

DESIGN AND FIRST STEPS IN THE DEVELOPMENT OF OPALIS

A. Bergmann, H. Gärtner, M. Breunig, A.B. Cremers, R. Dikau
Institute of Computer Science III and Department of Geography
University of Bonn

ABSTRACT

Integrated access to the many heterogeneous and distributed sources of Geo data is an important precondition in order to facilitate the investigation of global processes like temporal development of ecosystems. The object oriented paleoecological information system OPALIS aims to support interoperability of various data sources. While retaining the advantages of local data maintenance and formats, it will provide uniform access to information by means of integrated queries and object oriented database views. In this paper, we present the component based architecture of OPALIS being designed to support extensibility and adaptability to future requirements. Furthermore, we describe the derivation of the relevant Geo domain knowledge used to guide the integration process as well as characterize the query types that will be supported in OPALIS.

1 INTRODUCTION

Ongoing efforts and high investments have been made in recent years to develop models and techniques for the acquisition, maintenance and analysis of Geo data. The increasing use and development of modeling approaches in Geo sciences has caused a demand for, and the production of a vast amount of data. Additionally, many new case studies produce a considerable variety of field data. A great diversity of methods such as hydrological modeling, analyses of sediment covers, geomorphological mapping or dating studies provide heterogeneous data that are stored in diverse formats and in different databases [VINK92; DIKA92]. Furthermore, there are still no satisfactory solutions for handling 3D/4D models within standard two dimensional GIS. This scenario, common to many application fields [BBCR95], obviates the need for open and extensible information systems supporting the integration and administration of heterogeneous Geo data.

In 1996 a combined project named IOGIS (Interoperable Geoscientific Information Systems), consisting of six German research groups in the fields of Geography, Geology, Climatology, Computer Sciences and Remote Sensing was founded in Germany. This project, which is supported by the German Research Foundation, deals with the integration of data and methods for spatial and temporal modeling of geoscientific problems [IOGIS98].

The topics of IOGIS concentrate on object-oriented modeling, 3D/4D extensions, standardisation and structuring of geoscientific methods in GIS and plausibility of Geo objects.

As a part of IOGIS the OPALIS¹ project aims to develop object oriented concepts and mechanisms as well as an open GIS architecture facilitating the integrated use of isolated data pools to support Geo scientists in characterizing the temporal development of ecosystems and comparing the importance of different processes leading to actual conditions.

The remainder of the paper is organized as follows. In the second chapter we propose an open system architecture for OPALIS and give a brief description of the main

components that will support interoperability in the heterogeneous environment. We will then focus on the main building blocks of OPALIS providing integrated data access. The third chapter describes the process of deriving the needed Geo scientific domain knowledge from data formats and standards utilized by the different Geo groups to deal with semantic heterogeneity. We present a sample integrated model that was developed for an application to investigate drilling data from different sources. In the fourth chapter we classify "integrated queries" that should be answered by OPALIS. Finally we give an outlook over the next steps of OPALIS.

2 THE OPALIS SYSTEM DESIGN

The OPALIS system aims at facilitating integrated queries that are posed against diverse distributed data sources. Each of the source systems manages certain possibly overlapping portions of the data needed to answer the overall query according to its local needs. Furthermore, OPALIS should provide a platform for rapid development of information technology [BBBC+98] for a variety of Geo scientific applications that need integrated access to a broad base of information sources and available Geo services. In the following we discuss the design issues arising from this scenario and propose an appropriate system architecture.

2.1 Design issues

OPALIS will be used in an environment that is characterized by *different levels of heterogeneity*. Geo scientists use diverse computer based tools and systems for maintenance and analysis of their data depending on the special needs of the different disciplines and projects. Furthermore, highly developed and sophisticated tools are often only available on one or at most some of the existing computer platforms. In order to serve as a common access interface to integrated Geo information, OPALIS has to deal with heterogeneity on a technical level concerning different computer hardware, networks and operating systems.

In addition, OPALIS has to deal with two aspects of heterogeneity on a logical level. On one hand, the single source systems store data using different formats fitting the local needs. This leads to incompatible data schemata in many ways. Information of the same type is modeled

¹ The OPALIS project (Open Paleoecological Information System) is funded by the German Research Foundation (DFG)

using different names of attributes, values are stored in different units and using different standards, integrity constraints applicable to one data set do not hold for another, etc.. OPALIS has to incorporate mechanisms and the necessary Geo domain knowledge that are able to solve these kinds of semantic conflicts.

