Integration of Heterogeneous Databases and Ontologies

Adriana dos Santos Aparício[1], *Oscar Luiz Monteiro de Farias*[1], *Neide dos Santos*[1, 2]
[1] Post-Graduation Program in Computer Engineering
[2] Department of Computer Science
Universidade do Estado do Rio de Janeiro
adriana.aparicio@globo.com, fariasol@eng.uerj.br, neide@ime.uerj.br
Rio de Janeiro - Brasil

Abstract

Research in interoperability has been motivated by the growing heterogeneity of computing systems. Heterogeneity can occur in many levels and each level of heterogeneity requires an isolated or integrated approach for solution. In this paper, we propose the specification of a formal ontology for the information related to a specific domain of a database system, to work together with a global scheme, developed as software layer among the different databases under consideration. To test this approach we elaborated a case study, based upon hypothetical queries submitted to relational and heterogeneous databases, with data on soil domain, aiming at identifying the kinds of soil most appropriate to a certain culture. The case study demonstrated that the semantic conflicts were circumvented and the integration of the databases was easily reached.

1. Introduction

Research in interoperability has been motivated by the growing heterogeneity of computing systems and the need to interchange information and processes among heterogeneous computing systems environments (Yuan, 1998). Sheth (1998) identifies the system, syntactic, structural and semantic levels of heterogeneity. The system level includes incompatible hardware and operating systems; the syntactic level refers to different languages and data representations; the structural level includes different data models and the semantic level refers to the meaning of terms using in the interchange.

Wache et al (2001) have classified several types of semantic heterogeneity. Cui, Jones and O'Brien (2002) argue that many technologies have been developed to tackle these types of heterogeneity.

Cui, Jones and O'Brien' (Cui, Jones and O'Brien, 2002) solution to the problem of semantic heterogeneity is to formally specify the meaning of the terminology of each system and to define a translation between each system terminology and an intermediate terminology. They specified a system and intermediate terminologies using *formal ontologies* and the translation between them using *ontology mapping*s. A formal ontology consists of definitions of terms. It usually includes concepts with associated attributes, relationships and constraints defined between the concepts and entities that are instances of concepts.

In the realm of database systems, different ways of representing reality lead to different conceptual models. But, once organizations (firms, universities, etc.) don't adhere to a common conceptual model, it is not always possible to interchange information among different data base systems.

In this paper, we propose a solution for integration of heterogeneous databases that is similar, in some aspects, to the solution proposed by Cui, Jones and O'Brien. We also specify a formal ontology for the information related to a specific domain of a database system. However, instead of developing a system to map ontologies, we promote the integration of conceptual schemes, developing a software layer among the different databases under consideration. To test the solution, we elaborated a case study, based upon hypothetical queries submitted to relational and heterogeneous databases, with data on soil domain, aiming at identifying the kinds of soil most appropriate to a certain culture. The results showed up that the semantic conflicts were circumvented and the integration of the databases was easily reached.

2. Interoperability in Heterogeneous Databases

Interoperability among different software applications and system components is a key to the successful integration of digital information. Nowadays, there are several interoperability specifications and standards at various stages of development and adoption, promoted by a number of organizations and consortiums. Even so, a lack of interoperability mechanisms among heterogeneous platforms remains. Proposed issues recommend open service architecture to build standard-driven distributed and interoperable systems, based on the definition of open software interfaces for each subsystem in the architecture, avoiding any dependency from specific information models (Anido-Rifón et al, 2002).

Casanova, Brauner, Câmara e Lima Júnior (2002) argue that the real interoperability demands solutions able to deal with heterogeneous data in its format and structure as well as in its interpretation and meaning. The authors report three strategies to solve the integration of heterogeneous data. The first one consists in to generate mappings among pairs of data source. The strategy grows in complexity in accordance with the growing of the sources number. The second one uses a community description of data, by means of the global scheme, the mediated scheme or reference scheme, depending on the adopted approach to reach interoperability, the integration of conceptual schemes, developing a software layer among the different databases under consideration.

Summarizing, the approach maps the data source description, called local schemes or schemes for exportation. It avoids creating two-to-two mappings among the local schemes, but the strategy requires that all data source is known *a priori*.

