Proceedings of the IASTED International Conference
DATABASES AND APPLICATIONS, pp. 13-18,
Innsbruck, Austria, February 17-19, 2004. ACTA Press.
ISSN: 1027-2666

# Database Schema Detection and Mapping on Mobile Applications: an Ontology-based Approach

Villie Morocho
LSI-SI, Universitat Politècnica de Catalunya,
Jordi Girona 1-3. Barcelona, Spain.08034.
vmorocho@lsi.upc.es

Lluís Pérez-Vidal
LSI-IG,Universitat Politècnica de Catalunya,
Av. Diagonal ,647. Barcelona, Spain.08028.
lpv@lsi.upc.es

Fèlix Saltor
LSI-SI, Universitat Politècnica de Catalunya,
Jordi Girona 1-3. Barcelona, Spain.08034.
saltor@lsi.upc.es

**ABSTRACT**

In the growing field of Mobile Information Systems, there is a need for systems architectures that facilitate access to *heterogeneous information components* to guarantee the availability of information stored in databases. A mobile system should access different databases depending on user characteristics (e.g. GPS localization, user profile, user preferences). The semantic understanding of information, by means of *understanding the database schema*, will allow the extraction of the essential information for mobile applications. In this work we present a framework to detect the essential DB schema from heterogeneous spatial databases and GISs. The detection DB schema is used for a *federated architecture*. A model for assessing semantic similarities by means of ontologies, is applied to the mapping between the schema required for the application and the *essential schema*. The paper describes the components of the framework and proposes XML-based technology.

**KEY WORDS**

Mobile database, spatial database, semantic integration, ontologies, database schema mapping

## 1 Introduction

Semantic understanding is necessary to discover and extract the essential information from data sources into a structure suitable for the user's application. We believe that semantic understanding can be achieved if we focus on a specific domain. In this work we focus on Geographic Information Systems and Spatial Databases. Doing so, we limit the domain [6] and this will help us understanding related problems. The next delimitation adopted is by means of *user application characteristics* where there exist a profile and preferences that will imply the use of a specific domain-dependent ontology.

In a federated database architecture [14] there is an integration level, and at this level the export schemas are integrated in a federated schema [12]. In this paper we adhere to the approach to integration supported by federated architectures. In the framework presented here, we propose to reach a mapping between "any" database schema from a user's application to an essential spatial database schema in the context of mobile applications.

When a mobile application has to be developed, the process involves the design of a particular DB schema. However, in many situations, it is possible for the application to require access to different DBs with structures *unknown* to the application: A user application may ask for a DB schema that is necessary to achieve a specific task. We call *essential schema* the minimum DB schema required by an application to work properly. This means that the application should be able to gain access to main data stored in the DB. This paper investigates a framework based on Federated Architecture and standards as XML, XMI, GML and a model from OpenGIS (http://www.opengis.org) to facilitate the retrieval *an essential schema detection and mapping in a spatial DB for mobile applications*.

The organization of the remaining sections is as follows. Section 2 presents related work. Section 3 presents the operations within the spatial database bank directory to search the spatial DB capable of providing required data. Section 4 introduces the framework inspired by federated approaches for searching the essential DB schema. Section 5 shows how to construct the XMI schema from an application using a Canonical Data Model. Section 6 illustrates, through an example, the suitability of our approach and how the mapping is done. Conclusions and future research directions are given in the last section.

## 2 Related work

A recent survey about ontology mapping is presented in [4]. The ontology mapping has been addressed using different approaches: One-to-one approach, where a set of translating functions is provided for each ontology to allow the communication with the other ontologies without an intermediate ontology. In this approach the problem is the complexity degree for computing. With single-shared ontology the problems are similar to apply any standard.

In the database community, ontologies have been used in an attempt to reconcile the semantic and schematic perspectives. At present days, intelligent integration has been applied to heterogeneous database integration. In the artificial intelligence world, this is often achieved by means of agents or mediators that provide intermediary services by linking data resources and application programs. An architecture purposed by [7] is information-brokering: This adapts and extends the concepts of *federated environments* and *mediator architecture*. In the present work we try to follow this trend. In the domain of spatial information, different approaches for semantic integration were proposed (see, among others, [6, 13, 5]; and WFS-based ). [13] is based on a similarity-based analysis of concepts described in independent ontologies. Unlike OBSERVER [9], the solution does not create new ontologies, but creates links between similar entities across ontologies. Fonseca in [5] takes a top-down approach by starting from ontologies and using the concept of *role* to handle different conceptual views of Geospatial Information Communities (GIC). In our research, we profit from the power of ontologies for solving the semantic problem in the construction of the federated schema.

