FIELD DATA AND THE GAS HYDRATE MARKUP LANGUAGE

Ralf Löwner[1]*, Georgy Cherkashov[2], Ingo Pecher[3], and Y. F. Makogon[4]

[1]**GeoForschungsZentrum Potsdam, 14473 Potsdam, Germany*
Email: loewner@gfz-potsdam.de
[2]*VNIIOkeangeologia, St. Petersburg, Russia*
Email: Cherkashov@mail.ru
[3]*Herriot-Watt University, Edinburgh, Scotland*
Email: ingo.pecher@pet.hw.ac.uk
[4]*Texas A&M University, College Station, Texas, USA*
Email: makogon@pe.tamu.edu

ABSTRACT

Data and information exchange are crucial for any kind of scientific research activities and are becoming more and more important. The comparison between different data sets and different disciplines creates new data, adds value, and finally accumulates knowledge. Also the distribution and accessibility of research results is an important factor for international work. The gas hydrate research community is dispersed across the globe and therefore, a common technical communication language or format is strongly demanded. The CODATA Gas Hydrate Data Task Group is creating the Gas Hydrate Markup Language (GHML), a standard based on the Extensible Markup Language (XML) to enable the transport, modeling, and storage of all manner of objects related to gas hydrate research. GHML initially offers an easily deducible content because of the text-based encoding of information, which does not use binary data. The result of these investigations is a custom-designed application schema, which describes the features, elements, and their properties, defining all aspects of Gas Hydrates.

One of the components of GHML is the "Field Data" module, which is used for all data and information coming from the field. It considers international standards, particularly the standards defined by the W3C (World Wide Web Consortium) and the OGC (Open Geospatial Consortium). Various related standards were analyzed and compared with our requirements (in particular the Geographic Markup Language (ISO19136, GML) and the whole ISO19000 series). However, the requirements demanded a quick solution and an XML application schema readable for any scientist without a background in information technology. Therefore, ideas, concepts and definitions have been used to build up the modules of GHML without importing any of these Markup languages. This enables a comprehensive schema and simple use.

An extensive documentation ensures the usability of the "Field Data" module consisting of a detailed explanation integrated in the application schema, an HTML-based document, and a detailed documentation. Because of the close collaboration of gas hydrate experts and specialists in Geoinformatics, the application schema of GHML is user-oriented and contains all possible aspects of this research field. The usability is the assessment factor for GHML.

Keywords: Hydrate, GHML, XML schema, Field data, Data exchange, Information management

1 INTRODUCTION

A critical necessity to manage and exchange worldwide existing data and information resources throughout the different research fields is creating sustainable scientific results and avoiding duplicated work. New Technologies and the Internet permit the integration of worldwide distributed data sources in a virtual data infrastructure.

A number of gas hydrate research databases currently exist in different institutes and countries. They are all strongly proprietary, heterogeneous, and build upon different data models incompatible with each other. To integrate all these databases in one single data infrastructure in order to consult, compare, and combine all the existing data and information, a common communication language or exchange format is crucial as a first implementation step. This language could be described as a central virtual new data model, which integrates all the different data structures and thus enables communication between them. The present work describes the realization and structure of such a data model. It concerns the field data model as a part of gas hydrate research and as a component of the Gas Hydrate Markup Language (GHML), a data model for all possible gas hydrate research investigations. This field data module is used for all data and information coming from the field. It is one of three constituent modules being laboratory data (Smith et al., 2007) and hydrate modeling data (Wang et al., 2007), which are not object of this paper.

The present work is a result of a collaboration of three different institutes in the framework of the task group "Data on Natural Gas Hydrates" of CODATA (Committee on Data for Science and Technology). The Center for Hydrate Research at the Colorado School of Mines, USA, the Computer Network Information Center (CNIC) of the Chinese Academy of Science (CAS), China, and the GeoForschungsZentrum Potsdam (GFZ Potsdam), Germany, combine efforts in the development of a Gas Hydrate Markup Language (GHML).

The cooperation with CODATA is based on the work for the ICDP Mallik2002 database, developed at the GFZ Potsdam. This database was created for an international drilling project for gas hydrate research in North-West Canada in 2002, with the contribution of ICDP (International Continental Scientific Drilling Program, www.icdp-online.de). The ICDP Mallik Data and Information System provide a database prepared for most project data and its secure restricted dissemination and distribution. It operates as a communication platform between the project members through the Web portal of the ICDP Information Network (Löwner et al., 2005).

Both the field data part and the complete Gas Hydrate Markup Language are described in an XML application schema. With regard to the requirements, a model-driven process by automated translation to XML schema from conceptual models defined in other conceptual schema languages such as UML could not be implemented. In contrast, the present schema is constructed by hand using a specialized XML schema editor and taking the XML schema as a conceptual schema language. This approach is recommended as one of the two possible methods for schema construction in the specification of GML (Cox et al., 2004).

The work on the data model was realized in close collaboration with scientific experts with wide-ranging experience in gas hydrate research and with potential users of the resulting products. This promotes the acceptance and usability of the results. The modeling work has been done with the minimum of abstraction needed for a robust and consistent model. A detailed description of the elements and types is included in the

XML schema file as annotations and permit an easy understanding of the content. This effort could be considered a pilot project and could become a model for other fields of scientific research investigations.

