

# Applying Ontologies in the Integration of Heterogeneous Relational Databases

Adriana S. Aparício<sup>1</sup> Oscar L. M. Farias<sup>2</sup> Neide dos Santos<sup>3</sup>

Post-Graduation Program in Computer Engineering,  
Rio de Janeiro State University,  
Rua São Francisco Xavier 524, bloco D, sala 5028, Maracanã  
20550-900 Rio de Janeiro, Brazil

<sup>1</sup>adriana.aparicio@globo.com, <sup>2</sup>fariasol@eng.uerj.br,  
<sup>3</sup>neide@ime.uerj.br

## Abstract

Interoperability is a key point for integrating heterogeneous computing systems. A usual approach proposes the integration of the conceptual databases schemes into a global conceptual scheme, to resolve syntactic and structural heterogeneity. Moreover, semantic problems can remain. Ontologies have been largely designated to overcome semantic heterogeneity. We propose the specification of a formal ontology about the specific knowledge domain to be shared among several database systems build a posteriori of the ontology specification. We reach this integration through a global scheme, developed as a software layer among the databases under consideration. To test this approach we elaborated a case study, based upon hypothetical queries submitted to heterogeneous databases, with data on soil domain, to identify the soil most appropriate to a certain culture. The results are promising, but crucial, in our approach, is the acceptance for a given community of a common vocabulary and its relationships and that are captured by the ontology and transformed to the target conceptual models.

## 1 Introduction

Ontology is generally considered to provide definitions for the vocabulary used to represent knowledge. The ontology role is to reflect a community's consensus on a useful way to conceptualize a particular domain. In this way, ontology can be a useful support for software reuse, since it establishes a joint terminology between members of a community of interest. It is also a useful support for search engine design since it provides means for machine-based knowledge sharing and reuse between applications. It can be seen as a schema that provides precise and complete models of particular domains. In the process of building ontology, each term must be defined by means of a formal and informal description, as well as the specification of the relationships among the terms, shaping a semantic net.

Copyright © 2005, Australian Computer Society, Inc. This paper appeared at the Australasian Ontology Workshop (AOW 2005), Sydney. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 58. T. Meyer and M. Orgun, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

Building ontology, however, can mean different things to different practitioners. The distinction of how one carry out describing something reflects a progression in ontology from: simple lexicons or controlled vocabularies to categorically organized thesauri; to hierarchies called taxonomies where terms are given distinguishing properties; to full-blown ontologies where these properties can define new concepts and where concepts have named relationships with other concepts.

According to Fikes (1996), adopting reusability as a primary goal for ontology has a significant impact on the tools and methodologies that are needed for ontology creation and use. For example, developers need to make ontology accessible and understandable to a community of use. So, new techniques are needed for translating ontology between representation formalisms and for describing the competency of ontology. Also, when knowledge is encoded in a new ontology specifically for use inside a community, it is reasonable to expect that this community will have a serious involvement in the encoding process. So, tools that support collaborative development and evolution of ontology would appear to be important in achieving desired levels of reusability.

On the other hand, interoperability among heterogeneous systems remains as a serious problem. 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 (1999) identifies system, syntactic, structural and semantic levels of heterogeneity. The system level includes hardware and operating systems incompatibilities; 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 used in the interchange. Wache et al (2001) have also classified several types of semantic heterogeneity.

Cui et al (2002) point out that many technologies have been developed to tackle these types of heterogeneity. Cui et al 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 mappings*. 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 database systems. In this paper, we propose a solution for integration of heterogeneous databases that is similar, in some aspects, to the solution proposed in [5] to cope with semantic heterogeneity. We specify a formal ontology for the information related to a specific knowledge domain that is, *a priori*, shared by several heterogeneous databases. However, there is no need to develop a system to map ontologies, since all databases already share the same ontology. We only need to promote the integration of conceptual schemes, developing a software layer among the different databases under consideration. This is a feasible task, because all heterogeneous databases refer to the same concepts, definitions and constraints, as established by the common formal ontology. What is still need in the integration of the conceptual schemes, is merely defining the relationship among data pertaining to different databases. To test this 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 promoted by a number of organizations and consortiums, at various stages of development and adoption. Even so, a lack of interoperability mechanisms among heterogeneous platforms remains. Proposed issues recommend open service architectures 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 et al., 2002).

