

# ONTOLOGIES: SOLVING SEMANTIC HETEROGENEITY IN A FEDERATED SPATIAL DATABASE SYSTEM

Villie Morocho, Fèlix Saltor  
Universitat Politècnica de Catalunya - LSI  
c/Jordi Girona 1-3. 08034 Barcelona, Spain  
{vmorocho,saltor}@lsi.upc.es

Lluís Pérez-Vidal  
Universitat Politècnica de Catalunya - IG  
Av.Diagonal 647,8.08028 Barcelona, Spain  
lpv@lsi.upc.es

Key words: Semantic Heterogeneity, Geographic Integration, Interoperability, Federated Database, XML, UML, GML

Abstract: Information integration has been an important area of research for many years, and the problem of integration of geographic data has recently emerged. This paper presents an approach based on the use of Ontologies for solving the problem of semantic heterogeneity in the process of the construction of a *Federated Schema* in the framework of geographic data. We make use of a standard technology (OMT-G based UML, XMI based XML, GML from OpenGIS)

## 1 INTRODUCTION

Interoperability and integration of heterogeneous data have been some of the goals to achieve during the last few years. From software corporations to world scientific institutions, researchers are working on it.

This paper presents a framework based on BLOOM (Abelló et al., 1999) which is based on Federated Database Architecture (Sheth and Larson, 1990). The BLOOM architecture especially adds security levels. In this paper, moreover, we change the scope from traditional Databases to spatial Databases. Inside of this Federated Architecture, at the level of schema integration, we make use of Ontologies for solving Semantic Heterogeneity.

Semantically rich information (i.e. metadata, context information) is added to *Native Schema* at the bottom level and this information will aid for assessing semantic similarity across ontologies in order to allow the construction of the *Federated Schema*. In this framework, after the *Geospatial Schema* level, (in which all models are native), these schemas are transformed into a Canonical Data Model. A Canonical Data Model (Castellanos et al., 1992) is a common model for all *Component Schemas*. We make use, in the first solution, the OMT-G (Borges et al., 2001) as CDM, and we take advantage of the features of this model. OMT-G provides some primitives used for modeling the geometry and topology of geographic data, providing support for “whole-part” topological structures, network structures, multiple

views of objects and spatial relationships. In a second solution, we make use the abstract model from OpenGIS (OpenGIS, 1999).

The other part of this paper, proposes to materialize the models from OMT-G or OpenGIS in XMI (OMG, 2002). The main purpose of XMI is to enable easy interchange of metadata between modeling tools (based on the OMG-UML(OMG, 2001)) and metadata repositories(OMG-MOF based), in a distributed and heterogeneous environment.

Once the model in XMI is materialized, we construct the Ontologies for the objects in the model. Then each object should have its own ontology and, afterwards, match the ontologies of the different schemas to integrate. In this matching process, it is possible to know whether there is a correspondence between Ontologies, and which object is semantically parallel to another. In this way, it is possible to achieve a *semiautomatic* schema integration. Continuing with levels of the framework, the Federated Schema should be authorized at the level *Authorized Schema*. They should also be filtered through the *External Schema* to finally obtain a *User Schema* at the top of the framework.

In this paper we first analyze the integration problem in the section 2 and related research in section 3. We present our architecture in the section 4 and then study the use of OMT-G,GML and XMI in it. Finally, we consider future work in the last section.

## 2 THE INTEGRATION PROBLEM

Numerous geographic information integration projects have become relatively important. Projects for the integration of geographic information have been developed in many countries and organizations, from NASA's Jet Propulsion Laboratory (<http://www.jpl.nasa.gov/srtm>) to integrated geographic data systems like Kyoto's Digital City (<http://www.digitalcity.gr.jp>). Many big software companies (ESRI, MapInfo, Autodesk and so on) also invest in research in this field. Database specialized software packages, have added characteristics to manage spatial data (i.e. Oracle Spatial, DB2 Spatial Extender, Informix Spatial DataBlade Module).

