Schema Matching Based on Attribute Values and Background Ontology

Abadie Nathalie
Institut Géographique National, Laboratoire COGIT
2 Avenue Pasteur, 94160 Saint-Mandé, France
Université de Paris Est
Cité Descartes, Champs-Sur-Marne, 77454 Marne-La-Vallée cedex 2, France
nathalie-f.abadie@ign.fr

ABSTRACT

This paper focuses on the specific problem of geographic database schema matching, as a first step in an integration application. We propose, a schema matching approach based on attributes values and background ontology. We follow the intuition that comparing only schema classes is not sufficient and that there are specific class attributes in geographic database classes whose role consists in specifying the exact nature of each class instance. Their enumerated values refer to geographic concepts. We assume that it is possible to take advantage of this additional knowledge for upgrading the level of granularity of schema classifications by making it explicit in local ontologies created from each database schema that we have to match. Moreover, we propose to use external domain knowledge, namely background ontology, to improve our schema matching strategy, which implies local ontology matching. We lastly present and discus the results of two schema matching tests based on this schema matching strategy.

1. INTRODUCTION

Nowadays, technologies such as positioning systems and Internet make it easier to produce and access to geographic information. In recent years this has resulted in an increasing availability of diverse, heterogeneous and distributed geographic data sources containing useful and complementary information for GIS applications. However, even if these data represent the same topographical real world, there is a great heterogeneity between them caused by their "design autonomy" (Sheth and Larson, 1990). Consequently, sharing and integrating information from such heterogeneous sources is not a straightforward task. To fully achieve information integration, several aspects of data heterogeneity must be solved (Bishr, 1998). Systems must be able not only to exchange data, but also to understand the meaning of interchanged data. This latter aspect, known as semantic interoperability, is a key issue for successful geo-information integration.

Schema matching poses challenges in many fields of research, such as schema integration, data integration, data warehousing, or catalogue integration. It consists in identifying schemas entities that are semantically related (i.e. schema entities that represent the same real world features) (Rahm and Bernstein, 2001). The use of ontologies as tools for specifying the exact meaning of terms within a community has been acknowledged as a valid approach to overcome the problem of semantic heterogeneity (Hakimpour and Timpf, 2001, Partridge, 2002). In a classical general information integration scenario, the data sources are wrapped to local ontologies which are matched against a common ontology (Euzenat and Shvaiko, 2007).

This paper focuses on the specific problem of geographic database schema matching, as a first step in an integration application. We follow the intuition that comparing only schema classes is not sufficient and that there are hidden geographic concepts in schemas. We propose to use this additional information for upgrading the level of granularity of schema semantics in an ontology-based schema

matching process. Lastly we discuss the results of two schema matching strategies based on the use of these hidden concepts: the former is a simple lexical process, the latter relies on background ontology.

This paper is structured as follows. Section 2 introduces the problem of semantic heterogeneity for schema matching. Section 3 describes our schema matching strategy based on hidden geographic concepts. Section 4 presents practical tests on two geographic database schemas that we carried out respectively for each of the proposed strategies. In this section, we also discuss the results of our tests.

2. SEMANTIC HETEROGENEITY IN GEOGRAPHIC DATABASES

An important issue for schema matching is the ability to understand what the differences between the databases are. For that purpose, one first requires to understand what each database exactly contains, i.e. their semantics. Like for any database, the content of geographic databases is first described by their schema.

Geographic database schemas result from an abstraction of the real geographic world into an object-oriented model (Fonseca et al., 2003). Classes are named by common geographic words, which usually refer to geographic concepts. Geographic features are represented by class instances and are described by attributes and a geometrical representation (point, line or polygon).

However, for each data producer, there exists a precise meaning beyond the words used for naming schema entities. Indeed, different communities can have different points of view about the same real world feature, or they can use different terms or labels to name equivalent geographic concepts. Moreover, depending on their application domain, geographic databases are associated to a certain level of detail, and only the most relevant geographic features are captured. Consequently, if a class is named *River*, it may designate only permanent rivers in a database, or only natural rivers in another one. Besides, the name *river* may designate only rivers that are wider than 10 meters.