Casanova et al (2005) argue that 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 generate mappings among pairs of data sources. This strategy grows in complexity, in accordance with the growing number of sources. The second one uses a community description of data, by means of a global scheme, a mediated scheme or a reference scheme; depending on the adopted approach to reach

interoperability, the integration of conceptual schemes is reached through the development of a software layer among the different databases under consideration.

Summarizing, this approach maps the data source description, called local schemes or schemes for exportation into a global schema. It avoids creating mappings among each pair of 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. A global scheme emerges from the integration of the different local conceptual schemes and it consists in an intermediate software layer providing 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:

- (i) comparison of local schemes, where equivalencies and conflicts are identified;
- (ii) adequacy of schemes, where the eventual conflicts are solved; and, (iii) 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 ontologies to formalize the reference scheme and the local schemes (Casanova et al, 2005). Semantic interoperability could be solved via the use of classes derived from the ontologies, where all handling of information should be based on the definition of terms met on the ontologies.

Ontologies can help in solving the interoperability problems among heterogeneous databases, 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 schema that provides precise and complete models of particular domains.

## 3 Semantic Interoperability: Ontology and Global Scheme

To share and interchange information among different database systems involve 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.

Usually there are two ways to resolve semantic conflict using the concept of ontology. One way is to build the ontology for each system, then directly to create the mapping between these ontology-based systems. Another way is to create common ontology to specify the vocabularies and terms to describe and interpret shared information among its users, and build the mapping

between the systems through the common ontology (Huang, 2002). In our approach, ontology describes and organizes the semantic of the common standard vocabulary and the relationship between them, including concepts, instances, relationship between concepts, and relationship between instances.

The integration of heterogeneous databases calls for the specification of a formal ontology for the information related to a specific domain (that can be thought as the union of the domains of all databases involved), and the integration of conceptual schemes, through a software layer. The different database schema can be easily integrated, since they share the same terminology, derived from the formal ontology. This software layer will basically establish the relationship among entities pertaining to different databases, inasmuch all the syntactic and semantic conflicts were eliminated by using a common ontology. 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 future conceptual models.

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.

### 3.1 Ontology Specification

Ontology development is necessarily an iterative process. The first steps to build a specific ontology imply delimitating the ontology scope, and acquiring and validating the domain knowledge. Briefly, we must: I) determine the domain and scope of the ontology; ii) enumerate important terms in the ontology; iii) define the classes and the class hierarchy and the properties of classes—slots and the facets of the slots; iv) create instances.

In a pragmatic point of view, ontology is a set of concepts, properties and restrictions. Each concept has some properties describing features and attributes of the concept (slots, sometimes called roles or properties). Concepts are the focus of most ontology, because they describe the classes in the domain. A class can have subclasses that represent concepts that are more specific than the super-class. Slots describe properties of classes and instances. From this point of view, developing ontology includes: I) defining classes in the ontology; II) arranging the classes in a taxonomic (subclass/super-class) hierarchy; III) defining slots and describing allowed values for these slots; IV) Instantiating classes and filling in the values for slots. In the next subsection we start to specify the soil ontology, studying the related concepts and the domain.

**Concepts:** the main concepts related to soil classification are 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 Brazilian System of Soil Classification (1999) organizes or groups soils into a hierarchy of six levels, composing taxonomy of six categorical levels (in brackets we have the total number of classes involved): Order (14), Sub-order (44), Great Group (150), Subgroup (580), Family (unforeseen) and Series (unforeseen).

**Soil domain:** soil is a complex mixture of mineral matter, organic matter and living organisms. It 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. They 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 soil 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 Brazilian System of Soil Classification (1999), 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 (color, texture, consistency and transition). They provide the basis for defining the diagnostic horizons. The profile allows the study of environment features (Relief, Erosion, Drainage, Primary Vegetation, Roots and Biological Factors). From this theoretical basis, presented here in a very succinct way, the ontology was represented. We used, as a tool for building the ontology, the software Protégé 2000.