For these reasons, one of the main goals is to eliminate the problems in the integration and interoperability of geographic and spatial data. These problems have been studied by different authors. One of the biggest problem is the resolution of heterogeneity (Sheth and Larson, 1990) in federated databases, which most complicated part to resolve is *Semantic Heterogeneity* (Samos et al., 1999)). This problem is studied under new light with the integration of Geographic and Spatial Databases. Actually, Geographic Information Systems (GIS) are used at a very large scale. The user tries to get information from very different systems instantly, with only one snap show. The user would like to see some information about locations from city's databases and maps from GIS bank on a mobile device, and often on a very small-sized screen. Also, he may want to use GPS technology (Global Positioning System) to say whether this place is near or far from his actual position. He may also be willing to make a reservation or to buy something. With more complicated features, the user could revise home devices before some purchase. These applications are intensive in internal processing and quite transparent to the user.

One strategy to face the problem of interoperability and integration is the adoption of standards. If they are adopted, semantic heterogeneity will be less and the information loss in the integration process will be also reduced. Many associations, committees, consortiums, organizations, and programs, work on standardization of this kind of information XML (W3C, 2000), GML (OpenGIS, 2002), FDGC-UDK (Günter and Voisard, 1998) and so on. The OGC (OpenGIS Consortium <http://www.opengis.org>) is one of them. The OpenGIS includes a model for standardization of geographic data (OpenGIS, 1999), and also defines a language for it, the Geographic Markup Language GML (OpenGIS, 2002) (based on eXtensible Markup Language). In modeling languages actually the main is the Unified Metadata Language UML (OMG, 2001). In our framework, we will take advantage of this trend, and our design uses OMT-

G (Borges et al., 2001) (UML based) XMI (OMG, 2002)(XML based) and UMT (Oldevik, 2002) (UML-GML based), and is thereby standard based.

## 3 RELATED WORKS

Schema Integration refers to integration schemas into a single schema (eg., federated schema development by integrating schemas in a bottom-up FDBS development process). Many approaches and techniques for schema integration reported in the literature. Sheth and Larson in (Sheth and Larson, 1990) remark the unfeasibility of the complete automatic schema integration process. This is not possible because, among other reasons, (1) *the data models are unable to capture a real-world state completely*, (2) *it will be necessary to capture much more information than is typically captured in a schema*, and (3) *there can be multiple views and interpretations of real-world state; and the interpretations change with time*. One of the main problems is the comparison step (identifying naming conflicts, homonyms, or synonyms). From the view point of integration of spatial databases there is a light advantage to resolve this problem. Now there is a lot of information available for add as semantic information for schema integration. With the arrival of new technology and new possibilities of storage, it is possible to talk of near automatic solving of semantic heterogeneity in schema integration.

Around this approach there are a lot of approximations. Tools developed to perform schema integration are reported in (Hayes and Ram, 1990), (Sheth et al., 1988). Sheth (Sheth et al., 1988), for example, describe a forms-based interactive tool to integrate EER schemas with a way semiautomatic, guiding to the integrator through the process of integration. Recently the focus has changed from schematic to semantic integration of heterogeneous sources (Sheth, 1999). Many approaches for solving semantic problems are considering the use of ontologies as the best solution. Examples are: OBSERVER (Mena et al., 1998), applies ontologies to replace terms in user queries with suitable terms in target ontologies; SHOE (Heflin and Hendler, 2000) and Ontobroker (Benjamins and Fensel, 1998) are in the framework of semantic Web. They use ontologies to improve the searching abilities on the web. DB-MAIN (Thiran et al., 2000) is a CASE tool. It try to integrate all the aspects of the Federated Information System development, but in the process of solving semantic for the integration, is necessary much intervention of expert.