Furthermore, it is a common modeling practice for geographic databases to simplify the schema structure by merging semantically close classes into a single class. In such cases, the specific nature of each instance of the class is described more accurately by an attribute (usually named “nature” or “type”). Most of the time, this attribute’s values are terms that designate geographic concepts. As a result, the meaning of the class can be understood, not only thanks to the class’s name, but also by reading this attribute’s enumerated values.

Thus the nature of geographic features captured in each class is not only designated by their class’s name. In some cases, attributes values can give additional information on the exact nature of the geographic features stored in the database.

3. GEOGRAPHIC SCHEMA MATCHING STRATEGIES

The main idea of our schema matching strategy is to use geographic concept labels stored in some attribute values for making the content of each class explicit. A first technical step in our schema matching scenario consists in building a local ontology for each schema. Then, these ontologies are aligned. The correspondences provided by this alignment step are lastly used to find mappings between schema entities.

3.1. Making hidden geographic concepts explicit in local ontologies

This step of building a local ontology from a geographic database schema is fairly intuitive. Each schema class represents an object-oriented abstraction of real world features. Therefore, in our local ontology, each database class will be translated into a concept, whose label corresponds to the class

name. Concept properties and relations are straightly derived from their respective class attributes and associations. Our local ontology is structured following a subsumption hierarchy. Specific concepts are linked to their more general concepts by *isA* relations. Thus schemas inheritance relations between classes are translated into *isA* relations between their associated concepts.

Besides, the main purpose of this step is to clarify the meaning of each database class: what kinds of geographic features are represented by each class? To reach this goal, we will use the specific class attributes which define the exact nature of each class instance, called "*determining properties*" by (Manoah et al., 2004). Most of the time, these attribute enumerated values refer to geographic concepts and represent a rich information source about the content of the class, which is usually hidden in the database structure. In our local ontology, such attribute values will therefore be translated into concepts, subsumed by the concept associated with their respective class. This aims at making classes semantics more explicit for the system by extending the schema classification's level of granularity in our local ontology. We assume that these hidden concepts can be successfully exploited after the ontology matching step to detect schema correspondences.

3.2. Matching local ontologies

Ontology matching, also named ontology alignment, like schema matching, aims at finding correspondences between semantically related named entities, such as classes, properties, individuals, or more complex expressions such as definitions, from different ontologies. These correspondences, also called mappings, can be of several types: equivalence, subsumption, consequence, etc.

Many ontology matching approaches have been developed (Kokla, 2006, Euzenat and Shvaiko, 2007). They mainly focus on two aspects: lexical matching and structural matching. Ontology lexical matching uses string-based and linguistic techniques to compare ontology elements' labels and detect correspondences. Ontology structural matching uses the structural relations between elements within the ontologies to compute similarities.

3.2.1. Lexical and structural matching approach

As a first approach, we propose to use a basic lexical technique to match our local ontologies. We follow an approach proposed by (Hamdi et al., 2008). This matching process is oriented. It aims at finding mappings from source ontology (O_{source}) to target ontology (O_{target}) between single concepts from these ontologies. When mappings between two concepts C_{source} and C_{target} are established, the type of relationship between these concepts is explicitly stated. This approach aims at detecting three types of relationship between concepts: equivalence relationships (isEq), subclass relationships (isA or isGeneral), and semantically related relationships (isClose).