2 REQUIREMENTS

The goal of this project was the construction of a common language based on a data model for data and information transport and storage. The focus was set on a global, easily-readable description of both data and metadata including all aspects of research on gas hydrates and their additional research fields. The construction of the field data model as a component of GHML was conducted by requirements caused by deadline constraints and the acceptance in the scientific community. However, the result is a “quick win solution”, which is self-contained, while keeping the capability of integration into GHML. Due to the self-explanatory nature of the schema model, the field data part is clearly comprehensible for all researchers without any background in Information Technology.

Therefore, the realization is based on a relatively simple schema model, which dispenses with substitution groups, abstract types, imports, and includes. No namespaces other than default (xsd:) and target namespace (ghml:) are defined. Most of the elements are optional, which makes the schema very flexible. The names of the elements are describing their denotation and were determined together with the scientific community.

The naming convention corresponds to the other parts of GHML. The definition of complex types use UpperCamelCase notations followed by the word “Type” (e.g., FieldDataType), simple types = definition uses lowerCamelCase notations followed by the word “Type” (e.g., doubleOrNullType), and elements are described by an UpperCamelCase notation (e.g., FieldData).

3 STRUCTURE

Modeling was performed directly in XML schema. Therefore, the structure of GHML is strictly hierarchical, and the field data portion is set as an element (FieldData) on the second level under the main GHML element (see figure 1), alongside the other GHML components.

The “FieldData” element is optional and could be infinitely repeated within an XML instance document. An identification number could be assigned to each instance of this element. It consists of three other optional elements on a third level (see figure 2): the “ProjectMetaData” element contains all principal metadata of the project from which field data results. This could be information about the investigators or the objectives of the project. The other two elements define both, data and metadata.

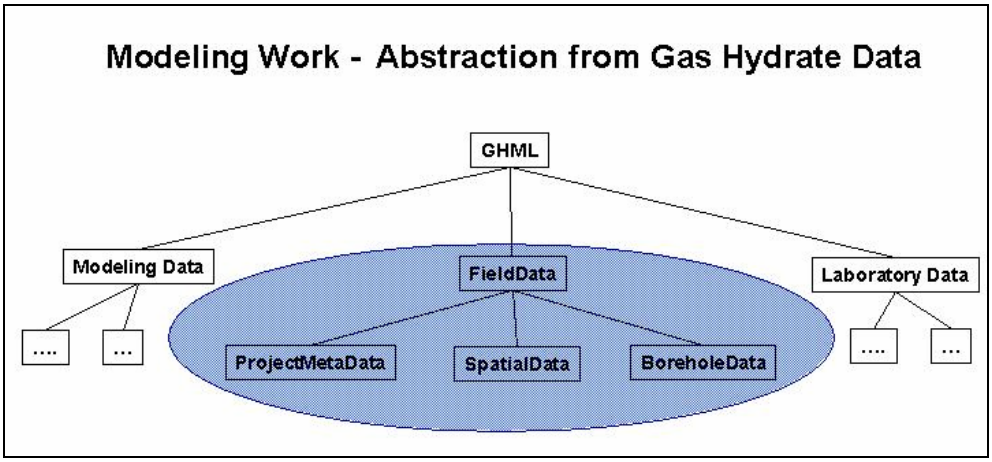

Figure 1. Field data portion within the GHML (Gas Hydrate Markup Language)

The "SpatialData" element describes all spatial relevant data and metadata content gathered in the field. In here, all values are related to a surface e.g., outcrop analyses. The "BoreholeData" element describes all data and metadata content coming from borehole investigations. This implies that values are mainly vertically related, e.g. to depth or time. All these elements contain optional identification numbers as attributes which enables a better adaptation to databases.

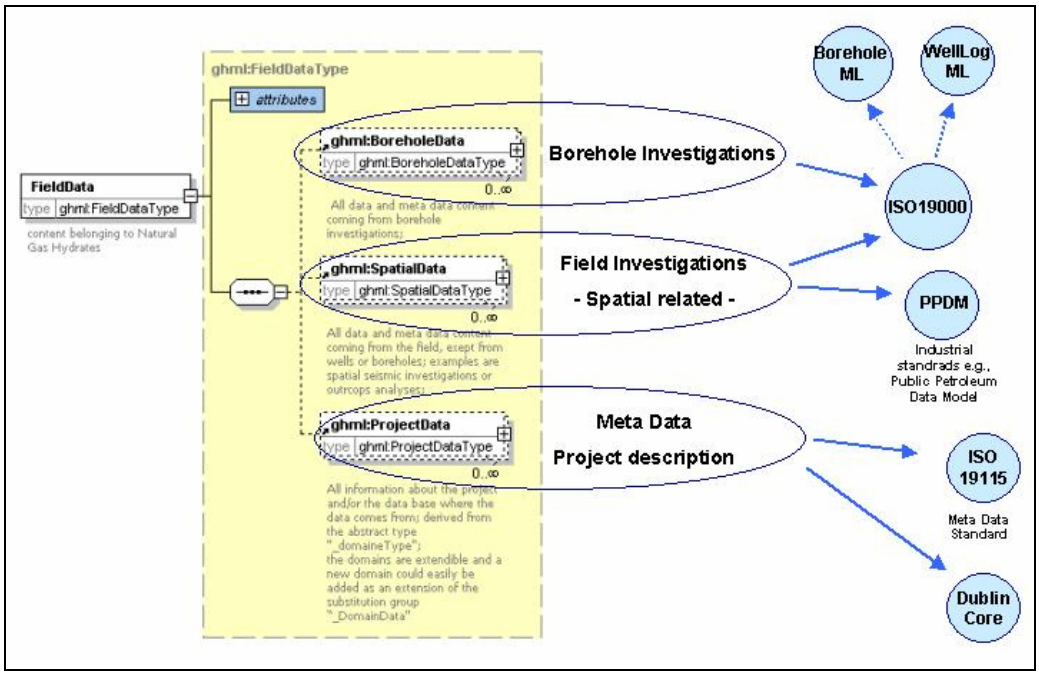

Figure 2. Field data portion with the three different elements on the third level, showing also the relations to already existing standards