Export schemas are considered as a subset of definitions in a global ontology such that two concepts from different schemas can be related via a common super concept; this is similar to the treatment of different ontologies in [13] that are semantically interconnected.

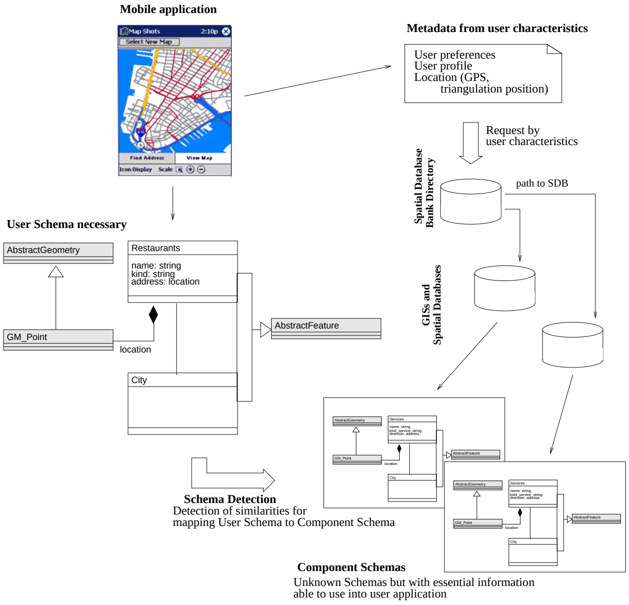

Figure 1. Scenario of a Mobile Application searching for information from Spatial Databases and GISs

## 3 Searching Spatial Database

Our work is concerned with the way to use databases whose data is available for access by some application but whose structure is not known by the application. Figure 1 shows the main scenario of our work. A user application can make a request of a special structure needed for its task. This structure can be represented as a spatial DB schema. Additional information for the application and the user is represented in preferences, profiles and other relative information.

We take as an example the application used for the searching of Restaurants. And from this example it is possible to infer consequences for a wide variety of applications. Among others searching possible infection focus, searching fire points, tracking transport shipping, tracking animal studies, tracking weather changes and so on.

1. **User preferences** is the user information about.

    - Language or idiom preferred by the application i.e. English, Spanish, French, and others.

    - Micro-domain, is the part of geographic information domain requested. This is named whit different names depending on where is the context in which it is defined. Many authors refer to this as *GICs according to their conceptualization of the world*. In Cyc (http://www.opencyc.org) , for example, those are defined as *microtheories*. The advantage in Cyc is the hierarchy of this microtheories. Then it is possible to know what is the “super” microtheory for another microtheory. ESRI divide the wide spectrum of GIS applications and attempts a classification by their use, it also gives Data Models.

      In our work, we take as micro-domains the following: Address, Historic Preservation and Archeology, Base map, Hydrology, Biodiversity, International Hydrographic, Organization (IHO) S-57, Census-Administrative Boundaries, Land Parcels, Defense-Intel, Local Government, Energy Utilities, Marine, Energy Utilities - MultiSpeak TM, Petroleum, Environmental Regulated Facilities, Pipeline, Forestry, Telecommunications, Geology, Transportation, Water Utilities.

    - Content Purpose, which is the main purpose of contained information.

    - Time Period, when any request for the application is valid.

    - Spatial Representation: data type i.e. vector, raster, and so on.; data format i.e. shapefile, coverage among others.

    - Spatial Domain information about Bounding Coordinates, or information with level of Country, City with translation to coordinates. Projection coordinates, and projection system.

    - Detail scale and Accuracy.

2. **User profile** defines characteristic user information. i.e. access level, security level, type of device, and so on.

3. **Intelligent track-selection** in both local device and server should save a ranking of previous click stream. This information could be applied in input of a Bayesian network to compute the probability of find the DB capable of supplies the request. The part referent to Bayesian network is out of the scope of the present work but for more information please refer to (http://www.lsi.upc.es/events/sitsd/).

