# Semantic Integration on Spatial Databases
## SIT-SD prototype

Villie Morocho[1], Lluís Pérez-Vidal[2], and Fèlix Saltor[1]

[1] Departament de LSI-SI, Universitat Politècnica de Catalunya,
Jordi Girona 1-3,08034. Barcelona, Spain.
{vmorocho,saltor}@lsi.upc.es
[2] Departament de LSI-IG, Universitat Politècnica de Catalunya,
Av. Diagonal ,647 08028. Barcelona, Spain.
lpv@lsi.upc.es

**Abstract.** Progress on data integration has come to the point that we must deal with the need of semantic integration. In this work we present a prototype of an integration tool of information sources (Semantic Integration Tool for Spatial Data). It focuses on schema integration of spatial databases. This tool is able to recognize the similarities and the differences between entities to be integrated. It is based on a Federated Architecture and, therefore, it is a leveled architecture. We obtain the XMI models from metadata of spatial sources that will be integrated. A parser takes the entities and attributes from XMI models by means of a semantic match process and we assess the similarities and differences between the objects. A domain-dependent ontology is created from the FGDC standard (Federal Geographic Data Committee) and domain-independent ontologies (Cyc and Wordnet). This ontology is represented in OWL. We make use of a ratio model (ontology nodes distance) for the assessment of the similarities and differences between terms. The prototype takes advantage of standardization in spatial data tools. For instance, ESRI or others that incorporate a lot of standards into their applications. We also make use for this work of the latest technology: XML, XMI, GML, OWL, and others.

## 1 Introduction

Geographic applications are an example of the need to bring data integration to a big scale. This is the case for the studies of weather, environment, sustained development, terrain use (ground use), mobile applications and more. When we attempt to integrate this kind of data we find that there are some advantages (many developed and applied standards) and some disadvantages (various types of geographic data). Semantic understanding is necessary to discover and extract the essential information into a structure suitable for integration from the sources of data. Researchers show the need to focus on a specific domain to achieve the main goal of semantic understanding [21]. In this work we focus on the domain of Geographic Information Systems and Spatial Databases. We first begin to work with shape data and thus we can narrow the domain even further.

Due to the large quantity and broad range of types of Geographic data it would be impossible to integrate on an application without losing some information. In the search

for the best solution for this integration we have taken advantage of the standards that have been actually accepted by the main spatial software companies. As an example we take the case of the OpenGIS Consortium standard for representation of geographic features. ESRI follows this standard for their representation of features. On the other hand, the FGDC (Federal Geographic Data Committee) has approved many content standards for geographic applications. An example is their standard for representing metadata. With the above hypothesis, we have developed a prototype to integrate heterogeneous spatial data sources. This first phase in the prototype tries to integrate simple features. This term is taken from OGC (OpenGIS Consortium) which geometric properties are restricted to *simple geometries*. For these simple geometries coordinates are defined in two dimensions and the delineation of a curve is subject to linear interpolation [18]. This work is the continuation of [16] which presented a framework that was based on a federated architecture for schema integration.

The evolution of federated database systems has identified two approaches to managing distributed data: installing a distributed DBMS and adding a software layer above existing DBMSs to create an FDBS. This evolution process is divided into three phases [20] : (1) *preintegration*, (2) *developing a federated database system*, and (3) *federated database operation*. Section 3 refers to the development of a federated database system where we define the mappings between various schemas. In this work we focus only in the second phase where we achieve a federated schema from two heterogeneous spatial sources. SIT-SD is a prototype capable of implementing the above theory developed in our work group. We take spatial data worked in ArcGIS [1] and from these files we extract the model in XMI. A parser identifies the main objects as entities and attributes and subsequently submits them to the process for assessment of similarities. By means of a Java program and Cyc, we take names of entities and attributes from the FGDC standard for generation on OWL from one part of the ontology that will be used in the assessment process.