Research approach from view point of spatial and geographic sources are (Rodríguez, 2000; Fonseca, 2001; Fonseca et al., 2002). (Rodríguez, 2000) is

based on a subset of two ontologies, unlike OBSERVER, the solution does not create new ontologies, but creates links between similar entities in distinct ontologies. Fonseca in (Fonseca, 2001) directs the research from ontologies but not from database schemas. In our research, we profit the powerful of ontologies for solving semantic problem in the construction of the federated schema.

## 4 FEDERATED ARCHITECTURE WITH ONTOLOGIES

In this section we present our framework proposal for the use of Ontologies inside a Federated Architecture in order to resolve the semantic heterogeneity problem in the schema integration. In the Fig.1 we present the framework.

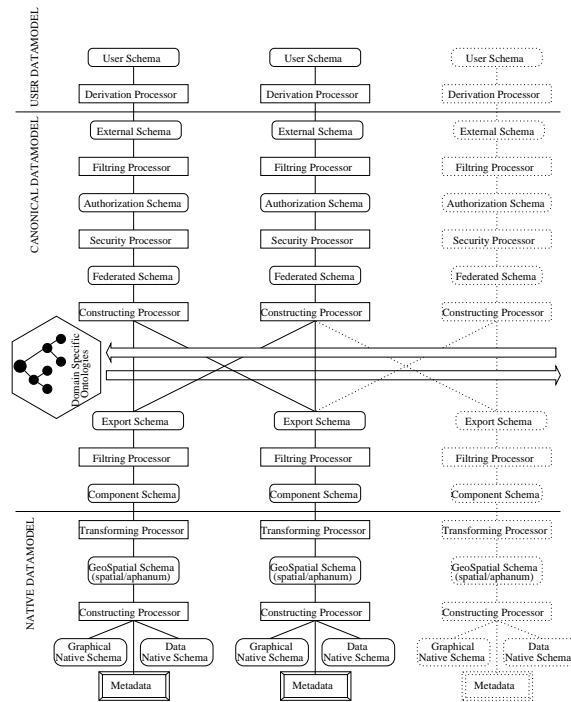


Figure 1: Federated Architecture with Ontologies

### 4.1 OMT-G as Canonical Data Model

Applications to model geography were guided by existing GIS internal structures, forcing the users to use the structures for interpretation of spatial phenomena according to whatever structures were available. Thereby, the modeling process did not offer mecha-

nisms that would allow for the representation of reality according to the user's mental model. OMT-G (Borges et al., 2001) (Object Modeling Technique for Geographic Applications) was initially based on the classic OMT (Rumbaugh et al., 1991) class diagram notation, and later adapted to match the concepts and notation of the Unified Modeling Language (UML (OMG, 2001)). OMT-G offers primitives that provide the means for modeling the geometry and topology of geographic data, making the modeling of geographic applications easier.

In this article we study the use of OMT-G for the representation of *Component Schema*. The framework takes advantage of the characteristics of OMT-G because it provides primitives to model the geometry and topology of geographic data, providing support for "whole-part" topological structures, network structures, multiple views of objects, and spatial relationships. Besides, the OMT-G model allows the specification of alphanumeric attributes and associated methods for each class. The main strong points of the OMT-G model are its graphic expressiveness and its representation capabilities, since textual annotations are replaced by the drawing of explicit relationships. For an extended explanation of OMT-G you can refer to (Borges et al., 2001). Between the big problems are the multi-representation and multi-resolution of spatial and geographic information (Spaccapietra et al., 2000). Those problems are covered in the model of OMT-G, by means of the definition of *generalization and specialization, aggregation and conceptual generalization*. A separation between levels of abstraction is necessary. For geographic applications, three levels of abstraction can be considered: the *conceptual representation level, presentation level and implementation level*. OMT-G primitives lead to three diagrams: class, transformation and presentation. We profit from this characteristics of OMT-G, then in the level of *Component Schema* we make use of OMT-G for representing the different components sources. Once we obtain the component schema, it is possible obtain the model in XMI representation.