The integration of the conceptual schemes is a well establish and utilized approach. This global scheme emerges from the integration of the different local conceptual schemes and it consists in an intermediate software layer proving access to the involved databases. Queries into the global scheme are mapped to the local scheme, where the needed information is stored in an integrated and non-redundant way. The integration requires:

a) the comparison of local schemes, where equivalencies and conflicts are identified;
b) (ii) the adequacy of schemes, where the eventual conflicts are solved; and,
c) the integration and restructuring of schemes, where local schemes are integrated by means of common concepts.

Global scheme strategy can solve syntactic and structural heterogeneity, but it does not guarantee the solving of semantic interoperability.

The third strategy adopts ontology to formalize the reference scheme and the local schemes (Casanova, Brauner, Câmara e Lima Júnior, 2002). Semantic interoperability could be solved via the use of classes derived from ontology, where all handling of information should be based on the term definition met on the ontology.

Ontology can help searching for interoperability in the heterogeneous databases integration since it establishes a joint terminology between members of a community of interest. Ontology is generally considered to provide definitions for the vocabulary used to represent knowledge. It can be seen as a scheme that provides precise and complete models of particular domains. We particularly used the notion of domain ontology (Guarino, 1998) that describes precisely the basic concepts found in particular domains. The model that specifies the domain terms, shaping a semantic net of terms, is called features model. The features model (Cohen, 1994) captures the general features of available software application in specific domains and allows the insertion of new terms in as domains grow up.

Cui, Jones and O'Brien' solution to the problem of semantic heterogeneity is to formally specify the meaning of the terminology of each system and to define a translation between each system terminologies and an intermediate terminology. They specified a system and intermediate terminologies using *formal ontologies* and specified the translation between them using *ontology mapping*s. A formal ontology consists of definitions of terms. It usually includes concepts with associated attributes, relationships and constraints defined between the concepts and entities that are instances of concepts.

To test this approach we elaborated a case study, based upon hypothetical queries submitted to relational and heterogeneous databases, with data on soil domain, aiming at identifying the kinds of soil most appropriate to a certain culture. First, we

modeled and built soil ontology, supported by the well-known ontology editor *Protégé 2000*. Second, we developed a software layer integrating the different conceptual models (schemes), in order to create a global virtual scheme.

3. Semantic Interoperability: Combining Ontology and Global Conceptual Scheme

To share and interchange information among different database systems involves the availability of a common vocabulary, because semantic conflicts emerge from the lack of standardization (consistency) in the meaning of concepts, terms and structures found in the data source. Ontology can help the calling for standardization since it demands a precise semantic representation.

In our point of view, the integration of heterogeneous databases calls for the specification of a formal ontology for the information related to a specific domain of a database system and the use of conceptual schemes, developing a software layer among the different databases under consideration. The terms of the ontology offer the support to model queries in the available heterogeneous databases as well they should help databases developers to compose the future conceptual models, i.e., classes and attributes.

Ontology specification

Ontology development is necessarily an iterative process. The first steps to build ontology imply delimitating the ontology scope, and acquiring and validating the domain knowledge. We must:

- determine the domain and scope of the ontology,
- enumerate important terms in the ontology,
- define the classes and the class hierarchy and the properties of classes—slots and the facets of the slots, and
- create instances.

In a pragmatic point of view, ontology is a set of concepts, properties and restrictions. Properties of each concept describe features and attributes of the concept (slots, sometimes called roles or properties).

Concepts are the focus of most ontologies, because they describe the classes in the domain (Noy and McGuinness). A class can have subclasses that represent concepts that are more specific than the superclass. Slots describe properties of classes and instances. From this point of view, developing an ontology includes:

- defining classes in the ontology,
- arranging the classes in a taxonomic (subclass–superclass) hierarchy,
- defining slots and describing allowed values for these slots,
- filling in the values for slots for instances.

We start to specify the soil ontology studying the related concepts and the domain.

Concepts

The main concepts related to soil classification were layers and horizons. The soils are composed of parallel sections, called horizons or layers. The formation of the layers or soil horizons is a result of the environmental forces that have acted upon the soil during its formation, often for thousands of years. The color, texture and structure of each horizon and often its chemical characteristics are used to group soils and form the basis of most systems of soil classification. Almost all systems of soil classification are based on the morphology of soils. The systems organize or group soils into a hierarchy of five levels, composing a taxonomy in six categorical levels: Order, Sub order, Great Group, Subgroup, Family and Series.