With regard to the complexity and the "quick win solution" aspects of the project, GHML does not import any other international standards. The field data portion, however, integrates some structures or concepts from other standards to provide exchangeability and extensibility. Various relevant standards were analyzed and compared during the requirement analysis. Primarily, basic elements of the Geography Markup Language (GML) and other parts of the ISO19000 series are used for spatial related features. Also other standards were analyzed, e.g., WellLogML for elements related to drilling activities.

According to most of the accepted international standards, the use of global elements and types permits the exchangeability and extensibility of the field data portion. The structure in all of the three elements is defined by a relatively small number of global types. The result is an easily understandable data model, which does not need the knowledge of each element or parameter. Because of the similarity of each element, the following notes describe only the "BoreholeData" element as an example.

The optional "BoreholeData" element consists of a number of attributes and other elements (see Figure 3). The attributes describe direct information about the special borehole, such as position and identification number. Only the name of the well hole is set as mandatory. Besides the attributes, one can find a list of other optional unbounded elements, which represent research fields or domains. Geological, Geophysical, Geochemical, and Biological investigations are provided, but this list could be enlarged in future. These elements are listed under the fourth level of XML hierarchy.

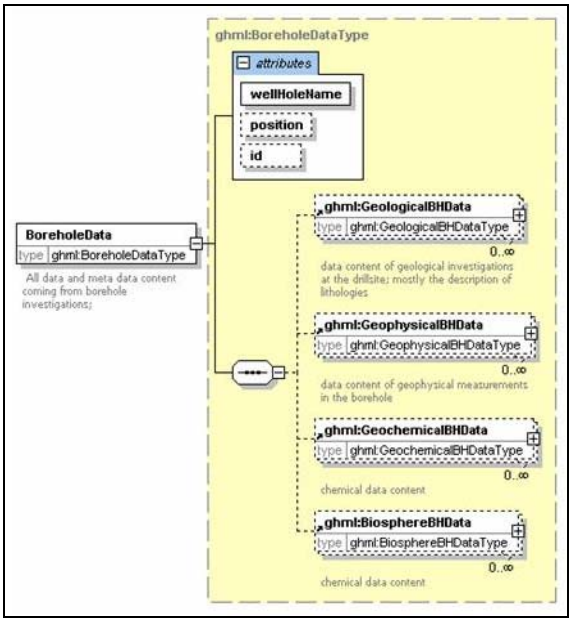

Figure 3. Borehole data element and its different underlying optional elements representing the different Gas Hydrate research investigation in a drilling well.

The different investigation fields or domains contain activities themselves under a fifth level (see Figure 4), e.g. the Geochemical investigation contains analyses of gas data (GasData) and analyses of the water column (WaterColumnData). These lists can be enlarged in future and adapted to other and new research fields and activities.

All the activities are built up the same way. They can be added or deleted according to the requirements. This enables a flexible model, which can be adapted to each new scientific research field. New types and elements can be created by the extension of two types: the "InvestigationDataType" and the "MeasurementDataType" (see Figure 5). The first describes the metadata of the data set, and the second contains the data set itself. Only a few elements are included in each type, so the extension will be adapted to each data set.

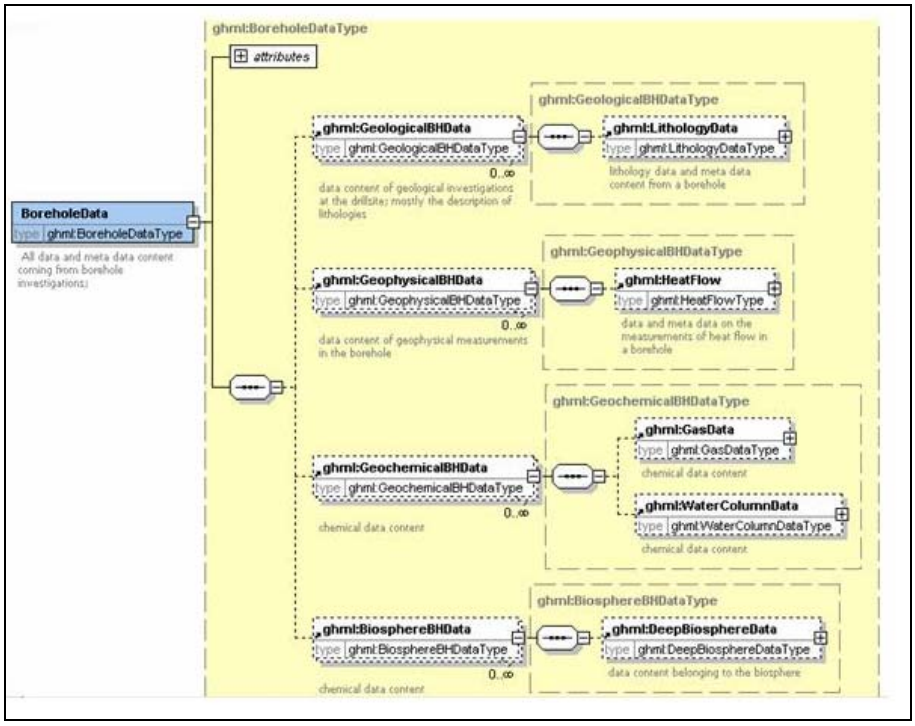

Figure 4. Borehole data element, its different gas hydrate research investigation elements, and underlying research activities