The organization of the remaining sections is as follows. Section 2 presents related work. Section 3 follows with the proposed architecture which we use technology XML based and OpenGIS and FGDC standards. Section 3.1 introduces the similarity-based strategy for schema integration. Conclusions and future research directions are given in the last section.

## 2 Related work

Ontologies provide significant benefits for the design and use of geographic information. Ontologies define semantics independently of data representation and reflect the relevance of data without accessing them [9]. Such a high-level description of the semantics of geographic information provides more and new means for comparing and integrating spatial data. In addition, ontologies enable knowledge reuse by semantically describing data that were derived from consensus reached by different GIS communities [24].

In the database community, the ontologies has been used in an attempt to reconcile the semantic and schematic perspectives. Kashyap and Sheth in [11] present a semantic

[1] All commercial marks are registered trademarks by owners

taxonomy to demonstrate semantic similarities between two objects and related this to a structural taxonomy. At present days, intelligent integration has been applied to heterogeneous database integration. From artificial intelligence world often it is achieve by means agents [5] or mediators [24] that provide intermediary services by linking data resources and application programs. Otherwise, from databases world has been purposed an architecture named information-brokering ([12],[10]) that adapts and extends the concepts of *federated environments* and *mediator architecture*, in the present work we follow this trend. In the domain of spatial information, there are research approaches

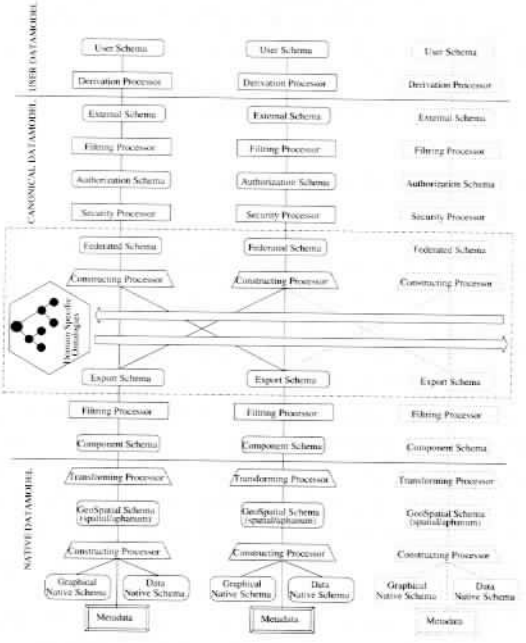


**Fig. 1.** Federated Architecture combined with mediators. Ontologies for solving semantic heterogeneity

for semantic integration [19], [8], [7]. [19] is based on a similarity analysis of concepts described in independent ontologies. Unlike OBSERVER [14], the solution does not create new ontologies, but creates links between similar entities across ontologies. Fonseca in [8] 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 first construct a framework with a federated architecture, and in the construction process of schema federated we take a domain-dependent ontology for finding similarities among entities and attributes to be integrated.

## 3 The SIT-SD Architecture

Here we present the Semantic Integration Tool for Spatial Data *SIT-SD*. We assume that the preintegration process exists before and it has decided what are the sources to be integrated. Fig.1 describes the architecture based on [1]. It is divided in seven different levels from the sources to be integrated until user application. The architecture is a bottom-up process because we consider that the data sources exist before. This work is focused only on the process of schema integration on this architecture (dashed line part in Fig. 1).

From the view point of Federated Information Systems it is necessary to find a "Canonical Data Model" CDM, capable of representing all schemas with a minimum loss of information from the Native Data Model(bottom level). In [16], we studied the possibility of using OMT-G [3] and OpenGIS models as CDM. We use the model from OpenGIS Consortium OGC [17] as CDM. ESRI with ArcGIS follows the OpenGIS specifications, and this reason we use ArcMap in this prototype.