On the other hand, to be able to interoperate with the existing tools mentioned above, OPALIS has to provide access to integrated information by means of adapted interfaces. That is, OPALIS needs a mechanism supporting different views of an integrated model. The ability to define adapted views of integrated data relieves us from the task to build the non realizable "universal model" in Geo sciences.

Furthermore, this technique supports another important design issue - *extensibility*. The internal representation of integrated information and the integration process is decoupled from the access interface of the data consumers. Thus OPALIS can be easily extended and adapted to upcoming requirements like compatibility to OpenGIS standard [OGI96]. Extensibility is also required with respect to Geo services and methods provided by third party vendors. OPALIS must be able to easily make use of new functionality on the server side to extend the integration mechanism as well as to enhance client side features.

In order to preserve advantages like local maintenance and processing of data related to local projects, OPALIS should not encompass a centralized data store. Thinking of the rapidly growing amount of data gathered in the Geo domain, it is preferable to provide an integration infrastructure connecting data sources with consumers. So OPALIS should have a *distributed system architecture* supporting "thin" and platform independent clients at the data source and data consumer sites communicating over a data integration server. Nevertheless, in order to enhance performance of access to integrated data and to ensure the availability of important information, it may be necessary to store certain parts of the data in a central repository.

Summarizing our discussion, the OPALIS system architecture should support the following features:

- heterogeneity on different technical and semantic levels
- open and extensible architecture
- distributed Client/Server architecture
- Geo domain knowledge and Geo data repository

2.2 Open system architecture

This chapter surveys the main "building blocks" of the OPALIS system shown in figure 1. We propose a design based on object oriented architectures supporting distributed access to heterogeneous software systems.

The main information infrastructure of OPALIS is a CORBA® [OMG97] compliant *object request broker*. CORBA enables transparent access to methods and data of objects provided in a computer network. This will not only lead to full location transparency for communication within the system, but it will also bridge the gaps between possibly different computer platforms, operating systems, and programming languages used to build the various components of OPALIS.

The OPALIS server encompasses three main components. The first two components are the "Integrator" and the "Geo-Warehouse" implementing respectively a mechanism to decompose queries appropriate to the involved data sources and a repository for the "integration

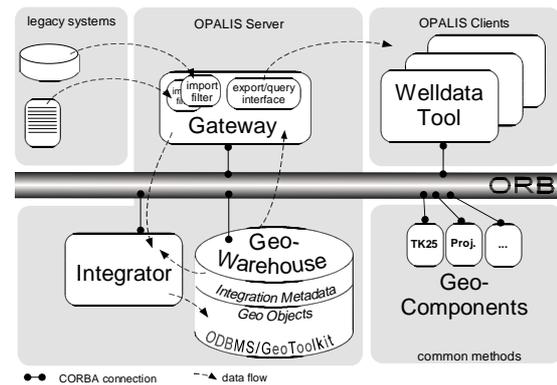


figure 1: The OPALIS system architecture

meta data" used to control integration tasks. Examples for this kind of meta data are Geo domain knowledge or the specification of integrated symbol keys for drill data. Together these two components perform the actual integration process which is described in more detail in section 2.3 UNTER.

The storage component of the OPALIS server is built upon the object oriented database system ObjectStore® [ODI97] and GeoToolkit [BBC97] - a class library for the management of spatial objects within an object oriented database. In this way OPALIS adopts the essential features of database systems like efficient maintenance of huge amounts of data, concurrency control, and retrievability of data.

The third main component, called "Gateway", is responsible for transporting all data streams out of and into the server. Here we have to deal with many different kinds of legacy storage systems ranging from flat files at most associated with a textual format description to sophisticated database systems providing full access to schema information. We need a mechanism that supports easy development of import filters connecting the manifold storage systems and formats to the integration framework. The filters perform syntactic checks and convert source data to prepare it for the incorporation into the system. Furthermore, a query processor is located in the "Gateway" facilitating the access to integrated information from OPALIS client applications.

An additional building block of OPALIS is the extensible collection of Geo methods like cartographic projection routines or the information server for the TK25 grid of Germany. These are implemented as CORBA services, thus being available to the OPALIS server as well as to client applications.