### 4.2 XMI for Codifying the Model

The main purpose of XML Metadata Interchange XMI (OMG, 2002) is to enable easy interchange of metadata between modeling tools (based on the OMG-UML (OMG, 2001)) and metadata repositories (OMG-MOF based) in distributed heterogeneous environments. XMI allows metadata to be interchanged as streams of files with a standard XML-based format. The XMI specification supports the interchange of any kind of metadata that can be expressed using the MOF specification, including both model and metamodel information. The specification supports

an encoding of metadata consisting of both complete model and model fragment, as well as a tool-specific extension metadata. The XMI specification has two major components:

- The *XML DTD Production Rules* for producing XML Document Type Definitions (DTDs) for XMI encoded metadata.
- The *XML Document Production Rules* for encoding metadata into an XML compatible format. The production rules can be applied in reverse to decode XMI documents and reconstruct the metadata.

We propose a representation of the OMT-G model by means of XML, and obtain a materialized model capable of being parsed. From there it is possible to obtain the Ontologies for each object of the model. Once we get all the models to be represented in XMI, we construct the Ontologies for the classes and the attributes of the model obtained in XMI. Afterwards, we construct the trees Ontologies and it finally makes a matching of Ontologies and resolve, in certain degree, the semantic heterogeneity.

### 4.3 Solving Semantic Heterogeneity

At this level, Ontologies trees were obtained from all component schemas. It is also necessary to add to the trees all additional information from metadata, context and information correlation, considering that this information is able to semantically enrich the Ontology.

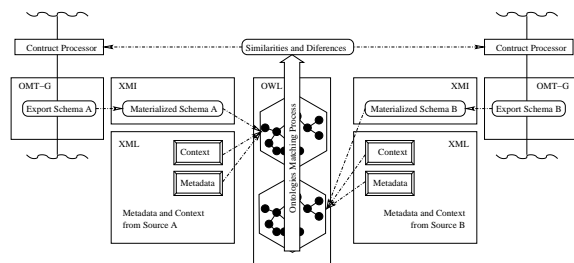


Figure 2: Creating Ontologies from schemas with OMT-G model

In the Fig. 2 we present how the Ontologies should be created from materialized schemas and how it gather information from metadata and context. Metadata and Context information should be expressed in XML. It is then possible to use the actual technology from W3C to express the Ontologies.

The OWL Web Ontology Language (W3C, 2002) is being designed by the W3C Web Ontology Working Group in order to provide a language that can be used for applications that need to understand the content of information instead of just understanding the

human-readable presentation of the content. OWL facilitates a greater machine readability of web content than XML, RDF, and RDF-S support by providing an additional vocabulary for term descriptions. Thereby, it is possible to create metadata, context information and Ontologies from schemas represented in XMI, because all of them are using XML technology.

Thereby, we propose the use of OWL to express the Ontologies and to use the matching technics (Kilpeläinen and Mannila, 1995) for searching similarities and differences between the objects that have to be integrated.

### 4.4 GML the most Standard

Another solution for our framework is the use of a model from the OpenGIS Consortium. It is based on the OGC Abstract Specification (<http://www.opengis.org/techno/specs.htm>). This might be a better solution than the OMT-G-based one, because it currently is the most solid specification, and it has many technological features additional. The main GIS corporations software are compliant with the specifications of GML (ESRI, MapINFO, ORACLE, Galdos and more). Thereby we think that any solution which GML is part of, will be the most capable to achieve the objective of searching “Interoperability and integration of spatial source”.

Actually, the GML version 2.1.2 is an OpenGIS recommendation paper. Version 3 is now in draft, but it is for sure the main trend of standardization for representation of spatial and geographic data. The GML is an XML encoding protocol for the transport and storage of geographic information, including both the spatial and non-spatial properties of geographic features (geographic features can be considered as “an abstraction of real world phenomenon; it is a geographic feature if it is associated with a location relative to the Earth”).

In GML the way to face the problem of multi-resolution and multi-representation is achieved by using some application schema, or style XSLT(eXtensible Stylesheet Language Transformations (X3C, 1999)). This allows the creation of many versions of a set of spatial or geographic data.

