
OriginDate	Indicates the date in which the sample was either extracted or created in the laboratory environment
OriginLocation	Indicates the location of where the sample was sourced from. If it is a natural sample, the location may include latitude, longitude, water depth, permafrost depth, metres below sea floor. If synthetic, it includes the research center name and description.
OriginConditions	Indicates the conditions that existed at the source including the in-situ pressure and temperature.
RecoveryMethod	Indicates the method of recovery such as: piston drop core, autoclave, pressure core, pressure temperature core, ROV
InvestigationData	Information related to the source dataset, which includes the analysis date, dataset name, data file, owner, contact information, responsible parties, researcher, comments.

3.4.2 Preservation History Block

This block contains information related to the history of the preservation of the sample. The following outlines the top level tags for this block:

Technique	Indicates the technique utilized for the preservation of the sample such as internal pressurization, external pressurization, liquid nitrogen
PressurizationGas	Indicates the gas used for pressurization such as He, N2, CH4
ContainerConditions	Indicates the conditions of the container such as the pressure and temperature
PTBehaviour	Indicates the pressure and temperature behaviour over time. Data considered

	include the phases (e.g.: LHc, V, H1, H2, Hh, I), pressure, temperature, component and mole fraction.

3.4.3 Macroscopic Block

This block contains information related to the macroscopic qualities of the sample. The following outlines the top level tags for this block:

Appearance	A physical description of the sample
Colour	A description of the color of the sample described by the researcher.
WaterSedimentRatio	Indicates the ratio of water mass to sediment mass
GasWaterRatioSTP	Indicates the ratio of gas volume to water volume
PTBehaviour	Indicates the pressure and temperature behaviour over time. Data considered include the phases (e.g.: LHc, V, H1, H2, Hh, I), pressure, temperature, component, and mole fraction.
Gas	Description of the gas component(s) of the sample as well as isotopic analysis data. The gas component portion contains both the gas (such as Ethane, CH4, etc.) and its related mole percent. The isotopic analysis portion contains the gas component, the isotope as well as delta value information.
Water	Contains information regarding ion concentration as well as isotopic analysis. Ion concentration contains the ion, mass percent, mole percent and parts per million. The isotope analysis portion contains the gas component, the isotope as well as delta value information.

3.4.4 Microscopic Block

This block contains information related to the microscopic qualities of the sample. Though currently under development, the following outlines the top level tags for this block thus far:

Morphology	Currently under R&D
XRayDiffraction	Currently under R&D

3.4.5 Mesoscopic Block

This entire block, which is currently under consideration and development, is intended to contain information related to that which does not clearly fall within the realm of the microscopic or macroscopic categories.

3.5 Uncertainty Data

It was realized during the development of the GHML that there needed to be provisions to exchange uncertainty data for the various measured values used throughout. As such, each measured value has the provision to carry with it any related uncertainty data..

By a careful review of NIST technical note 1297 entitled 'Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results,' a comprehensive schema for the uncertainty data was architected. This document was chosen as a basis for two reasons: a) It is based on another document entitled "Guide to the Expression of Uncertainty in Measurement" (a.k.a. GUM) by the International Organization for Standardization (a.k.a.: ISO), and b) By adopting these standards, possible integration with ThermoML (Frenkel, et al., 2006) might be more readily facilitated since ThermoML also adopts similar standards.

The following are the uncertainty information which is carried along with each measured value:

- Evaluator
- Evaluation Method
- Standard Uncertainty (Uc)
- Expanded Uncertainty (U)
- Coverage Factor (k)
- Confidence Level

For detailed explanation of the above items please refer to NIST document TN 1297 (Taylor & Kuyatt, 1994).