2.3 The Integration process

While we can rely on an established interoperability standard like CORBA to overcome heterogeneity on the technical level, there remains much work to be done to enable interoperability on the data level. In the following we describe the integration process designed for OPALIS in more detail.

The actual integration process is performed in several steps, which correspond to the different components of the OPALIS server (see figure 2). At first, each data source needs a connection component, called "Wrapper". These are on one hand composed of a generic part, which hides the particularities of the different types of

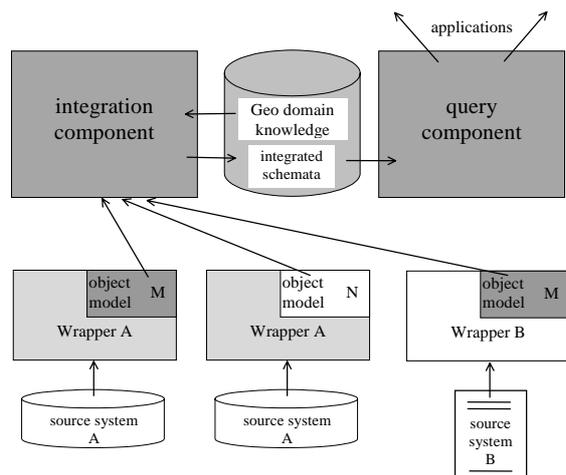


figure 2: data flow in the integration process

storage systems and provides a uniform interface for data import. On the other hand a wrapper includes an object model depending on the respective data schema implementing a mapping from the source data model into an adequate object oriented model. This technique reduces the costs for establishing connections from OPALIS to new data sources. The generic part merely depends on the storage system for the source data e.g. ASCII file or relational database. Hence it must be implemented only once per source. The schema dependent part is merely used to perform simple format testing in order to ensure a structurally correct instantiation of the respective object model. A meta object protocol and mechanisms needed to ensure certain integrity constraints are built into OPALIS providing a framework that supports the rapid realization of adequate object models.

The next step deals with semantic integration. On this level integrated schemata are defined reflecting the relationships between the object models of the data sources. The integration process is controlled, exploiting the integration meta data found in the Geo domain knowledge database of OPALIS. This knowledge includes diverse standards and their interrelationships, inter model integrity constraints, attribute name mappings, conversion function for measure units or coordinates, etc..

The third integration step is implemented in the query component of the OPALIS server. OPALIS supports the specification of more than one schema by means of an object oriented view mechanism [CBR94]. The query component has to provide client applications with a uniform interface for accessing integrated data by mapping queries on the appropriate integrated schema.

The process for semantic integration of heterogeneous data as described above does not determine the location where the integrated information will effectively reside. During the design of OPALIS we have looked into two strategies that can be associated with federated databases [CON97] and data warehouse systems [LZW+97], respectively. Federated systems rely on data storage and maintenance at the source sites. Queries are decomposed and forwarded to the data sources. The answers are collected and integrated in real time. In this way information actuality can be ensured at the cost of generally higher response times and possibly incomplete answers due to temporarily unavailable data sources. On

the other hand, data warehouses integrate and store necessary data in a central repository in advance. Therefore precomputed results are readily available at query time. This allows for relatively short response times and information derived from all source data sets. The disadvantages of this approach are storage of possibly outdated information and less flexibility. Information that was not modeled at integration time cannot be handled by the system.

As a result in OPALIS we are using a mixed strategy. Because of the huge amount of data and copyrights in the Geo sciences, most of the data cannot be stored in a central place. But there exists a variety of meta information, which can be computed in OPALIS, like "How many drillings stored at the connected source sites are located in a certain area?" or "Determine where data or publications about a certain project can be found?". In these cases a pure distributed approach would lead to intolerably long response times. Thus an important question in order to adequately model the Geo domain is, what information should be integrated in advance and stored centrally?

3 MODELING DOMAIN KNOWLEDGE

A precondition for data modeling is knowledge of the domain the data are related to. In case of OPALIS the domain is Geo Science. Modeling domain knowledge as the central part of data modeling in OPALIS involves transferring specific knowledge of a subject into a data model. This knowledge is concentrated in and expressed by specific standards that most of the datasets are based on. Taking these assumptions into account, OPALIS started with the sampling and documentation of existing data and data handling methods of different German research groups within the IGBP-PAGES programme. The research topic of these groups is the development of sediment covers and the reconstruction of palaeo surfaces in different environments.