We construct a *spatial database bank directory SDBBDir* taking information from metadata of available sources, spatial databases and GIS metadata. Metadata can be represented in standard formats (http://www.esri.com) : it will be used in the searching process [6]. Inside SDBBDir the major information on identification (name, description, purpose, version, location, *UTM coordinates*) is stored; quality (accuracy, completeness, currentness); lineage (sources, processing steps, previous versions); contact personnel; and, most important, the *path to the database server*. In SDBBDir the application searches for any DB capable of offering the information required (it has a slight similarity to services that are now in operation [1]).

Once the SDBBDir finds databases that satisfy the user characteristics, it returns to the application that requested the information, the path to different Spatial Databases or GIS servers. The search engine in the SDBBDir can be as complex as necessary. The process can be *personalized* in many ways. The SDBBDir system will keep statistical information on user preferences, profiles, and application access. Consequently, the SDBBDir system will be able to manage Data Mining techniques like “click stream” among others.

When the application knows which DB will be able to offer the information requested, it will search for the schema. Our main hypothesis is based on the actual trend to standardize the representation of Geographical Information with OpenGis specifications. Waiting for this standardization to be effective, the spatial databases and GISs should have a GML or XMI representation of their schema. Actually, many applications on (databases and GISs) have this representation. The next section deals with the derivation of the essential schema for GIS and spatial databases.

## 4 A Federated Approach, Hybrid Development

Our work is based in BLOOM architecture [1], and the approach in this work is a loosely coupled Federated Database System FDBS, due to the fact that the schema may be managed on the fly by a user. The federated schema can be quickly created and dropped. Inside of the development process of a FDBS there are two approaches: Bottom-Up and Top-Down. In the case of this work a *hybrid development process* (figure 2) is used, because the databases already exist (Bottom-Up) but it is necessary to support user requirements (Top-Down).

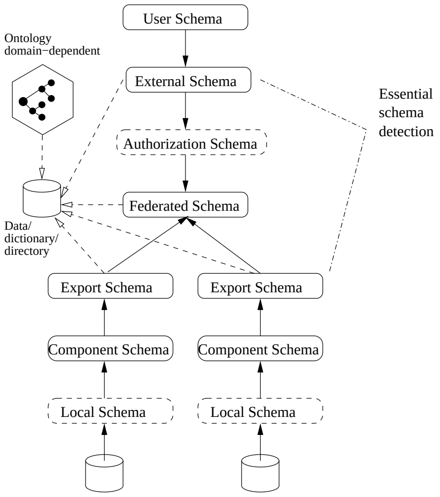

Figure 2. Hybrid FDBS development process. Top-down from the user application and bottom-up from the spatial databases

- Part I: Bottom-up development process
  - Translate schemas: Translate the local schema of the spatial database into the OpenGis abstract model. We consider that all spatial database components can be represented by or have already a GML or XMI interface; therefore they can be represented in an abstract model from OpengGIS (see section 5).
  - Define export schemas: The administrators (DBAs) of the respective component DBSs should authorize what part of their spatial DB will be available for access by the user application.
- Part II: Top-Down development process
  - Define a user application schema: Analyze requirements from user applications to define an external schema. The user application schema should be represented in GML or XMI too.
  - Detect essential schema: Analyze and compare the user application schema with possibles schemas to take necessary information (see section 3). Compare relevant federated schemas with the external schemas.
- Part III: Integrate schemas
  - Make the mapping between user application schema and DBs schemas.

[1] e.g, Geography Network http://www.geographynetwork.com

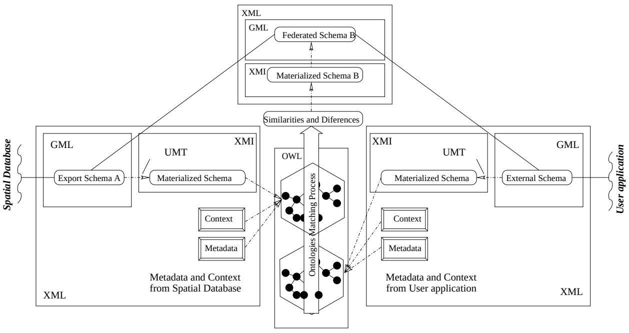

Figure 3. Creating Ontologies from GML Schemas

## 5 A Canonical Data Model for Spatial Databases

From the point of view of Federated Information Systems it is necessary to find a “Canonical Data Model” CDM, capable of representing all schemas with minimum loss of information from the Native Data Model. Some requirements for a geographic data model are presented in [3].