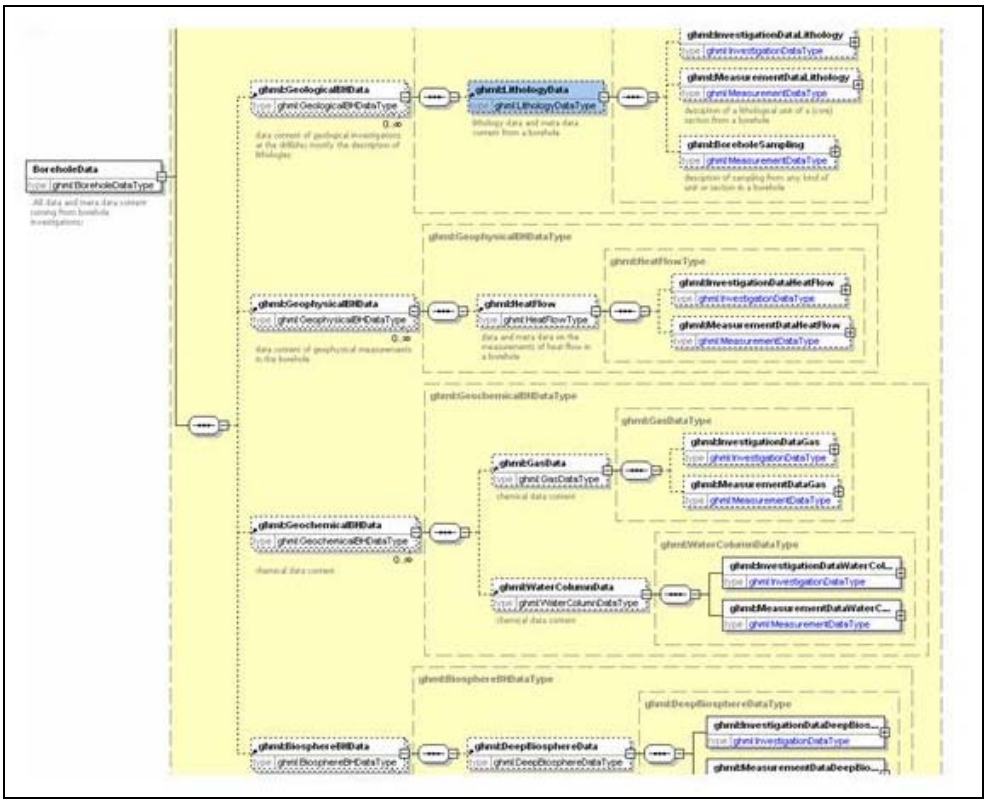

Figure 5. The entire expansion of the "BoreholeData" element with the "InvestigationDataType" and the "MeasurementDataType" for each activity

Regarding the metadata description element (InvestigationData), an optional attribute for the identification number and several elements of simple types are included, all optional (see Figure 6). These elements give the information about the data set as well as any copyright item. By extension, the specific fields of the different activities are included (see Figure 7).