Questionnaires were directed to different group members to get relevant information about both the aims of their research projects and meta data corresponding to the datasets that were made available to us. The datasets are founded basically on borehole data. Therefore it was an important aspect for the development of OPALIS to start data modeling based on one- and two-dimensional data.

3.1 Conceptual Data model

After documenting these data including all associated meta information (this means "Core Meta Data" and "Domain Meta Data" with respect to the definition of the Open GIS Consortium – Open GIS Project Number 97-110R1 -) we started to construct a conceptual meta data model.

It was necessary to structure the datasets on different levels of information. This meant predefining categories of meta data and data which will represent the basic entities of the conceptual model. These entities are derived from the internal structure of the datasets. They represent the basic categories for creating the objects of the object-oriented model.

For getting a precise overview it is essential first to represent all information about the aim of the project, associated institutions, persons dealing with data and the equipment used in the project. Furthermore we had to identify the different types of data (1 to 4-dimensional), the kind of data (e.g. stratigraphic information derived from

boreholes or sediment profiles), the methods used for data capture, processing and storage as well as interpretations of analyzed data.

The main resulting categories (objects) are as follows:

- Project / Aim of project
- Type of data
- Methods
- Archives
- Primary data
- Tools
- Interpretations / References

The most important information with respect to data quality apart from the hitherto cited examples relates to standards used for data capture in the field. These standards comprise, other than specific instructions for data capture, encoding descriptions for handling field data. These vary whether the field work (for example drillings) is carried out by Geologists or Soil Scientists with respect to their specific research interests. The importance of isolating these standards is obvious if you take into account that the description of stratigraphic information of a borehole will differ whether it is analyzed by geologists or soil scientists.

All datasets captured in the field are usually based on these standards (e.g. "Symbolschlüssel Geologie [SYM91]; Bodenkundliche Kartieranleitung [KA496]; etc.) and the primary data is stored and often processed as encoded information.

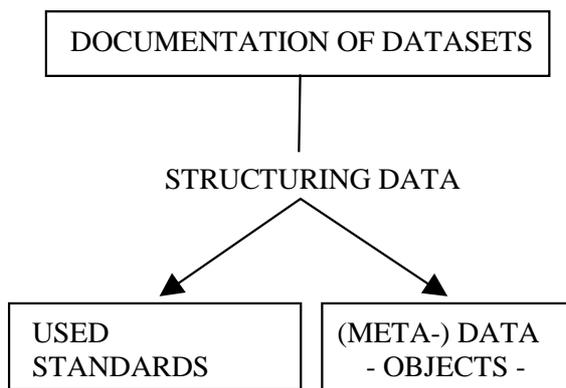


figure 3: Steps in creating the conceptual model

The diagram shown in Figure 3 reflects the central task of building a conceptual model in preparation for the object-oriented modeling of geo scientific data. Besides predefining objects that represent the dataset itself, it is of great importance to identify standards used.

3.2 Logic Data model

By structuring the datasets the preconditions for generating a detailed data model using object-oriented modeling techniques are met. The representation of semantics (- in a wider sense the theory of the inherent meaning of a model or a Language -) within a data model is a measure of the accuracy of the model that has to be achieved.

Within Geosciences this way of modeling may lead to a closer approximation to real world conditions because of the consideration of the semantics of geo objects.

Before modeling the datasets and the standards used a common modeling language is necessary because the OO-models are created by Geomorphologists and should be transferred into a physical model by Computer Scientists. In spite of very good personal communication many projects profit from the use of a graphic modeling language that is able to include and express semantics of geo data without further explanation. So the syntax and the semantics of the language should be easy to learn and well known to all participants. In addition semantic modeling enables the integration of spatial and temporal data independent of implementing details like single point representations. For this reason the integration of semantics (e.g. the interaction between landslides, precipitation and soil conditions) is the base for modeling complex processes in Geosciences.

Within the OPALIS project the Unified Modeling Language (UML) [UML97] is used as a base for the design of geobjects and their interacting relationships, to develop a concept for an open palaeoecological information system and to realise a prototype of an open object-oriented database.