The possibility of OMT-G [3] and of the OpenGIS model as CDM is studied in [11]. In this work, we use the model from OpenGIS Consortium OGC as CDM. The OGC defines in its abstract specification a model good enough to allow for the representation of geographic data. This International Standard (also norm ISO/DIS 19107) specifies conceptual schemas for describing the spatial characteristics of geographic features, and a set of spatial operations consistent with these schemas.

OpenGIS has developed additional technology to deal with geographic information. The Geography Markup Language (GML) is an XML encoding for the transport and storage of geographic information. GML provides mechanisms for the encoding of geographic feature data without considering how the data may be presented to a human reader. From a GML representation of geographic data, it is possible to obtain the model in XMI representation.

### 5.1 XMI for Interchange models

XMI allows metadata (model and metamodel) to be interchanged as streams of files with a standard XML-based format. Some tools as ArcGIS 8.3 let us obtain the schema in XMI directly by means of an extension for Visio or Rational Rose. Those models are based on the abstract model form OpenGIS, therefore we can obtaining a materialized model capable of being parsed. From there, it is possible to match with the Ontology for each object of the model.

In the case of a GML interface, the UML Model Transformation Tool UMT (http://www.modelbased.net/umt/) or others, is capable of transforming some XMI into GML and the opposite. From here, we can make use of UMT as a tool for coding and decoding the GML model. It is possible with UMT to take the XMI model and transform it into GML, as well as the reverse step (to obtain the GML from XMI). In Fig. 3 we present a framework working with this technology.

All information, metadata, context and ontologies, can be extracted from XML-based files in order to use this in a process to assess similarities. We propose the use of OWL to express the ontology and to use the matching techniques [4, 13] for searching similarities and differences between the objects that have to be integrated.

## 6 Making the mapping between schemas

Ontologies provide significant benefits for the semantic description of data. Many research efforts, where ontologies are the main element to give semantic sense to data, have been reported in the literature. The creation and maintenance of a global and domain-independent ontology capable of supporting all knowledge is very complex. WordNet [10] and Cyc [8] are some of them. In contrast, constructing domain-dependent ontologies may be the solution to reach the best result in integration. But in the process of integration the main problem is to solve similarities, in [13] there is an approach for the use of ontologies to assess the similarities between entities.

We follow this initiative because we take an ontology that has the main geographic definitions together with their respective semantic sense. It could be constructed, for example, from WordNet or Cyc and SDTS.

We extract the names from the XMI model by means of a parser. The matching is divided in two phases:

- *Stage one*: Search of elements from both user application schema and DB schema on the global ontology.

  – A syntactic search for entity name.
  – A syntactic search for entity parts and attributes names. The ontology elements have parts like in WordNet [10].
  – A semantic search for entity metadata by means of keywords.
  – A semantic search for attributes metadata by means of keywords.

  This result is stored as a first matching reference. It is possible to give a weight for each search and decide if the entity has a corresponding element in the ontology.

- *Stage two*: Each object from the user application schema will search the corresponding object in the DB schema.

  – Assessment of semantic similarity. This is carried out with a similarity function like the computational model for semantic similarity used in [13]. Taking each entity name from the user schema and comparing against all entity names from the DB schema.

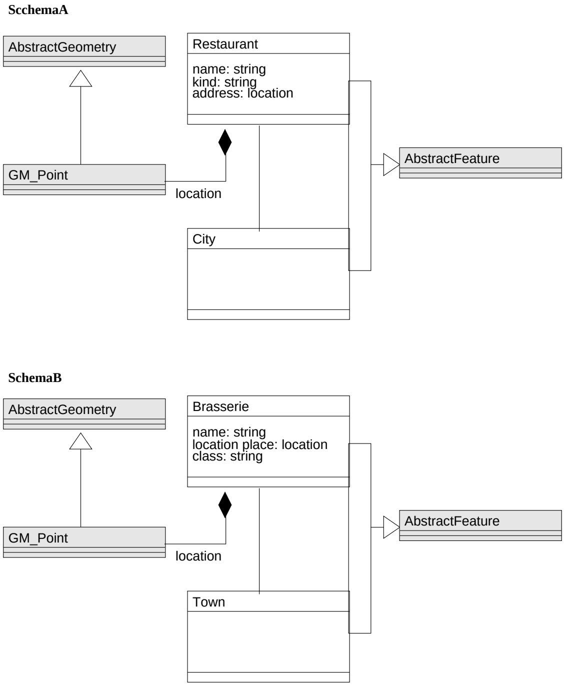

Figure 4. External Schema from a User Application *SchemaA* and Export Schema from Spatial Database *SchemaB*