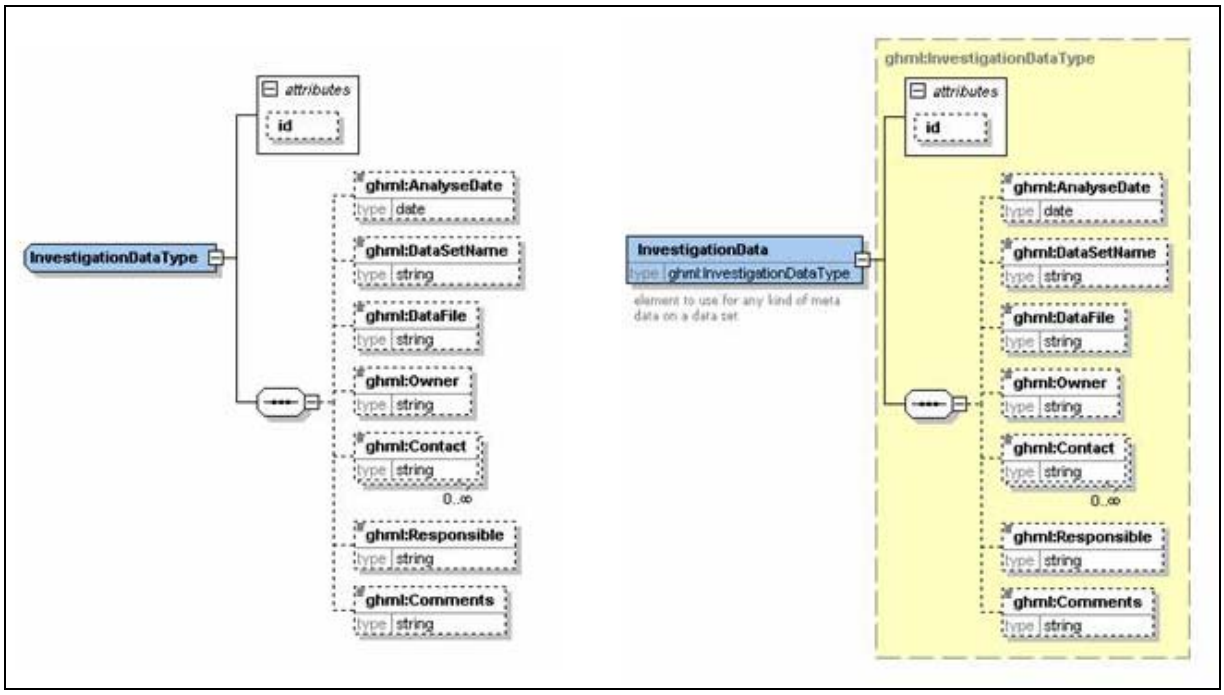

Figure 6. The global "InvestigationDataType" (left) and element (right).

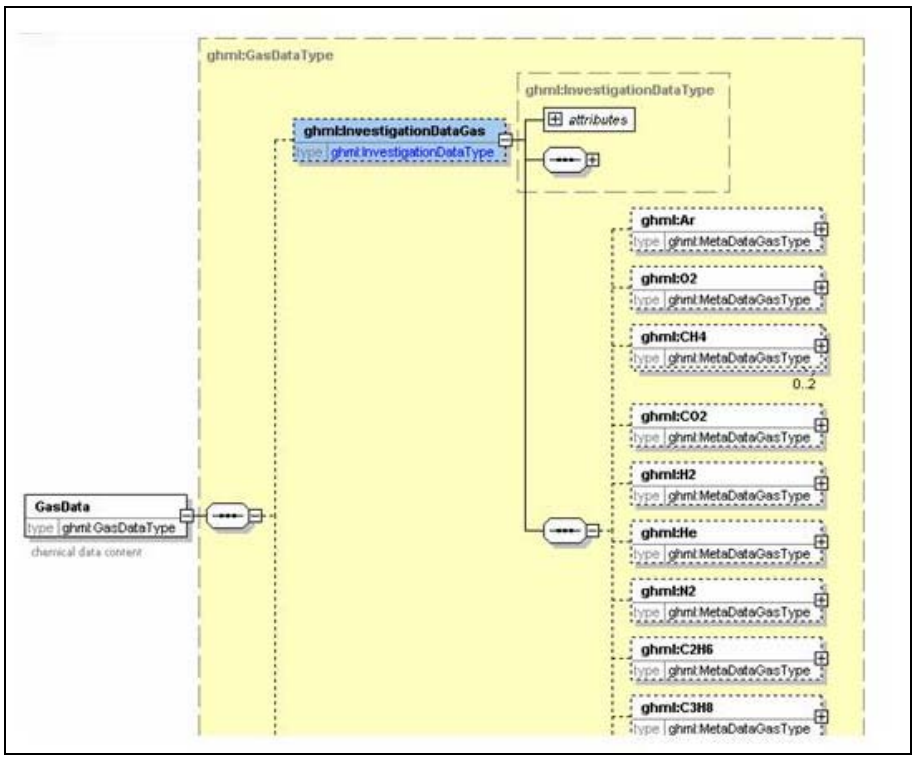

Figure 7. An example of the extension of the "InvestigationDataType": analysis of direct gas coming from the borehole while drilling

The "real data" could be described and stored in the XML instance document, which is built according to the "MeasurementDataType" (see Figure 8). This type consists of two optional attributes: an identification number and a persistent digital object identifier DOI. A DOI can be assigned to each data set. The DOI attribute enables publication of data and offers authors an incentive to publish data through long-term repositories (Klump et al., 2006).

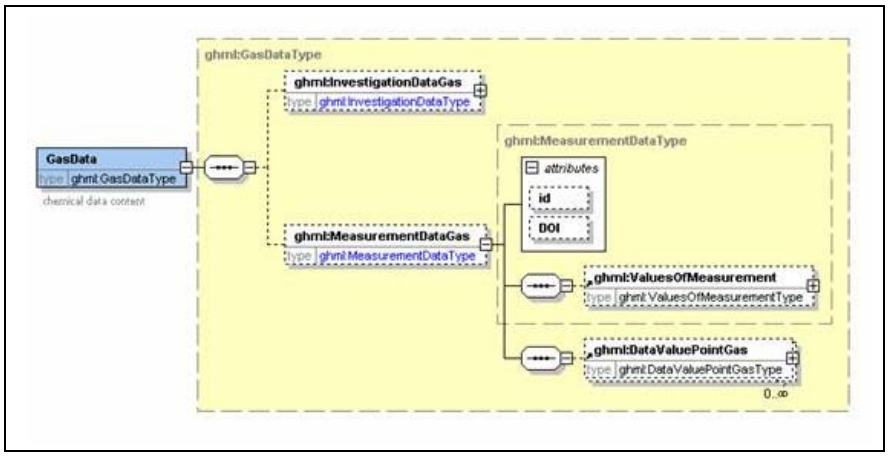

Figure 8. the "MeasurementDataType" extended for analysis of gas in a borehole

Additionally, this type includes an optional element, which can describe data in all existing formats: the "ValuesOfMasurement" element, accorded to GML (Cox et al., 2004). It includes solutions for the storage of data tables, lists and any kind of data files (see Figure 9). A detailed description is attached in the field data model.