Just as OMT-G divide the levels for presentation, GML has been designed to uphold the principle of separating content from presentation. GML provides mechanisms for the encoding of geographic feature data without considering how the data may be presented to a human reader. Since GML is an XML application, it can be readily styled into a variety of presentation formats, including vector and raster graphics, text, sound and voice. Generation of graphical output such as maps is one of the most common presentations of GML and this can be accomplished in a

variety of ways including direct rendering by graphical applets or styling into an XML graphic technology.

#### 4.4.1 Application Schema

GML provide three XML schema base documents: feature.xsd, geometry.xsd and xlink.xsd. These base GML schemas effectively provide a meta-schema, or a set of foundation classes, from which an application schema can be constructed. Also, the GML specification provide the guidelines and rules for developing the application schema. For the version 3 of GML, the group of OpenGIS is preparing the use of profiles and of more base schemas ; however, it has no incidence on our solution in interoperability considering that we are working at the XMI level.

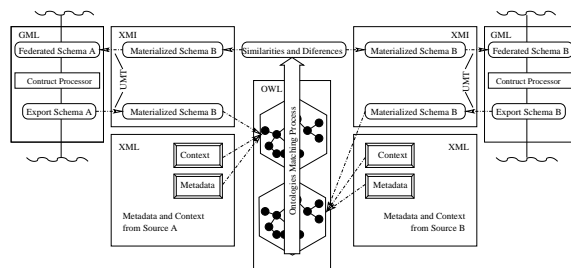


Figure 3: Creating Ontologies from GML Schemas

The UML Model Transformation Tool, UMT (Old- evik, 2002) is capable of transforming some XMI into GML and opposite. With UMT, we can obtain the model in XMI, then the process is similar than applied to OMT-G. In the Fig. 3 we presented a framework working with this technology.

From here, we can make use of UMT as a tool for coding and decoding the GML model, in order to obtain the XMI model and transform it into GML, as well as for the reverse step (to obtain the GML from XMI).

## 5 FUTURE WORK

In this paper we have presented a framework for integration of spatial data, spatial databases and most spatial sources. But we are working on a product software that could be applied to the framework purpose. We are waiting the results from the W3C for adopting OWL for representing the ontologies and the new version of UMT for transforming the XMI into some GML v3. Also, the time dimension is out of the scope of this paper. Thereby, as a future work, it is necessary to take the temporal dimension into

account.

## Acknowledgments

Part of this work has been supported by: Ibero-Americana Spanish Agency of Cooperation (spanish acronym AECI) MUTIS Program; and the Spanish Research Program PRONTIC under projects TIC2000-1723-C02-01. Special acknowledgments to Michael Gould (coordinator of OGC documentation subcommittee), to Etienne Canaud (LISI Université Claude Bernard) for their valuable contributions for this paper.

## REFERENCES

- Abelló, A., Oliva, M., Rodríguez, M. E., and Saltor, F. (1999). The syntax of bloom99 schemas. Technical Report LSI-99-34-R, LSI-UPC.
- Benjamins, V. R. and Fensel, D. (1998). The ontological engineering initiative ( $ka$ )<sup>2</sup>. In Guarino, N., editor, *Formal Ontology in Information Systems*, pages 287–301. IOS Press.
- Borges, K. A., Davis, C. A., and Laender, A. H. (2001). Omt-g: An object-oriented data model for geographic applications. *GeoInformatica*, 5(3):221–260.
- Castellanos, M., Saltor, F., and García-Solaco, M. (1992). A canonical model for interoperability among object-oriented and relational databases. In Özsu, M. T., Dayal, U., and Valduriez, P., editors, *Distributed Object Management: Papers from the International Workshop on Distributed Management (IWDOM)*, pages 309–314, Edmonton, Alberta, Canada. Morgan Kaufmann.
- Fonseca, F., Egenhofer, M., Agouris, P., and Câmara, C. (2002). Using ontologies for integrated geographic information systems. *Transactions in GIS*, 6(3).
- Fonseca, F. T. (2001). *Ontology-Driven Geographic Information*. PhD thesis, University of Maine, Orono, Maine 04469.
- Günter, O. and Voisard, A. (1998). *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*, chapter Metadata in geographic and environmental data management, pages 57–87. McGraw Hill.
- Hayes, S. and Ram, S. (1990). Multi-user view integration system (muvis). In *Proceedings of the 6th International Conference on Data Engineering*.