Soil domain

Soil is a complex mixture of mineral matter, organic matter and living organisms. Soil is a product of the environment, constantly changing, constantly evolving. It develops over time, sometimes very slowly in dry desert areas or more quickly in wet tropical regions.

Soils can be studied on physical, chemistry or biology perspectives. Soils are a complex three-phase system composed of solids, liquids and gases. The study of the physical behavior of these phases is called Soil Physics and includes: density and porosity, texture, structure, color, and movement. Soil Chemistry studies the chemical characteristics of soil, which depends on their mineral composition, organic matter and environment. Soil Biology is the study of the living component of soils. Numerous bacteria, fungi, worms, insects, small rodents and mammals inhabit the soil. Many

of these organisms help in maintaining the fertility of the soil by decomposing plant and animal residues, which recycle the nutrients.

The domain is complex, and the process of acquiring knowledge renders more difficult due to the heterogeneity of vocabulary found in the reports on soil. For our study of case, the ontology scope was limited to the classification of Brazilian soils. The domain knowledge was mainly obtained from the report System of Soil Classification, official source of soil information of Embrapa, the Brazilian governmental board on Agriculture.

Relevant features for modeling the soil ontology

The soil ontology was modeled from six main concepts or classes: Morphology, Profile, Diagnostic Attributes, Diagnostic Superficial Horizons, Diagnostic Sub-superficial Horizons and Classification. Soil classification begins describing the morphological features of soil profile, including color, texture, consistency and transition. They provide the base for defining the diagnostic horizons. The profile allows the study of environment features, such as Relief, Erosion, Drainage, Primary Vegetation, Roots and Biological Factors.

From this theoretical basis, the ontology was represented by the use of the ontology editor Protégé 2000 (www.protégé.stanford.edu). Protégé 2000 is a free, open source ontology editor and knowledge-base framework. Protégé is based on Java, is extensible, and provides a foundation for customized knowledge-based applications.

The analysis of some available tools aiming at building the soil ontology showed us that Protégé was useful for our purposes, since it works with a model of extensible knowledge, which allow redefine primitive classes (metaclasses) of a representational system, in a declarative way.

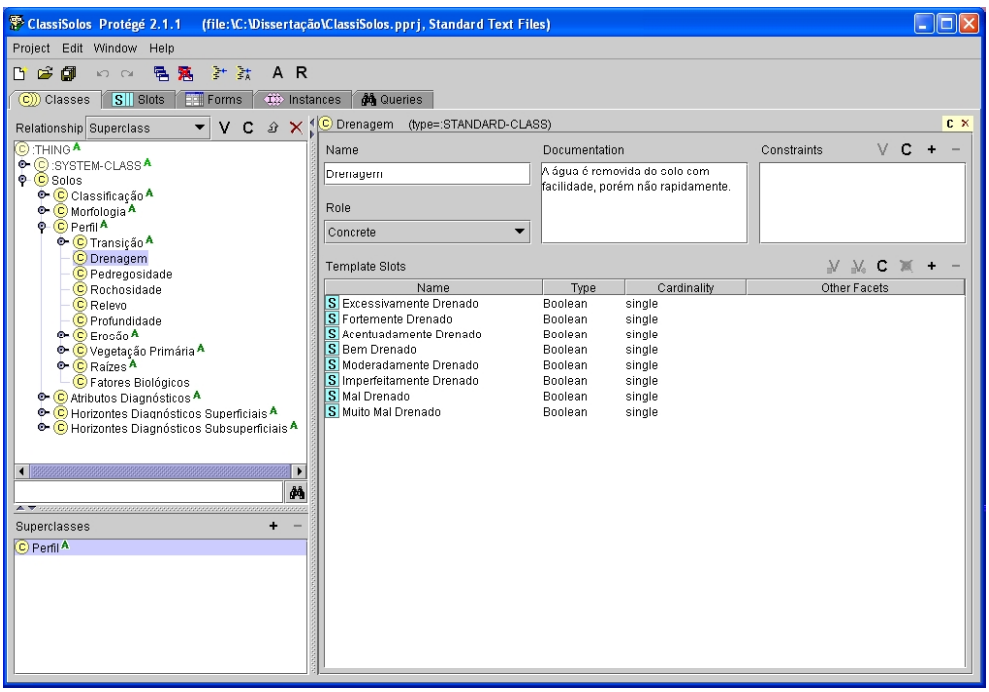

Figure 1. Modeling the Soil Profile: Partial Visualization

SIBDAR Prototype

The prototype SIBDAR was developed to allow users in the integration of distributed and heterogeneous databases. Our solution assumes that many governmental or non governmental organisms, universities and research institutes, interested in determined knowledge area, would be potential users of Heterogeneous Database Systems (HDBS), because they needed to share and

interchange information. If these organisms adhere to a ontology, our prototype would easily allow the integration of their databases.