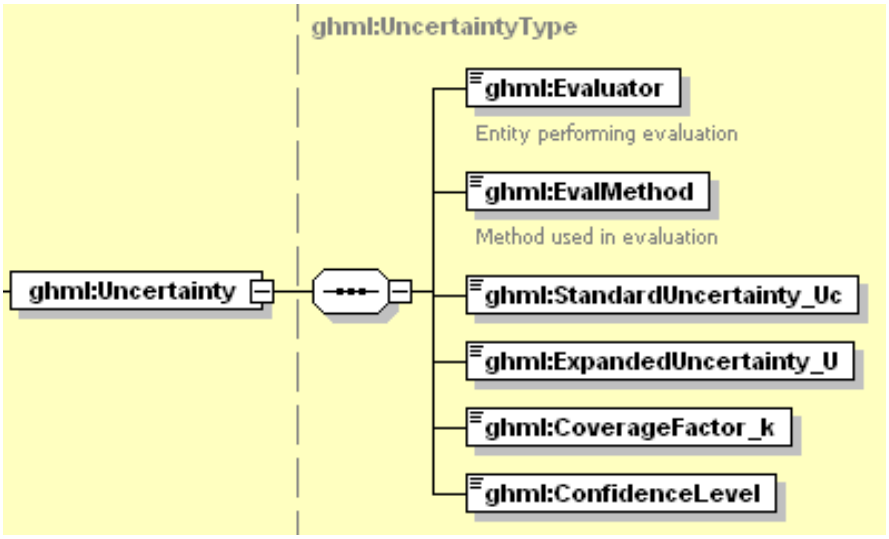

Figure 2. Top level tags for uncertainty information in GHML

4 OUTLINE OF A POTENTIAL GHML SYSTEM

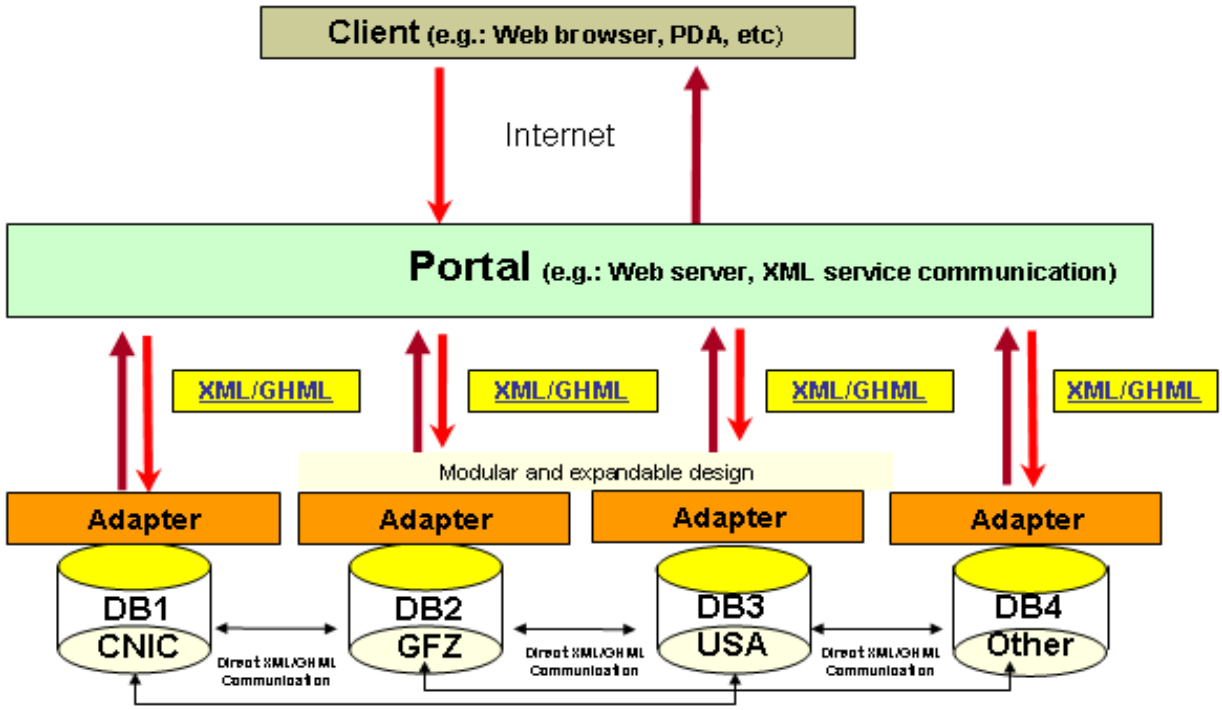

Figure 3. Diagram of potential GHML service system

The above diagram depicts one method by which the GHML can be leveraged. To begin, a user (utilizing a web browser) can connect to the 'portal' (scheduled to begin development during 2007) and request whatever information so desired. The portal would then create an XML document (encapsulated via SOAP) containing the request for said data. This document would then be communicated to the adapter of each of the participating organizations. Each adapter will then parse the XML request, search its associated database, place the response back into an XML document (conforming to the GHML), and then send that response on to the portal for the user's viewing.

The central theme of the portal is to communicate the user requests to all of the participating organizations, collect the responses, and then return those responses to the user. In this way the user is presented with a single standardized and central location by which to do their research rather than multiple ones with varying and questionable interfaces.

In addition to direct user requests via a web browser, it is envisaged that the portal will contain one or more web services capable of automatically communicating Gas Hydrate data amongst the community. By adhering to certain standards, organizations can create web services of their own, which would allow automated data sharing and exchange without the necessity of direct user intervention.

Furthermore one can also envisage communication occurring directly among the various participating organizations (via the adapters and XML) thereby bypassing the Portal directly. This ability is inherent in the architectural design of the system and can be implemented on an 'as needed' and/or 'as desired' basis. Reasoning for doing this might include data replication between databases, consistency/validation checks with other datasets, and so on. In summary, by implementing the above described architecture as well as incorporating certain industry standards/practices throughout, a true service oriented architecture (a.k.a.: SOA) can be realized.

5 CONCLUSION

The laboratory portion of the Gas Hydrate Markup Language is currently in beta revision and is an evolving work in progress. By careful integration with the other two portions of the GHML (i.e.: Field and Modelling), a fairly complete and robust markup language capable of communicating and sharing gas hydrate data across networks and the Internet results.

6 ACKNOWLEDGEMENTS

The authors gratefully acknowledge support by several agencies. CODATA contributed over the course of this project to support travel costs. The Computer Network Information Center (CNIC) of the Chinese Academy of Sciences donated two IT professionals to develop the GHML. GeoForschungsZentrum (GFZ) - Potsdam donated one full time IT professional to the project in 2006. The USA Department of Energy has supported the participation of an IT consultant and provided partial travel support for some members of the CODATA task group.

7 REFERENCES

Frenkel, M., Chirico, R.D., Diky, V.V., Dong,Q., Marsh, K.N., Dymond, J.H., Wakeham, W.A., Stein, S.E., Königsberger, & Goodwin, R.H., (2006) XML-based IUPAC standard for experimental predicted and critically evaluated thermodynamics property data storage and capture (ThermoML), *Pure Appl. Chem.,* 78(3), 541-512, doi:10;1351/pac200678030541.

Löwner, R., Cherkashov, G., Pecher, I., & Makogon, Y. F (2007) Field Data and the Gas Hydrate Markup Language, *Data Science Journal,* 6, GH1-GH12.

Taylor, B. & Kuyatt, E. (1994) Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, Physics Laboratory, National Institute of Standards and Technology, NIST Technical Note 1297

Wang, W., Moridis, G., Wang, J., Xiao, Y.,& Li, J., (2007) Modeling hydrates and the gas hydrate markup language, *Data Science Journal,* 6, GH25-GH36.