The method to extract mappings between a concept C_{source} in O_{source} and a concept C_{target} in O_{target} is based on a substring similarity between the labels of C_{source} and Ct_{arget} respectively:

S represents the set of strings. Let c_s and c_t be C_{source} and C_{target} labels. Let x be the longest substring of c_s and c_t. The substring similarity is the following:

$$s(ct, cs) = \frac{2 * |x|}{|cs| + |ct|}$$

Then, according to each pair of concept similarity values, relationships between these concepts can be established. Pairs of concepts with similarity values higher than a given threshold are considered equivalent. Other types of relationships are established depending on several criteria:

- Label inclusion: Let c_t be the label of O_{target} with the highest similarity score with c_s. If c_t is included in c_s, then we assume that c_s designates a more precise concept than c_t, and a (C_{source} isA C_{target}) relationship is generated. Inversely, if c_s is included in c_t, a (C_{source} isGeneral C_{target}) relationship is created.

- Relative similarity: Let c_{tmax} and c_{t2} be the two labels with the highest similarity measure with c_s. If the ratio of c_{t2} similarity value on c_{tmax} similarity value, called relative similarity, is lower than a given threshold, then:
    - A (C_{source} isClose C_{tMax}) relationship is generated if the similarity value of c_{tMax} is greater than a given threshold and if c_s is included in c_{tMax};
    - A (C_{source} isClose C_{tMax}) relationship is generated if the similarity value of c_{tMax} is greater than a given threshold;
    - An isA relationship is created between C_{source} and the father of C_{tMax} if the similarity value of C_{tMax} is greater than a given threshold;
- Structure: Let c_{tMax}, c_{t2} and c_{t3} be the three labels of O_{target} with the highest similarity measure with c_s. An isA relationship is generated if they have similarity values greater than a given threshold and if they also have a common father.

3.2.2. Background knowledge approach

In the approach presented above, the existence of either lexical or structural overlap between ontologies is needed for matching concepts. In the cases where ontologies have different terminologies and different structures, correspondences between elements that are semantically related can not be found. To overcome this problem, (Aleksovski et al., 2006) propose an approach to match two ontologies using a third comprehensive domain ontology as background knowledge. This background ontology is used to compensate the lack of lexical or structural similarity between the two ontologies that we want to align.

This approach works in two steps: anchoring and deriving relations. The former consists in matching the local ontologies against the background ontology. This can be done by using classical lexical and structural ontology matching techniques. The latter aims at discovering relationships between source and target concepts, by looking for relationships between their anchored concepts in the background ontology. Combining the types of relationships detected during the anchoring step between source or target concepts and their anchored concepts with the structural relationships between these anchored concepts within the background ontology enables us to derive relationships between source and target concepts.

In our approach, the anchoring step is processed with the lexical and structural matching algorithm presented above. Local ontologies built from databases schemas are matched against a more detailed taxonomy of geographic concepts. The alignments resulting from this step are then used to derive relationships between local ontology concepts.

Relationships between local ontologies are computed by analyzing their mappings with the background taxonomy, and the relationships between their anchored concepts within this taxonomy:

- If two concepts C_{target} and C_{source} from local ontologies are matched with the same background taxonomy concept C, and if C_{target} has an *isEq* relationship with C, then a relationship is generated between C_{source} and C_{target}. The type of this discovered relationship is the same as the relationship type between C_{source} and C.
- If two concepts C_{target} and C_{source} from local ontologies are matched with the same background taxonomy concept C, and if C_{target} and C_{source} have the same type of relationship with C, which is not an *isEq* relationship, then an *isClose* relationship is generated between C_{source} and C_{target}.
- If two concepts C_{target} and C_{source} from local ontologies are matched with the different background taxonomy concepts, respectively C_1 and C_2, with an isEq relationship, then we

check for relationships between C_1 and C_2. If C_2 subsumes C_1, then a (C_{target} isA C_{source}) relationship is generated. Else, if C_1 subsumes C_2, then a (C_{target} isGeneral C_{source}) relationship is generated.

3.3. Matching schemas on the basis of ontologies alignments

Once local ontologies have been matched, alignments are used to match schema elements. As a first approach, we will manage all our mappings as equivalence mappings. IsA, isGeneral and isClose relationships are thus regarded as equivalences in the schema matching process. However, they still remain, like the string similarity score, a good information source about the reliability of detected schema correspondences. Actually, schema correspondence computed on the basis of an isEq relationship ontology mapping is more likely to be valid than one computed from an isClose relationship mapping.