The prototype gets the local schemes of the each heterogeneous database under consideration, and creates an unique and virtual global scheme.

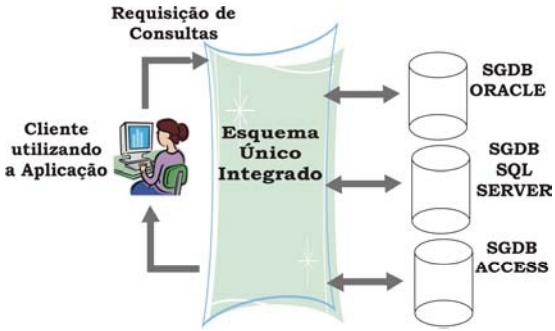

Figure 2. System Architecture

For that, the user specifies the drivers related to the database management systems (DBMS), with which he will work (from each heterogeneous database management system) and that he needs to consult all elements found in each DBMS. In addition, the user can establish relationships among entities of different databases that will be recorded in the virtual global scheme. From this moment, the user can formulate queries in SQL in the usual way.

To test the approach we elaborated a case study, based upon hypothetical queries submitted to relational and heterogeneous databases, with data on soil domain, aiming at identifying the kinds of soil most appropriate to a certain culture. In our case, the citric culture.

4. Case Study: Soils to Citric Culture

As mentioned, we specified a ontology on classification of Brazilian soils, called ClassSolos. From the ontology, we compose the data bases aiming at to answer the aiming at identifying the kinds of soil more appropriate to a the culture of citric and use the prototype SIBDAR. Based on the Brazilian System of Soil Classification, we identify the ideal features to the citric culture and verify that we need for information about the morphological and the environment features of soils. The citric adapt both to arenaceous and argillaceous soils. The soils more appropriate to a commercial culture of citric are the areno-argillaceous. The citric does not tolerate impermeable soils, and rasos soils or soils that make marshy easily must be avoided. Looking at the ontological terms of ClassSolos and the citric characteristics met at System of Soil Classification, we compose the table 1.

Table 1. Characteristics of Appropriate Soils to Citric Culture.

Main Characteristics	
Texture	Arenaceous, Areno-argillaceous or Argillaceous
Relief	Plan, Soft waved or Waved
Depth	Little deep and Deep
Draining	Strongly drained, Very strongly drained and Well drained

Identified the ontological terms, used in the definition of the morphologic characteristics and profile, it was easy to identify the tables that must be accessed: Texture, Relief, Depth, Draining and Soils. Figure 3 presents the model of entity relationship that will be used.

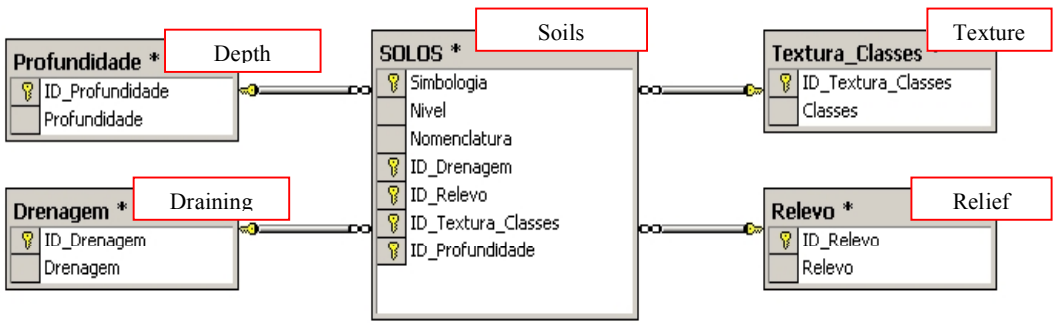

Figure 3. Entity Relationship Model

In this point of the case study, we use SIBDAR to access the databases, to define the relationships, to create the filters and to run the consult. Loaded the tables Texture_Class, Porosity_Size_Pores, Draining, Relief, Depth and a table named Soil, we created the relationships. In this specific case, we create the relationships: Soils with Draining; Soils with Depth; Soils with Texture_Classes; Soils with Relief; Soils with Porosity_Size_Pores.