3.2.1 Modeling datasets

While the conceptual data model is independent of any kind of database system, the task of object-oriented modeling is to create an object model that is able to be directly transferred into a physical model for a special database system. Object-based models are founded on the concept of objects that own properties, behaviours and relationships with other objects in space and time. For this reason it is necessary to revise the objects of the conceptual model with respect to their semantics.

Each object as an independent item and central entity of a data model has a unique identity that may differ from its value. So objects can be combined to create new objects.

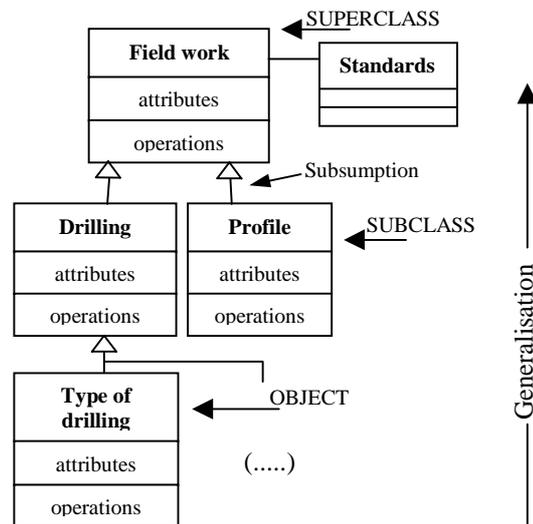


figure 4: .Generalisation of Geo data using UML

An object encapsulates structure (attributes) and behaviour (methods/operations). Objects with common attributes can be grouped to classes. Based on the principal of generalisation classes can be defined as special types of other classes. That is why they are called subclasses. These subclasses are subsumptions with

respect to structure and behaviour of the so called superclasses (Figure 4).

Apart from the illustration of generalisation of objects Figure 4 contains some of the redefined objects of the datasets used in the IGBP-PAGES programme. With respect to the complexity of the resulting data model, we are not going to describe the detailed object-oriented model in this paper. We will rather describe the fundamental assumptions that led to the final model. It was taken into account that the base of all datasets to be modelled is field work associated with (with no further specification in this model) the standard used for data capture. To specify information that is of general interest for all related classes and objects we derived new objects which are directly related to field work in general. These objects are "Site data" - containing all site information like coordinates, map information and all relevant data about the research area - and "Project data" containing all relevant information about the project itself. The attributes of these objects are clearly defined regarding the underlying dataset. Drillings will lead to borehole data basically expressed as stratigraphic information representing by different layers. The layers themselves contain further information that is related to each layer. This further information is expressed as objects/classes like "Components" (related to standardised archives, e.g. Pollen data base) "Chemical analyses", "Physical analyses" or "Datings" (as generalisation of different types of Datings, e.g. Dendrochronology, ^{14}C).

A lot of these objects/classes represent or contain discrete or continuous temporal data, that has to be integrated with the model. In the first stage of the project we handle temporal data as common objects with the attribute "time". The problem of 3D/4D extension by implementing separate spatio-temporal data models will be addressed after realising the prototype of OPALIS which is based on one and two dimensional data in three dimensional space.

3.2.2 Modeling standards

Treating a modeling standard like "Symbolschlüssel Geologie" as a kind of dictionary does not support efficient data integration. The standards contain various rules for different types of field work. These rules and the various shortcuts are associated on different levels with the objects (especially to their attributes) of the dataset.

To guarantee full data integration and to compare different datasets, these rules must be transferred to a data model that follows the same guidelines as the model of the related datasets.

Naturally the basic objects of "Symbolschlüssel Geologie" are defined according to the specifications of this standard. The resulting objects/classes are primarily Project/Site data, Stratum data and Sample data. The layer related to stratum data subsumes the subclasses "Depth", "Stratigraphy", "Petrography", "Genesis", "Colour", "Additional data" and "Special data", whereas the subsumptions of "Special data" are directly associated to **G**, **C**, **A** and one of the subsumed objects of **P**. As a result we gain a data model that is as complex as the models of the datasets described before. Using the UML we created OO-models for different standards like "Symbolschlüssel Geologie", "Bodenkundliche Kartieranleitung" and several mapping instructions. To prepare data integration we started to build up links between the models by defining special types of queries.