For each mapping between local ontologies, original schema entities from which the matched concepts are derived are analyzed:

- If both original schema entities are classes, then a mapping between these classes is generated. This means that we assume that these classes have the same meaning.

- If both original schema entities are attribute values (our hidden concepts sources), then a mapping between these attribute values is generated. This means that we assume that instances from both classes which are proved to have these specific attribute values have the same meaning.

- If one schema entity is a class, and the other is an attribute value, then a mapping is generated between this class and this attribute value. This means that we assume that instances from le latter class which are proved to have this specific attribute values have the same meaning as all the instances of the former class.

4. PRACTICAL TESTS OF SCHEMA MATCHING

The proposed schema matching strategy has been implemented and tested on real geographic database schemas. The French national mapping agency (Institut Géographique National) produces several geographic databases covering the French territory. Due to technical and historical reasons, these databases have been designed and produced independently. Consequently, there is a great heterogeneity between them, which provides us with a realistic test application for our schema matching strategy.

4.1. Schema matching tests

Our tests use two IGN geographic databases: BDTOPO® (IGN, 2002), a geographic database with a metric resolution and BDCARTO® (IGN, 2005), another geographic database with a decametric resolution. Their schemas are available and encoded according to the ISO/TC 211 19109 standard "Rules for application schemas" (ISO/TC, 2001).

A first step for our application consists in translating these schemas into local ontologies. Then, a generic translator has been implemented. It takes an ISO 19109 schema as input and output OWL ontology (W3C, 2004), developed with the protégé-owl API, according to the approach presented in section 3.1. This means that FeatureType objects are translated into OWLNamedClass, AttributeType objects into OWLDatatypeProperty, AssociationType and AssociationRole objects into OWLObjectProperty, etc.

According to our proposition to use hidden geographic concepts, if a FeatureType instance has an AttributeType which define the specific nature of each geographic object of that FeatureType, then this AttributeType is not translated into OWLDatatypeProperty. Instead of that, its

FeatureAttributeValues are used to create OWLNamedClass, with subClassOf relationships with the OWLNamedClass derived from the corresponding FeatureType.

Figure 1 shows how a piece of BDCARTO® schema is translated into a local ontology. First, FeatureTypes "Commune" (*Locality*) and "Zone d'habitat" (*dwellings zone*) are translated into OWLNamedClass. The AssociationType "Chef-lieu" (*Capital*), and its associated AssociationRole "Est chef-lieu de" (*is the capital of*) and "A pour chef-lieu" (*has capital*), binding these FeatureType are then translated into an OWLObjectProperty (having an inverse property) binding "commune" and "zone_d__habitat" OWLNamedClass. The "Zone d'habitat" FeatureType's AttributeType "Importance" (*importance*) has three possible enumerated values: "Hameau" (*hamlet*), "Quartier de ville" (*quarter*), and "Chef-lieu de commune" (*district capital*). As these values refer to geographic concept labels, they will be translated into OWLNamedClass, with a subClassOf relationship with the OWLNamedClass "zone_d__habitat". Others AttributeType which do not refer to geographic concepts (not represented on the *Figure 1*) are lastly translated into OWLDatatypeProperty.