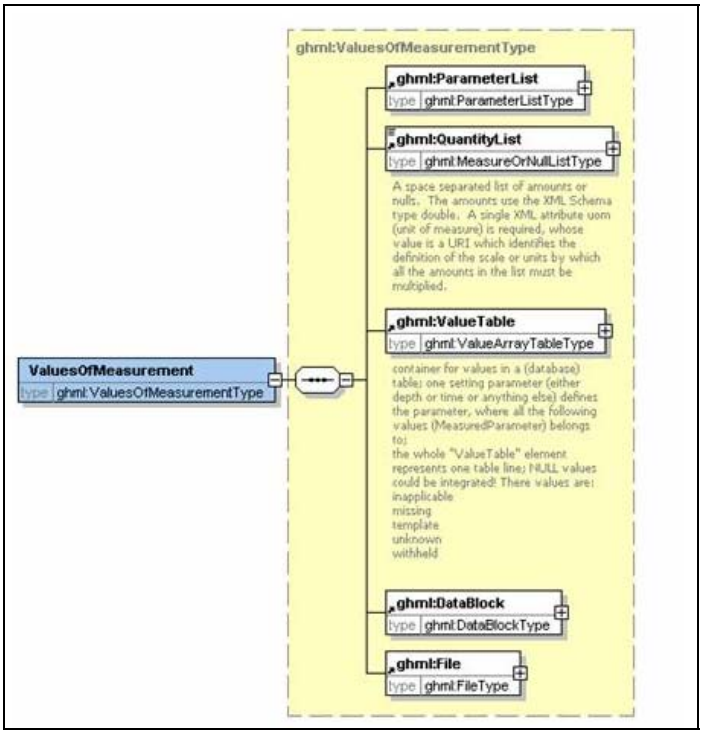

Figure 9. The global "ValuesOfMeasurement" element

If another description is desired, the extension of this type can be realized by the inclusion of "Data Value Points" (see Figures 8 and 10). The global "DataValuePointType" consists of one mandatory setting parameter (either time or depth) and a number of unbounded related measured parameters. The name of the parameter element describes the name of the parameter (e.g., Ar for Argon) and the attribute the unit of measurement (uom).

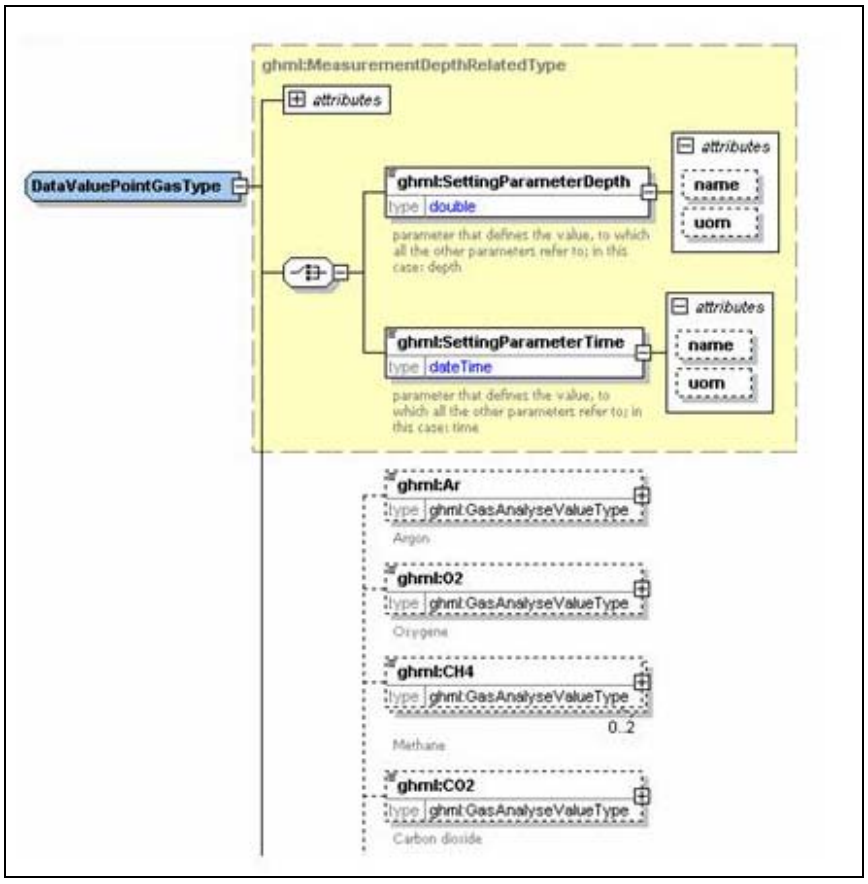

Figure 10. The "DataValuePointType" used and extended for gas analysis

In general, it is recommended to use the global "MeasureType" for any kind of data values. This type has a simple content (double), and the name and unit of the parameter are attributes. The setting parameter element, as with most of other data value elements, is built with this type. Restriction of possible data input is set by enumeration lists where it is required, e.g. a list of about 50 lithology therms for geological description.