### 3.1 Constructing the federated schema

Ontologies provide significant benefits in semantically describing data. Many researches, where ontologies are the main element to derive semantic sense from 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 [15] and Cyc [13] are some of them. In contrast, the construction of domain-dependent ontologies [25, 4] may be the solution for the best result in integration. In our case we construct a domain-dependent ontology taking entities from a very known standard like SDTS [23] (Spatial Data Transfer Standard) and applying functions from a domain-independent ontology as Cyc and WordNet. We chose this standard because we believe that it is the most used. This ontology is represented with OWL specification.

In [19] there is an approach for the use of ontologies to assess the similarities among entities. In this work we apply this foundation presented and extrapolate to apply over database schemas.

From an XML representation of data structure and database structure we extract the elements and relations for constructing the schemas on XMI (see example). The objects are extracted from the XMI model by means of a parser.

Later the matching process is applied which is divided into three phases (see Fig.2): Consider two schemas to be integrated: schemaA and schemaB from two different spatial databases.

- *Stage zero*: Translation to English. In this first version it makes a simple syntactic translation from different languages (we choose the English language as a standard for translation because we believe the translation process is easier). The metadata markup is taken because it indicates what language is used in the application. In the next version we will improve the translation process by maybe using a semantic search over a dictionary or thesaurus.
- *Stage one*: Search of elements on the SIT-SD ontology of two database schemas to be integrated.

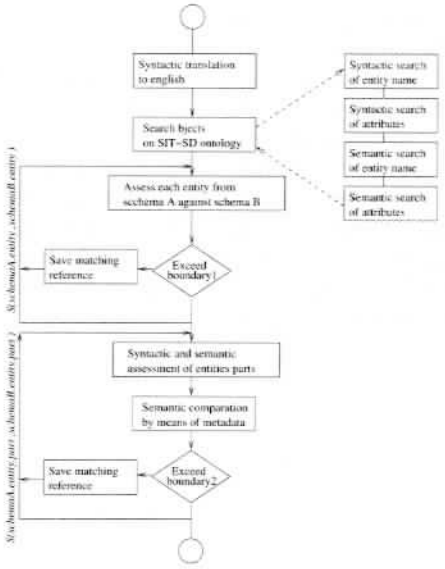


**Fig. 2.** Flowchart of schema integration between two export schemas in a Spatial FDB

- A syntactic search of entity name.
- A syntactic search of entity parts and attribute names. The ontology elements have parts like in WordNet. [15].
- A semantic search of entity metadata by means of keywords.
- A semantic search of 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 schemaA will search the corresponding object in schemaB.
  - Assessment of semantic similarity. This is carried out with a similarity function like the computational model for semantic similarity used in [19]. Taking each entity name from schemaA and comparing against all entity names from schemaB.
  - With elements whose similarity is accepted, we make the syntactic and semantic comparison against the parts of schemaA entity and the parts of schemaB entity. Afterwards, the similarity assessment is applied between entities and attributes metadata. The assessment delivers a set of possible pairs. Only the highest assessment will be accepted.
  - Complete mapping information is stored in Data/Dictionary/Directory [20]

The mapping is carried out with a similarity function that uses string matching of object names and object interrelations. Once this mapping has been done, an integrated schema should contain both original objects in local schemas as subclasses of a more

general common class. This common super class is determined by searching through semantic relationships in the SIT-SD ontology.

To show our approach, we introduce an example of integration of spatial data from different spatial databases and GIS. Our scenario is a compound of two different schemas from spatial databases with geographic information. In both schemas there is a class in the Export Schema that represents a "green zone". In **schemaA** there is a *fondo* class and in **SchemaB** there is a *background* class. Both classes have *fondo.codigo*, *fondo.area*, and *background.code background.surface* as properties respectively. *fondo* has a relationship with *ciudad* and *background* with *city*. All classes inherit geometry (Geometry Feature) from the *feature class*. In our example we try to obtain a federated schema **schemaC**, like that in Fig. 3.