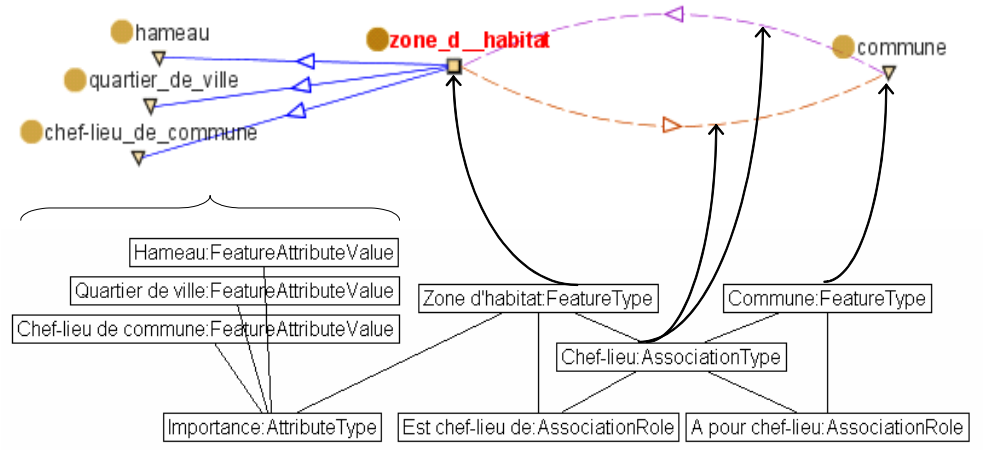

Figure 1: Translating ISO 19109 BDCARTO® schema (at the bottom of the image) into OWL ontology (piece of ontology visualized with Protégé, at the top of the image).

Then local ontology matching is performed. Both lexical and structural, and background knowledge approaches are used, in order to compare their respective problems and benefits. Background knowledge approach is performed with an existing OWL geographic taxonomy as external resource. This taxonomy, shown in Figure 2, which contains about 700 geographic terms, is organized in a hierarchy of *isA* relationships. It has been semi-automatically created by applying natural language processing techniques on geographic databases textual specifications (Abadie and Mustière, 2008).

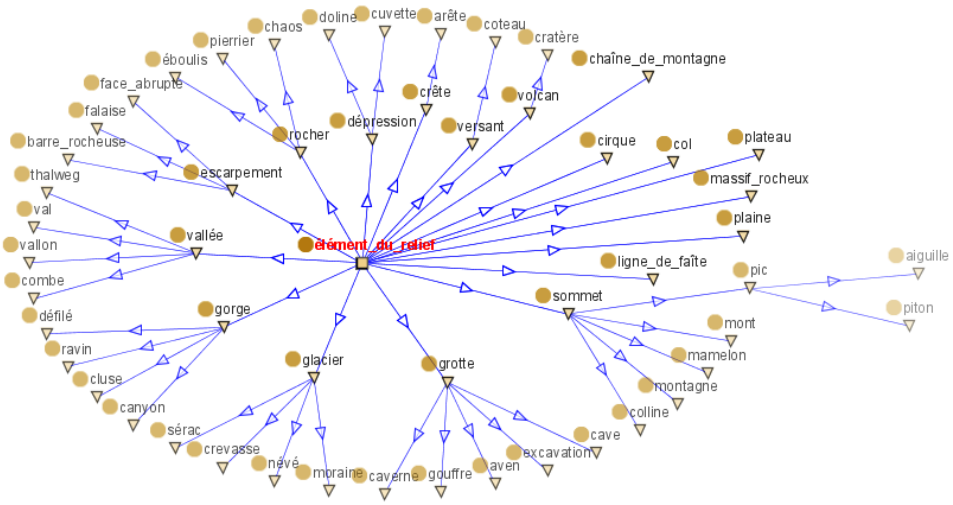

Figure 2: A part of the background taxonomy visualized with Protégé.

Each of the resulting alignment is lastly used to automatically retrieve schema correspondences, according to the approach presented in section 3.3.

4.2. Results

Automatic schema matching results have been evaluated in two different ways. First they have been interactively analyzed by experts having a good knowledge of both databases' specifications and contents. This analysis enables us to detect some interesting results and systematic mistakes. Second, in the particular case of relief points, the resulting mappings have been automatically compared with reference mappings to evaluate the benefits and problems of the background knowledge approach in relation to the lexical and structural approach. For readability reasons, in this section, matched schema elements will be presented according to the following format: Database name / FeatureType name / FeatureAttribute name / FeatureAttributeValue label.