4 A FIRST IMPLEMENTATION

An XML schema language is a formalization of the constraints, expressed as rules, or a model of structure that applies to a class of XML documents (van der Vlist, 2002). Therefore, an XML schema is the definition of the data model. Based on this data model, XML files will be created, which are thus instance documents of the XML schema, the concrete data objects. Therefore, these XML files use the XML schema as building instructions (see Figure 11) and contain the scientific data itself.

On the one hand, the XML schema, in our case the GHML and the field data portion, is used for the validation of the XML instance documents. On the other hand, these concrete objects can also be created based on the XML schema itself. In the present case, the XML file contains both, data and Meta data of any Gas Hydrate investigation.

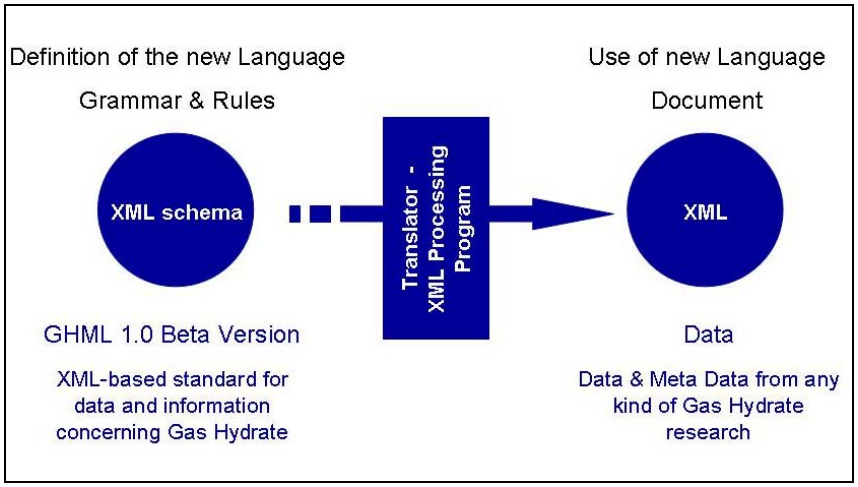

Figure 11. Relation between the XML schema definition and the XML instance document

Therefore, the resulting XML files are standardized in a uniform format enabling data exchange, data storage, data visualization, and data mining. A scientist is generally not obliged to handle the XML file itself. Tools and interfaces, which permit all desired data transactions, perform this instead. These functionalities could be provided by services accessible via an Internet based portal (see Figure 12).

Distributed heterogeneous national or international databases are connected by specific adapters to the data infrastructure. These adapters create the XML files and translate the standardized files into proprietary formats used by the local data provider and vice versa. Therefore, the data providers themselves do not need to make any changes to their internal data structure or model.

One first implementation of an adapter was realized at the GeoForschungsZentrum Potsdam and can be used via Internet (www-app1.gfz-potsdam.de/ghml/). It is the demonstration of a concrete use case. Instead of a database, a data file originated from a sensor in a borehole registering mass values of different gas types at different depths is connected to the portal and the provided services.

Therefore, an XML file based on the central new virtual data model has to be created. By the help of the internet based interface (see Figure 13), the user is led through a three step process, permitting the adapter to create the XML file and to store the data from the data file into the exchange format. The resulting data charged XML file can be saved to the hard disk of the user with the XML extension (.xml).

The boundary for a first implementation is reached here. Neither the portal function nor service is yet realized. This leads to future work on the portal itself.