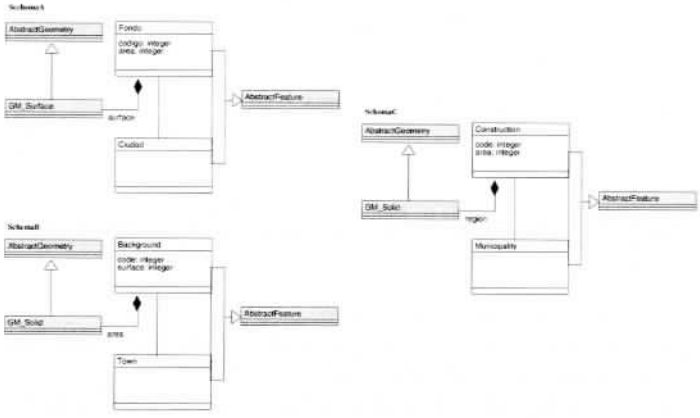


**Fig. 3.** Federated Schema *schemaC* from Export Schema *schemaA*, *schemaB*

On the architecture, in the construction level of federated schema, we extract the entities and attributes form XML files on ArcGIS.

```
<spdoinfo>
  <direct Sync="TRUE">Vector</direct>
  <ptvctinf>
    <esriterm Name="fondo">
      <efeatyp Sync="TRUE">Simple</efeatyp>
      <efeageom Sync="TRUE">Polygon</efeageom>
      <esritopo Sync="TRUE">FALSE</esritopo>
      <efeacnt Sync="TRUE">754</efeacnt>
      <spindex Sync="TRUE">TRUE</spindex>
      <linrefer Sync="TRUE">FALSE</linrefer>
    </esriterm>
    <sdtsterm Name="fondo">
      <sdtstype Sync="TRUE">G-polygon</sdtstype>
      <ptvctcnt Sync="TRUE">754</ptvctcnt>
    </sdtsterm>
  </ptvctinf>
</spdoinfo>
```

Above there is the code part of SDTS description for the *fondo* feature. The description for the *background* feature is the same way. 3.1 Below we display the description of the *fondo* attributes.

```
<eainfo>
  <detailed Name="fondo">
  <enttyp>
  <enttypl Sync="TRUE">fondo</enttypl>
  <enttypt Sync="TRUE">Feature Class</enttypt>
  <enttypc Sync="TRUE">754</enttypc>
  </enttyp>
  ...
  <attr>
  <attrlabl Sync="TRUE">AREA</attrlabl>
  <attalias Sync="TRUE">AREA</attalias>
  <attrtype Sync="TRUE">Number</attrtype>
  <attwidth Sync="TRUE">18</attwidth>
  <atnumdec Sync="TRUE">5</atnumdec>
  </attr>
  ...
  <attr>
  <attrlabl Sync="TRUE">CODIGO</attrlabl>
  <attalias Sync="TRUE">CODIGO</attalias>
  <attrtype Sync="TRUE">Number</attrtype>
  <attwidth Sync="TRUE">18</attwidth>
  </attr>
...
```

By using this data and applying a Java script we can construct a simple XMI model taking the entities and attributes names. Then we can apply the **stage zero** depicted above. In this case it translates to English with a simple dictionary and saves this translation for application of the next stages. Before obtaining the XMI model we constructed a domain-specific ontology derived from SDTS entities and applying the Cyc and Wordnet ontology functions. SDTS was adopted by the American National Standard Institute to provide a common classification and definition of spatial features used in the process of spatial data transfer. It contains a set of entity types and their corresponding attributes. The specific-domain Ontology takes synonym sets as well as hyponymy and meronymy relations from WordNet's and Cyc's definitions to complement definitions of entity types in SDTS. For this example we will consider that all the models belong to the same GIC, therefore the ontology is created over one specific domain.