The results obtained tend to prove that upgrading schema level of granularity by making hidden geographic concepts explicit in local ontologies significantly improves schema matching results. As a matter of fact, mappings that could not have been found just by comparing FeatureType names have been automatically discovered with this method. As an example, valid mappings between "BDTOPO® / Tronçon de chemin / Nature / Sentier" and "BDCARTO® / Tronçon de route / Etat physique de la route / Sentier" (*BDTOPO® / Trail section / Nature / Footpath* and *BDCARTO® / Road section / State of the road / Footpath*) or between "BDTOPO® / Oronyme / Nature / Grotte" and "BDCARTO® / Site et curiosité touristique / Nature / Grotte" (*BDTOPO® / Oronym / Nature / Cave* and *BDCARTO® / Touristic places and curiosities / Nature / Cave*) have been automatically detected, although in every cases FeatureType names (and in the first case AttributeType names also) in both databases are lexically and semantically different.

Moreover, combining schema level of granularity increase with the background knowledge approach provides better results. Even when schemas have totally heterogeneous terminologies or conceptualizations of the real world, valid mappings are found, thanks to domain knowledge provided by the taxonomic external resource. For instance, schema classes such as BDTOPO® / Zone arborée and BDCARTO® / Massif boisé (*BDTOPO® / Wooded area* and *BDCARTO® / clumps*) have been automatically matched although their FeatureType names and also their FeatureAttribute and

FeatureAttributeValue names are lexically different. Besides, in these databases, aqueducts are considered according to totally different points of view. Thanks to external knowledge, this cognitive difference has been identified and overcame. Thus, a mapping between BDTOPO® / Canalisation and BDCARTO® / Tronçon hydrographique / Nature / Aqueduc (*BDTOPO® / Canalization* and *BDCARTO® / Hydrographic section / Nature / Aqueduct*) has been automatically detected.

This improvement of the number and accuracy of detected mappings is confirmed by the comparison of mappings found by both approaches and reference mappings. The lexical and structural approach and the background knowledge approach were actually tested on the database specific themes of oronyms and relief points. The latter approach succeeded in detecting all expected correspondences, whereas the former approach failed in finding three mappings. These undetected mappings have appeared to be typical cases where external knowledge is needed. For instance, BDTOPO® specifications state that FeatureAttributeValue Oronyme / Nature / Gorge (*Oronym / Nature / Gorge*) represents either gorges, or cluses, defiles or canyons. In BDCARTO® specifications, cluses are stated to be represented in Point remarquable du relief / Nature / col, passage, cluse (*Outstanding relief feature / Nature / Moutain pass, passage, cluse*). The use of a background taxonomy provided the system with the information about the semantic proximity between gorge and cluse, which enabled the background knowledge approach to find this mapping.

However, even if combining schema level of granularity upgrade with the background knowledge approach significantly improves schema matching results, there still remain mappings that can not be found. Particularly, in some cases where the exact meaning of a class content given by the specifications implies geometrical selection conditions, like for the FeatureType “Surface d’eau” (*Water body*): “Watercourses wider than 7.5 meters are included”. The knowledge about the fact that some watercourses sections are represented in the same class as lakes or ponds is available in textual specifications only. Actually, specifications are a very rich source of knowledge about geographic database semantics, which would be useful in schema matching process.

5. CONCLUSIONS

We proposed, in this paper, a schema matching approach based on attributes values and background ontology. We start from the notion that there are specific attributes in geographic database classes whose role consists in specifying the exact nature of each class instance. Their enumerated values refer to geographic concepts and represent a rich knowledge source about the semantics of the class, which is usually hidden in the database structure. We assume that it is possible to take advantage of this knowledge by making it explicit in local ontologies created from each database schema that we have to match. A next step of our schema matching strategy consists in matching local ontologies thanks to additional domain knowledge, namely background ontology. Lastly, local ontologies alignments are used to compute schema correspondences.