So, we built the necessary filters to obtain the final result, by selecting all the types of soils whose characteristics are the desirable ones for the Culture of Citric. The creation of the filters requires following a simple set of steps, carried through from the graphical interface of SIBDAR:

- Select the types of soil that have texture arenaceous, texture areno- argillaceous and texture argillaceous
- Select the types of soil that have plain relief or wavy soft relief.
- Select the types of soil that have depth = little deep.
- Select the types of soil that soil Draining = Strong Drained or Draining = Very strong drained or Well drained

Created the filters, the option to "Run filters" and to "Update data" shows the answer: LATOSSOLOS and its determined levels of classification are the ideal types for the culture of citric.

5. Conclusions and Future Works

The arising and the fast dissemination of different database management systems have resulted in serious problems, such as interoperability among heterogeneous systems. Many solutions have been proposed, mainly solutions based on the development of a global scheme and on the specification of formal ontology. Ontology can be helpful for the effort of integrating heterogeneous databases, since it potentially solves the semantic conflicts, that the global scheme is not able to solve.

In this paper, we present an approach to lead with data heterogeneity by means of specifications of a formal domain ontology and the use of a global scheme, developed as software layer among the different databases under consideration. In our case, the global scheme is reached by the use of SIBDAR. It allows to access tables in its respective HDBSs, thus working with the concept of virtual database, creating, in the system memory, only a reference to the tables of information stores in original databases.

The case study demonstrated that our approach circumvented the semantic conflicts and make easy the database integration. Crucial, in our approach, is the widespread acceptance, for a given community, of the vocabulary, restrictions, and relationships related to the reality and that are captured by the ontology and transformed to the target conceptual models.

Our expectation is that, in a near future, organizations, researchers, practitioners and software developers walk together for creating a comprehensive and democratic repository of ontology, really allowing the sharing and interchanging of information among their heterogeneous systems.

References

[Anido et all, 2002]. L. Anido, J.Santos, J. Ródríguez, M. Caeiro, M. Fernández, M. Llamas. A Step ahead in E-learning Standardization: Building Learning Systems from Reusable and Interoperable Software Components.. *Proceedings of the 11th International World Wide Web.*

[Casanova et all]. M. Casanova, D. Brauner, G. Câmara, P. Lima Júnior. Integration and interoperability among geographical data source. *Geographical Databases*. (Eds. J. A Paiva, M. Casanova, R. Cartaxo e G. Câmara). In http://www.dpi.inpe.br/gilberto/livro/bdados. (acessed May 20 2005). In Portuguese.

[Cohen, 1994]. S. Cohen. Feature-Oriented Domain Analysis: Domain Modeling; Tutorial Notes; *3rd International Conference on Software Reuse*; Rio de Janeiro, November.

[Cui, Jones, O'Brien, 2002]. Zhan Cui, Dean M. Jones, Paul O'Brien: Semantic B2B Integration: Issues in Ontology-based Applications. *SIGMOD Record* 31(1): 43-48.

[Guarino, 1998]. N. Guarino. Formal Ontology in Information Systems. *Proceedings of FOIS'98*, Trento, Italy, June. Amsterdam, IOS Press, pp. 3-15.

[Noy and McGuinness]. N. Noy and D. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. In http://protege.stanford.edu/publications/ ontology_ development/ontology101-noy-mcguinness.html

[Sheth, 1999] Sheth, A P. Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. M. F. Goodchild, M. J. Egenhofer, R. Fegeas, and C. A. Kottman (eds.) *Interoperating Geographic Information System*s, Kluwer.

[Wache et all, 2001]. H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hübner, Ontology-based integration of information - a survey of existing approaches. In: H. Stuckenschmidt (ed.), *IJCAI-01 Workshop: Ontologies and Information Sharing*, Seattle, WA, pp. 108-117.

[Yuan, 1998] Yuan, X, Interoperability of Heterogeneous Geographic Information Processing Environment for Internet GIS. *Journal of Wuhan Technical University of Surveying and Mapping,* Wuhan, P.R China