We apply the **stage one** depicted above in order to search the objects inside of the ontology. Then, we can apply a matching-distance model [19], **stage two**, to compare components of entity classes in terms of a matching process. To obtain the objects for **schemaC** we assess the semantic similarity between two objects to integrate. In this case between *backdrop* and *code*. *fondo* was translated to its corresponding English word *backdrop*, and *codigo* was translated to *code*. For this assessment we use specific-domain ontology, Fig.4, and apply the Equation 1.

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 weights of the similarity values for parts, functions, and attributes, respectively. For each type of distinguishing feature 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 [22]. 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 corresponds 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 it obtains values in the range between 0 and 1. The final value of $\alpha$ is defined by a symmetric function (Equation 3). For the complete description of equations refer to [19].

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

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

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

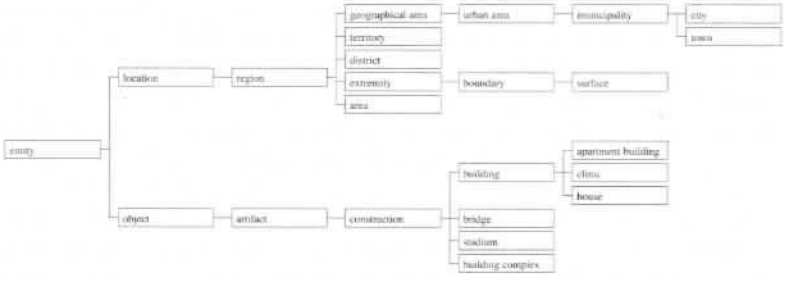


**Fig. 4.** Domain-specific ontology (geographic) derived from WordNet and SDTS

We put a limit to accept the similarity. If the similarity is acceptable then we can search the *minimum common node* in the ontology hierarchy. In the case of our example the m.c.n is *background*; this is then the object in **schemaC**. Likewise, it is possible to compare each attribute from both classes. A data dictionary/directory [20] stores mappings among schemas and another essential information. The additional information from Context and Metadata will help us to determine what is the sense of each object in the process of assessing. In a FGDC site (Federal Geographic Data Committee) there is large information about how metadata should be represented in geographic systems. We assume that the sources to be integrated keep this line as i.e, ESRI with ArcCatalog [6]. Thereby, it is possible to extract this information to be used in the entities and attributes metadata. A hierarchical structure is necessary for the attribute types and the geometry types, Fig. 5.

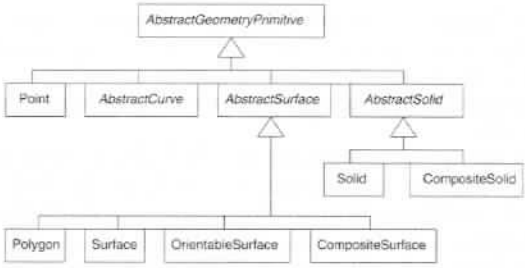


**Fig. 5.** Type Hierarchy of geometry types OpenGIS

## 4 Future Work and Conclusions

This paper has presented a SIT-SD (Semantic Integration Tool for Spatial Data). This is a prototype work that continues in progress. This prototype is a result of research and development in the research group where we are working on semantic integration for spatial data. We take advantage of the latest technology such as XML, XMI, OWL, and standards like OpenGIS and FGDC. These standards have allowed this first approach on semantic integration. This will be agreed upon by many companies and government offices on spatial data able to support standardization. Progress in the main tool (ArcGIS) used in this prototype for the access to GIS information should change the next version. The generation of the XMI model from the next ESRI application version will be able to represent the geodatabase on XML directly. This change will make the XMI model easier to generate. This simple prototype has a principal goal to demonstrate the theory developed above. The next version should work with more complex data because this present prototype only works with simple features. In the next version we will improve the translation process by using a semantic search over a dictionary or thesaurus.