### 3.2 The SIBDAR Prototype

The SIBDAR prototype was developed to allow users in the integration of distributed and heterogeneous databases. Our solution assumes that many governmental or non governmental organisms, 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 common ontology in their domain of interest, our prototype would allow the integration of their databases. The prototype gets the local schemes of each heterogeneous database under consideration, and creates a unique and virtual global scheme. Since all databases refer to the same ontology, theoretically there are not syntactic neither semantic conflicts. What the users of the prototype need to do is to establish the relationships among entities (terms in ontology vocabulary) pertaining to different database schema. For that, the user specifies the drivers related to the database management systems with which he will work (from each heterogeneous database management system) and that he needs to consult. Following, the user may establish relationships among entities of different databases that will be recorded in the virtual global scheme. Now, 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 culture of citric.

### 4 Case Study: Soils for Citric Culture

From *ClassSolo* and the technical data related to the culture of citric we built several relational heterogeneous databases (in *Oracle*, *SQL Sever* and *Access*) aiming at identifying the kinds of soil more appropriate to the culture of citric.

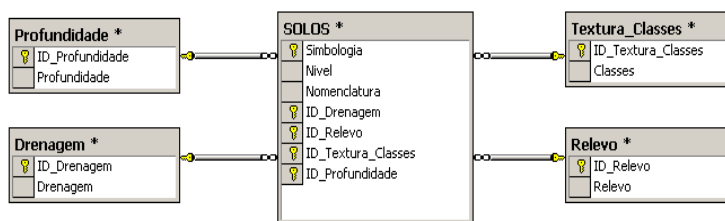
Based on the Brazilian System of Soil Classification (1999), we identified the ideal features for the citric culture and verified what kind of information, about the morphological and the environment features of soils, was needed. Citric adapt both to arenacious and argillaceous soils. The soils more appropriate to a commercial culture are the areno-argillaceous. The citric does not tolerate impermeable soils; superficial soils and marsh must be avoided. Looking at the ontological terms of *ClassSolos* and the citric characteristics met at the Brazilian System of Soil Classification, we compose the table 1.

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

**Table 1. Characteristics of Appropriate Soils for Citric Culture**

Identified the ontological terms, used in the definition of the soil morphologic characteristics and profile, it was easy to choose the tables that must be accessed: *Texture*, *Relief*, *Depth*, *Draining* and *Soils*.

Figure 1 presents the entity relationship model that was used.



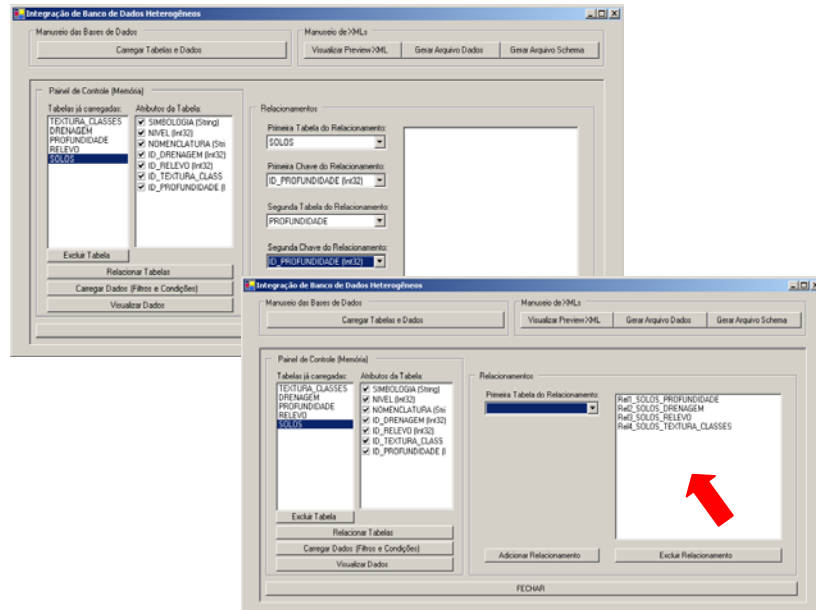
**Fig 1. Entity Relationship Model**

At this point of the case study, we used SIBDAR to access the databases, define the relationships, and submit the SQL query. Loaded the tables *Texture\_Class*, *Draining*, *Relief*, *Depth* and *Soil*, stored at different databases, we created the relationships: Soils with Draining; Soils with Depth; Soils with Texture\_Classes; Soils with Relief.