- Heflin, J. and Hendler, J. (2000). Semantic interoperability on the web. *Extreme Markup Languages 2000*. Retrieved from <http://www.cs.umd.edu/projects/plus/SHOE/pubs/extreme2000.pdf>.
- Kilpeläinen, P. and Mannila, H. (1995). Ordered and unordered tree inclusion. *SAIM J.COMPUT.*
- Mena, E., Kashyap, V., Illarramendi, A., and Sheth, A. (1998). Domain specific ontologies for semantic information brokering on the global information infrastructure. In Guarino, N., editor, *Formal Ontology in Information Systems*. IOS press.
- Oldevik, J. (2002). Uml model transformation tool. Retrieved from <http://www.modelbased.net/umt/>.
- OMG, O. M. G. (2001). Omt unified modeling language specification v1.4. Retrieved from <http://www.omg.org/>.
- OMG, O. M. G. (2002). Xml metadata interchange (xmi) specification. Retrieved from <http://www.omg.org/>.
- OpenGIS (1999). The opengis abstract specifications. topic 5: Features v4. Retrieved from <http://www.opengis.org/techno/abstract/99-105r2.pdf>.
- OpenGIS (2002). Geography markup language (gml) implementation specification v2.1.2. Retrieved from <http://www.opengis.org/>.
- Rodríguez, M. A. (2000). *Assessing Semantic Similarity Among Spatial Entity Classes*. PhD thesis, University of Maine, Orono, Maine 04469.
- Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., and Lorensen, W. (1991). *Object-Oriented Modeling and Design*. Prentice Hall.
- Samos, Abelló, Oliva, Rodríguez, Saltor, Sistac, Araque, Delgado, Garví, and Ruiz (1999). Sistema cooperativo para la integración de fuentes heterogéneas de información y almacenes de datos. *Novatica ATI*. (in Spanish).
- Sheth and Larson (1990). Federated database systems for managing distributed heterogeneous and autonomous databases. *ACM Computing Surveys*, 22(3).
- Sheth, A., Larson, J., Cornellio, A., and Navathe, S. (1988). A tool for integrating conceptual schemas and user views. In *Proceedings of 4th International Conference on Data Engineering*, pages 176–183.
- Sheth, A. P. (1999). *Interoperating Geographic Information System*, chapter Changing focus on Interoperability from System, Syntax, Structure to Semantics, pages 5–29. Kluwer Academic Publisher.
- Spaccapietra, S., Parent, C., and Vangenot, C. (2000). Gis databases: From multiscale to multirepresentation. In Choueiry, B. Y. and Walsh, T., editors, *Abstraction, Reformulation, and Approximation, 4th International Symposium, SARA 2000*, volume 1864 of *Lecture Notes in Computer Science*, pages 57–70, Horseshoe Bay, Texas, USA. Springer.
- Thiran, P., Chougrani, A., Hainaut, J.-L., and Hick, J.-M. (2000). Case support for the development of federated information systems. In *Proceedings of the 3rd Workshop EFIS 2000*, pages 106–113, Dublin, Ireland. EFIS, IOS Press.
- W3C (2000). Extensible markup language (xml) 1.0. Retrieved from <http://www.w3.org/TR/REC-xml>.
- W3C (2002). Owl, web ontology language 1.0 reference(draft). Retrieved from <http://www.w3.org/TR/2002/WD-owl-ref-20020729/>.
- X3C (1999). Xsl transformations (xslt). Retrieved from <http://www.w3.org/TR/xslt>.