Tests that we carried out tend to prove that it is possible to improve schema matching results by upgrading their level of granularity and using comprehensive domain taxonomy as knowledge resource. However, geographic databases modeling and capture follow specific rules, based not only on semantics, but also on geometrical, topological or cartographical criteria. These rules are stored in particular documents: the database specifications, which are the best available source of knowledge about geographic database semantics. Thus their use in schema matching should be further investigated (Mustière et al., 2003).

ACKNOWLEDGEMENT

This research is partly funded by the French Research Agency, through the GeOnto project ANR-O7-MDCO-005 on ‘Creation, Comparison and Exploitation of Heterogeneous Geographic Ontologies’ (http://geonto.lri.fr/).

BIBLIOGRAPHY

Abadie, N., Mustière, S., Construction d'une taxonomie géographique à partir des spécifications de bases de données. In Proceedings of SAGEO'08, Montpellier, France, 2008.

Aleksovski, Z., ten Kate, W., van Harmelen, F., Exploiting the structure of background knowledge used in ontology matching. In Proceedings of the Ontology Matching Workshop at 5th International Semantic Web Conference, Athens, Georgia, USA, pp 13-24, 2006.

Bishr, Y., Overcoming the Semantic and Other Barriers to GIS Interoperability. International Journal of Geographical Information Science, vol 12, n° 4, pp 299-314, 1998.

Euzenat, J., Shvaiko, P., Ontology Matching. Springer Verlag, 2007.

Fonseca, F., Clodoveu, D., Camara, G., Bridging Ontologies and Conceptual Schemas in Geographic Information Integration. GeoInformatica, vol 7, n° 4, pp 355-378, 2003.

Hakimpour, F., Timpf, S., Using Ontologies for Resolution of Semantic Heterogeneity in GIS. In Proceedings of 4th AGILE Conference on Geographic Information Science, Brno, Czech Republic, pp 385-395, 2001.

Hamdi, F., Zargayouna, H., Safar, B., Reynaud, C., TaxoMap in the OAEI 2008 alignment contest. In Proceedings of the Ontology Matching Workshop at 7th International Semantic Web Conference, Karlsruhe, Germany, pp 206-213, 2008.

IGN, BD Carto, Version 3, Spécification de Contenu, Edition 1, Institut Géographique National, Paris, France, 175 p, 2005.

IGN, BD Topo Pays, Version 1.2, Descriptif de Contenu, Edition 1, Institut Géographique National, Paris, France, 118 p, 2002.

ISO TC/211, Geographic Information – Rules for application schema, 2001.

Kokla, M., Guidelines on Geographic Ontology Integration. In Proceedings of the ISPRS Technical Commission II Symposium, Vienna, Austria, pp 67-72, 2006.

Manoah, S., Boucelma, O., Lassoued, Y., Schema Matching in GIS. In Proceedings of Artificial Intelligence: Methodology, Systems, and Applications, 11th International Conference, AIMSA 2004, Varna, Bulgaria, pp 500-509, 2004.

Mustière, S., Gesbert, N., Sheeren, D., A formal model for the specifications of geographic databases. In proceedings of GEOPRO Conference: Semantic Processing of Spatial Data, Mexico, 2003.

Partridge, C., The role of ontology in integrating semantically heterogeneous databases. Technical Report 05/02 LADSEB-CNR, Padoue, 2002.

Protégé-owl api, http://protege.stanford.edu/plugins/owl/api/ (accessed 16/01/2009).

Rahm, E., Bernstein, Philip A., A survey of approaches to automatic schema matching. The VLDB Journal, vol 10, n°4, pp 334-350, 2001.

Sheth, A., Larson, J., Federated database systems for managing distributed, heterogeneous and autonomous databases. ACM Computing Surveys, vol 22, n° 3, pp 183-236, 1990.

W3C, OWL Web Ontology Language, Overview, W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-features/ (accessed 16/01/2009), 2004.