- With elements whose similarity is accepted, make the syntactical and semantical comparison between the parts of the user schema entity and the parts of the DB schema entity. After this, similarity assessment between entities and attributes metadata. The assessment delivers a set of possible couples. Only the highest assessment will be accepted.
- Complete mapping information is stored in Data/Dictionary/Directory [14]

To show our approach, we introduce an example of an essential search of correspondence between an application and a spatial DB. Our scenario is a combination of two different schemas. In each one of them, there is a class that represents an “eating house”. In **SchemaA** there is a *Restaurant* and in **SchemaB** there is a *Brasserie*. Both of them have as property *Restaurant.address* and *Brasserie.location*, respectively. *Restaurant* has a relationship with *City* and *Brasserie* with *Town*. All of them have a geometry inherited Geometry feature from *AbstractGeometry*. In our example we try to obtain the mapping from SchemaA to SchemaB, Fig.4. We will consider that all the models belong to the same GIC; therefore, context information is the same.

*Stage one* is omitted because the entities names have correspondence with ontology elements. [13] presents a Matching-Distance Model to compare components of entity class in terms of a matching process. We can apply this model in our problem. First, for each entity name in **SchemaA**, we assess the semantic similarity against entity names in **SchemaB**. In this case between *Restaurants* against *Town* and *Brasseries*. For this assessment, we use WordNet Ontology, Fig.5, and apply equation 1. SDTS contains a set of entity types and their corresponding attributes. The WorkNet-SDTS Ontology takes synonym sets as well as hyponymy and meronymy relations from WordNet’s definitions to complement definitions of entity types in SDTS.

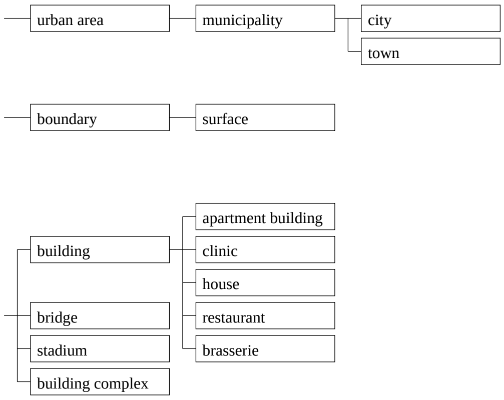

Figure 5. Partial ontology derived from WordNet

The global similarity function $S(c_1, c_2)$ is a weighted sum of the similarity values for parts, functions, and attributes; where $\omega_p, \omega_f$, and $\omega_a$ are the weights of the similarity values for parts, functions, and attributes, respectively. For each type of distinguishing features it uses a similarity function $S_t(c_1, c_2)$ (Equation 2). It is based on the ratio model of a feature-matching process [15]. In $S_t(c_1, c_2)$, $c_1$ and $c_2$ are two entity classes, $t$ symbolizes the type of features, and $C_1$ and $C_2$ are the respective sets of features of type $t$ for $c_1$ and $c_2$. The matching process determines the cardinality ($||$) of the set intersection ($C_1 \cap C_2$) and the set difference ($C_1 - C_2$), defined as the set of all elements that belong to $C_1$ but not to $C_2$. The function $\alpha$ is determined in terms of the distance between the entity classes $c_1$ and $c_2$ and the immediate superclass that subsumes both classes (or minimum common node m.c.n.). The minimum common node correspond to the least upper bound between two entity classes in partially ordered sets [2]. When one of the concepts is the superclass of the other, the former is also considered the m.c.n. The distance of each entity class to the m.c.n. is normalized by the total distance between the two classes, such that values in the range between 0 and 1 are obtained. The final value of $\alpha$ is defined by a symmetric function (Equation 3). For the complete description of equations refer to [13].

$$S(c_1, c_2) = \omega_p \cdot S_p(c_1, c_2) + \omega_f \cdot S_f(c_1, c_2) + \omega_a \cdot S_a(c_1, c_2) \quad (1)$$

$$S_t(c_1, c_2) = [|C_1 \cap C_2|]/[|C_1 \cap C_2| +$$

$$+\alpha(c_1,c_2)\cdot|C_1-C_2|+(1-\alpha(c_1,c_2))\cdot|C2-C1|] \tag{2}$$

$$\alpha(c_1,c_2) = \begin{cases} \frac{d(c_1,m.c.n.)}{d(c_1,c_2)} & d(c_1,m.c.n) \le d(c_2,m.c.n.) \\ 1-\frac{d(c_1,m.c.n.)}{d(c_1,c_2)} & d(c_1,m.c.n.) > d(c_2,m.c.n.) \end{cases} \tag{3}$$