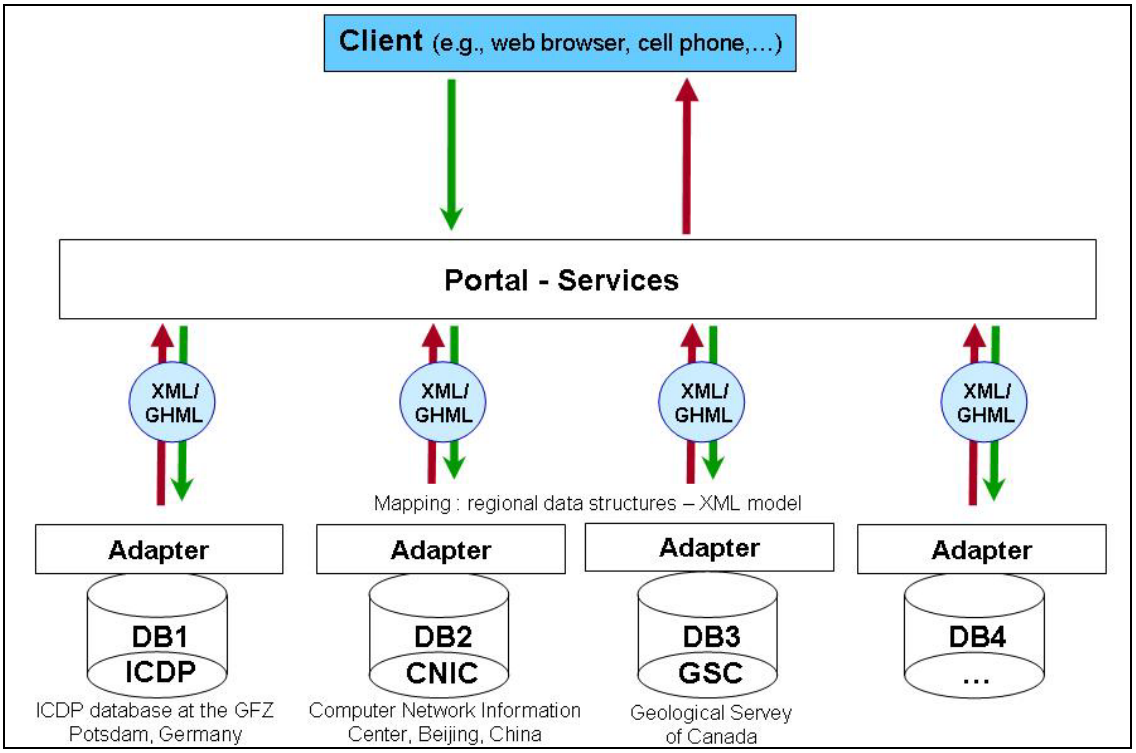

Figure 12. Structure of a future portal with the different components

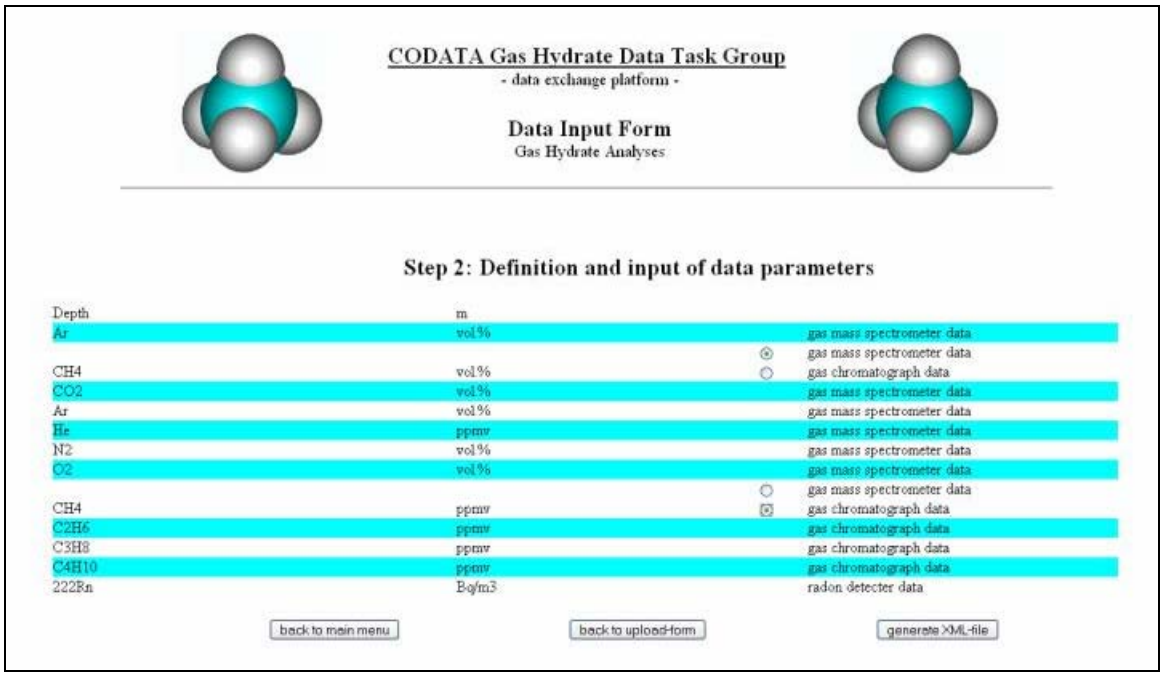

Figure 13. Internet based user interface for the creation of XML files based on the field data portion of GHML: Step two of the three step process

5 CONCLUSION

The present work leadS to a first step for a portal or a virtual data and information infrastructure for all aspects of gas hydrate research. The XML schema model enables the creation of XML instance documents, which can handle all manner of documents, databases, data files and formats. It is a quick solution, easily readable and usable.

Hence, in the future it is important that efforts are focused on the unification of the different components of GHML. Functionalities from XML (xPath, relations,...) should be added as well as interfaces to common standards.

This effort could stand as a pilot project and could also be a model for other fields of scientific research investigations.

6 REFERENCES

Cox S., Daisey P., Lake R., Portele C., & Whiteside A. (2004) *Implementation Specification - ISO/TC 211/WG 4/PT 19136 Geographic information – Geography Markup Language (GML),* version 3.1.0, OpenGis© Recommendation Paper.

Klump J., Bertelmann R., Brase J., Diepenbroek M., Grobe H., Höck H., Lautenschlager M., Schindler U., Sens I., & Wächter J. (2006) Data Publication in the open access Initiative. *Data Science Journal, Volume 5,* pages 79-83.

Löwner R. & Conze R. (2005) Mallik Data and Information System – development of a scientific data exchange platform in Scientific Results from the Mallik 2002 Gas Hydrate Production Research Well Program, Mackenzie Delta, Northwest Territories, Canada, (ed.) S.R. Dallimore and T.S. Collett; Geological Survey of Canada, Bulletin 585, p. 9.

Smith, T., Ripmeester, J., Sloan, D., & Uchida, T. (2007) Gas Hydrate Markup Language as it pertains to laboratory data. *Data Science Journal*, 6, GH18-GH24.

van der Vlist E**.** (2002) *XML schema*. O'Reilly & Associates, Inc., First Addition, ISBN: 0-596-00252-1

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.