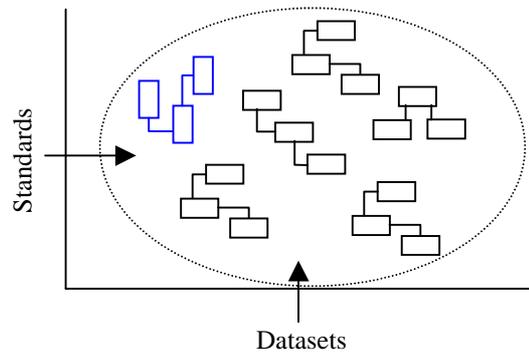


figure 5: Merging data sources for data integration

4 INTEGRATED QUERIES

Developing a query component for accessing integrated data (2.3 - third step of integration) is based on defining possible queries that may be made on the different data models. It is not realistic to anticipate all conceivable queries that may be made of datasets.

To build up this component adequately it is necessary to define possible types of queries. By analyzing the questionnaires and discussing the aims of different projects we defined basic types of queries:

- General queries on existence of specified data
- Queries on individual values of attributes
- Queries on ranges of an attribute
- Queries on multiple values
- Queries on spatial information, i.e. queries on geometry (1-4 dimensional) or topology of objects
- Queries on temporal information

These types are prerequisite for the design of the query component. Integrated queries on objects are composed from these types so they have to be implemented before programming special queries that are focused on different values.

The following example shows different possibilities of how an integrated query can be executed in OPALIS (see figure 6). Imagine the object model of one data source specifies an attribute named "numberTK25" and the object model of a second source does not. One way to perform an integrated query like "get all drillings located in the TK25 with number 2716" on both data sets is to use the attribute mentioned above. Because the second source cannot be queried this way, this would lead to a result only covering drilling data from the first source.

Alternatively, a spatial access query could be posed, using the extent of the TK 2716 as the selection predicate. In this case the TK25-Service provides the necessary coordinates and a projection service is responsible for comparable coordinate values. If we have enough domain knowledge, the query could be automatically transformed according to the object model of each data source. The different results computed this way can be integrated exploiting the mappings from the source models to an integrated schema. In OPALIS all integrated schemata are specified as views over the object oriented models of source data. The specification is stored in the meta data repository of the OPALIS server.

A third possibility is the data integration prior to querying. The second source model could be extended with the missing attribute "numberTK25". The respective value could also be derived using the TK25-Service. Here the models representing each data source may become very

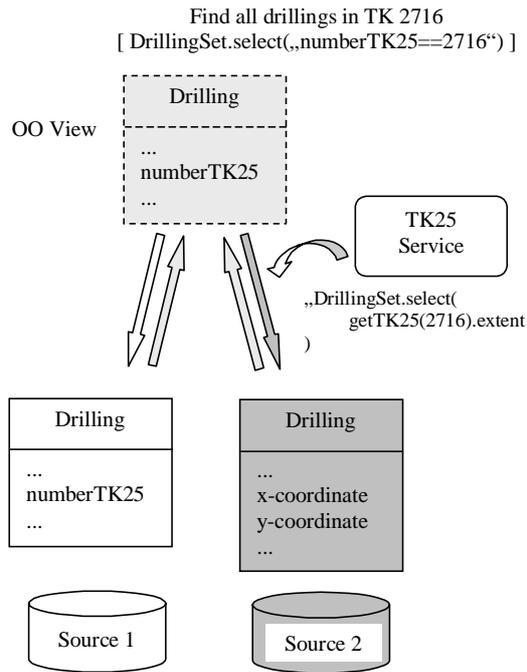


figure 6: A sample integrated query

large and complex. But the results of queries involving such an attribute could be computed more efficiently, due to partly precomputed predicates.

The specification of integrated queries has been the first step towards data integration in the OPALIS project. The development of the query component and its continuous evaluation are the next steps. This will involve a detailed analysis of the integration alternatives described above.

5 CONCLUSIONS AND FUTURE WORK

The aim of OPALIS to realize the integration of heterogeneous Geo data while using distributed data sources and to guarantee the integrity and extensibility of the system has been presented. By describing the design issues and the open system architecture (figure 1) related to the importance of accurate object-oriented modeling we want to focus on the need for a system that enables data integration without establishing a central database. The cooperation of Geo Scientists (OO Data Modeling) and Computer Scientists (System Architecture) is the most efficient way to handle the different aspects of data integration and the interoperability of the system. Interoperability will be realized in two aspects. On a technical level the integration of diverse GIS is supported. Moreover on the semantic level integration is supported by modeling domain knowledge and implementing different types of queries. Building up this system accurately and realizing the first prototype based on one and two dimensional data in three dimensional space is the foundation of future work. More complex datasets that are based on three and four dimensional data will be transferred to semantically based object models to realize the 3D/4D extension of the prototype by integrating these data to the system.