We put a limit to accept similarity. If similarity is acceptable, *i.e, Restaurant, Brasserie*, then we can assess the syntactical and semantical similarity between attributes *name, kind, location* with *name, address, class*. We also perform the assessment with attribute and entity metadata. A data dictionary/directory [14] stores mappings between schemas and other essential information. There is in FGDC (http://www.fgdc.gov/metadata/metadata.html) large information about how metadata should be represented in a geographic systems. We assume that the sources to integrate keep that alignment, as i.e, ESRI with ArcCatalog . Therefore, it is possible to extract this information from sources to be used in entities and attributes metadata. A hierarchical structure is necessary for attribute types and for geometry types.

## 7 Future Work and Conclusions

We have presented in this work a framework capable of detecting the essential DB schema to be used in a mobile application, so that this application may be able to access “any” DB that makes its data available. This framework is based on federated architecture and we have used standards; mainly those based on XML. We have used ontologies and a model for assessing the semantic similarity between two objects. We have presented a new approach to enable access from mobile applications to heterogeneous spatial sources in a nearly automatic way. As future work, it is necessary to expand the framework to introduce security levels. And also, a multi-language approach to translate from several languages (i.e. English, Spanish, French and more), with a minimum loss of semantic information.

## Acknowledgments

Part of this work has been supported by: the Spanish Research Program PRONTIC under projects TIC2000-1723-C02-01; FEDER and MCYT under the project TIC2001-2099-C03-01 . Special acknowledgments to Mohand-Saïd Hacid (University Claude Bernard, France), to M.Andrea Rodríguez (University of Ceoncepción, Chile) for their valuable contributions for this paper.

## References

[1] A. Abelló, M. Oliva, E. Rodríguez, and F. Saltor. The bloom model revisited:an evolution proposal. In *Proceedings ECOOP’99 Workshops & Posters*, Lisbon, Jun 1999.

[2] G. Birkhoff. *Lattice Theory*. American Mathematical Society, 3rd edition edition, 1979.

[3] K. A. Borges, C. A. Davis, and A. H. Laender. OMT-G: An object-oriented data model for geographic applications. *GeoInformatica*, 5(3):221–260, Sep 2001.

[4] Y. Ding and S. Foo. Ontology research and development part 2 - a review of ontology mapping and evolving. *Journal of Information Science*, 28(5):375–388, 2002.

[5] F. T. Fonseca. *Ontology-Driven Geographic Information*. PhD thesis, University of Maine, Orono, Maine 04469, May 2001.

[6] M. Goodchild, M. Egenhofer, R. Fegeas, and C. Kottman, editors. *Interoperating Geographic Information Systems*. Kluwer, 1999.

[7] V. Kashyap and A. Sheth. Semantics based information brokering. In *Proceedings of the 3rd International Conference on Information and Knowledge Systems*, pages 363–370, 1994.

[8] D. Lenat and R. Guha. *Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project*. Reading, Mass,Addison-Wesley, 1990.

[9] E. Mena, V. Kashyap, A. Illarramendi, and A. Sheth. Domain specific ontologies for semantic information brokering on the global information infrastructure. In N. Guarino, editor, *Formal Ontology in Information Systems*. IOS press, 1998.

[10] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[11] V. Morocho, F. Saltor, and L. Pérez-Vidal. Ontologies: Solving semantic heterogeneity in federated spatial database system. In *Proceedings of 5th International Conference on Enterprise Information System*, pages 347–352, Angers, France, Apr 2003.

[12] V. Morocho, F. Saltor, and L. Pérez-Vidal. Schema integration on federated spatial db across ontologies. In *Proceedings of the 5th International Workshop on Engineering of Federated Information Systems EFIS*, pages 63–72, Coventry, UK, Jul 2003.

[13] M. A. Rodríguez and M. J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456, Mar 2003.

[14] Sheth and Larson. Federated database systems for managing distributed heterogeneous and autonomous databases. *ACM Computing Surveys*, 22(3), 1990.

[15] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.