### Acknowledgments

Part of this work has been supported by: the Spanish Research Program PRONTIC under projects TIC2000-1723-C02-01 and TIC2001-2099-C03-01, and by FEDER. Special acknowledgments to M.Andrea Rodríguez (Department of Computer Science, University of Concepción, Chile) for her valuable contribution to 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. Garrett Birkhoff. *Lattice Theory*. American Mathematical Society, 3rd edition edition, 1979.
3. Karla A.V. Borges, Clodoveu A. Davis, and Alberto H.F. Laender. Omt-g: An object-oriented data model for geographic applications. *GeoInformatica*, 5(3):221–260, Sep 2001.
4. W. Borst, J. Akkermans, and J. Top. Engineering ontolofies. *International Journal of Human-Computers Studies. Special Issue on Using Explicit Ontologies in KBS Development*, (46):365–406, 1997.

5. M. Brodie. The promise of distributed computing and the challenges of legacy information systems. In *Proceedings of IFIP DS-5 semantics of interoperable database systems*, Lorne, Australia, 1992. Chapman & Hall.
6. ESRI. Esri profile of the content standard for digital geospatial metadata. Technical paper, ESRI, Jul 2001.
7. F. Fonseca, M. Egenhofer, P. Agouris, and C. Câmara. Using ontologies for integrated geographic information systems. *Transactions in GIS*, 6(3), Jun 2002.
8. Frederico Torres Fonseca. *Ontology-Driven Geographic Information*. PhD thesis, University of Maine, Orono, Maine 04469, May 2001.
9. A. Goni, E. Mena, and A. Illarramendi. Querying heterogeneous and distributed data repositories using ontologies. In P.-J. Charrel and H. Jaakkola, editors, *Information Modelling and Knowledge Base IX*, pages 19–34. IOS Press, 1997.
10. 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.
11. V. Kashyap and A. Sheth. Schematic and semantic similarities between database objects: a context-based approach. *The Very Large Database Journal*, 5(4):276–304, 1996.
12. Vipul Kashyap. *Information Brokering over Heterogeneous Digital Data: A Metadata Based Approach*. PhD thesis, Rutger University, 1997.
13. D. Lenat and R. Guha. *Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project*. Reading, Mass.Addison-Wesley, 1990.
14. E. Mena, V. Kashyap, A. Illarramendi, and A. Sheth. Domain specific ontologies for semantic information brokering on the global information infrastructure. In Nicola Guarino, editor, *Formal Ontology in Information Systems*. IOS press, 1998.
15. 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.
16. Villie Morocho, Fèlix Saltor, and Lluís Pérez-Vidal. Ontologies: Solving semantic heterogeneity in federated spatial database system. In *Proceedings of 5th International Conference on Enterprise Information System*, Angers, France, Apr 2003.
17. OpenGIS. The opengis abstract specification. topic 0: Abstract specification overview. Retrieved May 2001, from http://www.opengis.org/techno/abstract/99-100r1.pdf, 1999.
18. OpenGIS. Geography markup language (GML) implementation specification v2.1.2. Retrieved from http://www.opengis.org/, Sep 2002.
19. María Andrea Rodríguez and Max J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 5(2), 2003.
20. Sheth and Larson. Federated database systems for managing distributed heterogeneous and autonomous databases. *ACM Computing Surveys*, 22(3), 1990.
21. Amit P. Sheth. *Interoperating Geographic Information Systems*, chapter Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics, pages 5–29. Kluwer Academic Publisher, 1999.
22. Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.
23. USGS. View of the spatial data transfer standard (sdts). Retrieved May 2001, from http://mcmcweb.er.usgs.gov/sdts/standard.html, 1998.
24. G. Wiederhold. Interoperation, mediation and ontologies. In *International Symposium on Fifth Generation Computer Systems (FGCS94)*, Tokyo, Japan, 1994.
25. P. Zweigenbaum, B. Bachimont, J. Bouaud, J. Charlet, and J.F. Boisvieux. Issues in the structuring and acquisition of an ontology for medical language understanding. *Methods of Information in Medicine*, 34(1/2):15–24, 1995.