Then, we built the necessary filters to obtain the corresponding *SQL query*, 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 arenacious, 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; and,
- Select the types of soil that soil
  - Draining = Strong Drained or
  - Draining = Very strong drained or Well drained.



**Fig 2.** Creating relationships among soil table and depth table

Created the filters (figure 2), the option “Run filters” and “Update data” show the answer: LATOSSOLOS and its

determined levels of classification are the ideal types for the culture of citric.

Nível	Nomenclatura	ID_Drenag	ID_Rele	ID_Textura	ID_Profundid
1	LATOSSOLOS	2	1	2	3
1	LATOSSOLOS	3	1	2	3
1	LATOSSOLOS	4	1	2	3
1	LATOSSOLOS	2	1	3	3
1	LATOSSOLOS	3	1	3	3
1	LATOSSOLOS	4	1	3	3
2	LATOSSOLOS AMARELO	2	1	2	3
2	LATOSSOLOS AMARELO	3	1	2	3
2	LATOSSOLOS AMARELO	4	1	2	3
2	LATOSSOLOS AMARELO	2	1	3	3
2	LATOSSOLOS AMARELO	3	1	3	3
2	LATOSSOLOS AMARELO	4	1	3	3
3	LATOSSOLOS AMARELO	2	1	2	3
3	LATOSSOLOS AMARELO	3	1	2	3

**Fig. 3** Final result showing “Latossoilo” as the ideal soil to citric culture

## 5 Conclusions and Future Works

The arising and 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 a 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 specification of formal domain ontology and the use of a global scheme, developed as software layer among the different databases under consideration. In this layer the user needs only to specify the relationships among entities pertaining to different databases. In our case, the global scheme is reached by

the use of SIBDAR. It allows access to data in its respective HDBS, working with the concept of a virtual database, creating, in the system memory, only a reference to the tables of information stored in original databases.

The case study demonstrated that our approach can circumvent semantic conflicts and promote the database integration, under our study circumstances. Crucial, in our approach, is however 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 ontologies, really allowing the sharing and interchanging of information among heterogeneous systems.

## 6 References

- Anido, L., Santos, J., Rodríguez, J., Caeiro, M., Fernández, M., and Llamas, M. A Step ahead in E-learning Standardization: Building Learning Systems from Reusable and Interoperable Software Components. In *Proceedings of the 11th International World Wide Web -WWW2002*, 2002.
- Casanova, M., Brauner, D., Câmara, G., and Lima Júnior, P. Integration and interoperability among geographical data source. In *Geographical Databases*. (Eds. João Argemiro Paiva, Marco Casanova, Ricardo Cartaxo e Gilberto Câmara), 2005. Available in Portuguese at <http://www.dpi.inpe.br/gilberto/livro/bdados>, accessed in May 20, 2005.
- Cui, Z., Jones, D. M., and O'Brien, P.: Semantic B2B Integration: Issues in Ontology-based Applications. *SIGMOD Record* 31(1), 2002, 43-48.
- EMBRAPA. *Sistema Brasileiro de Classificação de Solos*. Serviço de Produção de Informação – SPI, Brasília, DF, 1999.
- Fikes, R. Ontologies: What are they, and where's the research? 1996. In <http://ksl-web.stanford.edu/KR96/FikesPositionStatement.html>, accessed in August 19 2005
- Huang, Y.. Ontology-based Land Use Information Service on the Semantic Web, 2002. Available in <http://www.ucgis.org/summer03/studentpapers/yuxiahuang.pdf>.
- Protégé home page. Available at <http://www.protégé.stanford.edu>, accessed in May 20, 2005.
- Sheth, A P. Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. Edited by Goodchild, M. F. Egenhofer, M. J., Fegeas, R., and Kottman, C. A. In *Interoperating Geographic Information Systems*, Kluwer, 1999.
- Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. Ontology-based integration of information - a survey of existing approaches. In *IJCAI-01 Workshop: Ontologies and Information Sharing*, Edited by Stuckenschmidt, H., Seattle, WA, 2001, 108-117.
- Yuan, X, Interoperability of Heterogeneous Geographic Information Processing Environment for Internet GIS. In *Journal of Wuhan Technical University of Surveying and Mapping*, 1998, Wuhan, P.R China.