ACKNOWLEDGEMENT

The authors would like to acknowledge the cooperation with the IGBP pilot groups of W. Andres and J. Wunderlich (University Frankfurt/Marburg), H. Streif and Ch. Hoselmann (NLfB Hannover) and G. Wagner and A. Lang (MPI and University of Heidelberg). We are also indebted to M. Assmann, J. Brinkmann and U. Radetzki for their implementation and modeling support.

REFERENCES

- [BBC97] Balovnev, O., Breunig, M., Cremers, A.B. (1997): From GeoStore to GeoToolkit: The Second Step. In: Proceedings of 5th Intern. Symposium on Spatial Databases, LNCS Vol. 1262, Springer, Berlin et al., pp. 223-237.
- [BBBC+98] Balovnev, O., Bergmann, A., Breunig, M., Cremers, A.B., Shumilov, S. (1998): A CORBA-based approach to data and systems integration for 3D geoscientific applications, submitted for 8th Intern. Symposium on Spatial Data Handling, Vancouver, July 13-15, Canada.
- [BBCR95] Bergmann, A., Bode, T., Cremers, A.B. Reddig, W. (1995): Integrating civil engineering applications with object-oriented database management systems. In: Proceedings of the 6th Intern. Conference on Computing in Civil and Building Engineering, Berlin, 12-15 July, Balkema Rotterdam/Brookfield, pp. 67-74.
- [CBR94] Cremers, A. B., Balownew, O. T., Reddig, W. (1994): Views in object oriented databases, International Workshop on Advances in Databases and Information Systems (ADBIS '94), Moscow.
- [CON97] Conrad, St. (1997): Föderierte Datenbanksysteme – Konzepte der Datenintegration, Springer, 331 p.
- [DIKA92] Dikau, R. (1992): Aspects of constructing a digital geomorphological base map. Geol. Jahrbuch, A122, pp. 357-370.
- [IOGIS98] IOGIS – Vision about a new generation of interoperable GIS. In German. Report of the participating groups at the Universities of Berlin, Bonn, Berlin/Bonn, Freiburg, Münster and Stuttgart.
- [KA496] Bundesanstalt für Geowissenschaften und Rohstoffe & Geologische Landesämter der Bundesrepublik Deutschland (Hrsg.) (1996): Bodenkundliche Kartieranleitung. 4. Auflage, Hannover, 392 p.
- [LZW+97] Labio, W.J., Zhuge, Y., Wiener, J.L., Gupta, H., Garcia-Molina, H., Widom, J. (1997): The WHIPS Prototype for Data Warehouse Creation and Maintenance, in: Proceedings of the ACM SIGMOD Conference, Tuscon, Arizona, pp.
- [ODI97] Object Design Inc. *ObjectStore C++ API User Guide*. ObjectStore C++ release 5.0 documentation.
- [OGI96] The OpenGIS™ Guide. Introduction to Interoperable Geoprocessing. Part I of the Open Geodata Interoperability Specification (OGIS). OGIS Project Technical Committee of the Open GIS Consortium Inc., Buehler K. and McKee L. (eds.),

OGIS TC Document 96-001, (1996)
<http://www.ogis.org/guide>.

[OMG97] Object Management Group (1997): *CORBA 2.0/IIOP Specification*. OMG formal document 97-09-01. <http://www.omg.org/corba/c2indx.htm>

[UML97] Rational Software Corporation: The Unified Modeling Language, Version 1.1, <http://www.rational.com>, September 1997.

[SYM91] PREUSS, H., VINKEN, R. & H.-H. VOSS (1991): *Symbolschlüssel Geologie*. Niedersächsisches Landesamt für Bodenforschung und Bundesanstalt für Geowissenschaften und Rohstoffe (Hrsg.), Hannover, 328 p.

[VINK92] Vinken, R. (Ed.) (1992): *From Geoscientific Map Series to Geo Information Systems*. Geol. Jahrb. A 122, Hannover, 501 p.