

ONTOLOGY-DRIVEN RESOLUTION OF SEMANTIC HETEROGENEITIES IN GDB CONCEPTUAL SCHEMAS

GUILLERMO NUDELMAN HESS, CIRANO IOCHPE
Universidade Federal do Rio Grande do Sul. Instituto de Informática
{hess,ciochpe}@inf.ufrgs.br

Abstract: As the geographic information system (GIS) community grows, more and more people need to share geographic information. As modern GIS data is stored in geographic databases their conceptual schemas have to be, at least, interoperable. Due to the fact that the databases are designed by many different people from different countries, using different languages and maybe with different definitions for the same phenomenon there is a high probability that the conceptual schemas have semantic heterogeneities between them. In order to handle these heterogeneities this paper suggests the use of ontologies as mediators to the semantic integration. A software architecture is proposed, which handles also syntactic heterogeneities using a standard language, the GML beyond the semantic ones. An ontology that represents a subset of the geographic reality was created and a similarity matching algorithm was developed to process schemas against it. The mathematical methods of similarity measurement modeling concepts have been tuned for a set of real GDB conceptual schemas.

Key words: Ontology, Conceptual Schema, Semantic Integration, Similarity Matching

1. INTRODUCTION

As the geographic information system (GIS) community grows, more and more people need to share geographic information. As modern GIS data is stored in geographic databases their conceptual schemas have to be, at least, interoperable. In this context, the conceptual modeling of Geographic Databases (GDB) has become a very important task due to both the

increasing exchange and reuse of geographic information (Burrough and McDonnell, 1997). The well-defined conceptual modeling offers a canonical conceptual representation of the geographic reality enabling the reuse of design sub-schemas. Furthermore, conceptual modeling is an important task for the understandability and extensibility of the database being developed.

Despite a well defined conceptual schema guarantees the reusability, extensibility and understandability, it does not guarantees that two conceptual schemas can be compared, integrated or merged, since they may have semantic differences, while modeling the same portion of the reality.

In order to establish a correspondence among different representations of a same real world concept, that were defined in different schemas, it is necessary to recognize the common concept through the identification of similarities as well as conflicts among those schemas (Gotthard and Lockemann, 1998). A semantic conflict, or heterogeneity, occurs when the same real world entity, modeled by two or more people, probably will not have the same modeling, even though it is representing the same phenomenon of the application's domain. In these cases occur what is called a conflict. A conflict is nothing else than a difference in the representation of the same concept.

To achieve this purpose, an ontology can be built to store the concepts concerning the application's domain. Due to the complexity of the geographic database modeling and repeatability of the modeled phenomena an ontology storing those phenomena and relationships may also be useful to the conceptual modeling of new applications (Sugumaran and Sorey, 2002). Furthermore, an ontology is semantically richer than a conceptual schema and thus closer to the human's cognitive model. While an ontology is used to represent the real world concepts a conceptual schema is developed to organize what is going to be stored in a database (Fonseca, Davis and Câmara, 2003).

More than just to construct an ontology, algorithms of similarity matching must be applied to achieve correct interpretation for the variations of known terms, and classification of new ones. This measurement of similarity must be balanced, since the input conceptual schemas may have different characteristics. We propose weights for the similarity measurement for some types of geographic database schemas, based on some case studies we have carried out.

The remaining of this paper is organized as follows. Section 2 presents a software architecture for the conceptual schemas integration, and briefly outlines the issues on the syntactic and semantic parts where ontology is applied. The methodology of the semantic integration is shown in section 3. A case study is presented in section 4. At last, the conclusions and future work are shown in section 5.

2. RELATED WORKS

Fonseca, Egenhofer, Agouris and Câmara (2002) proposed an ontology-driven GIS architecture to enable geographic information integration. In that proposal the ontology acts as a system integrator independently of the model (Fonseca, Egenhofer, Agouris and Câmara, 2002). As uses the hierarchical representation of ontologies to compare concepts, in many times only a partial integration is possible, when finding the same super-concept of two specialized concepts. In that methodology the concept attributes and roles are used as the integrations fields.

Hakimpour and Timpf (2001) propose the use of ontology in the resolution of semantic heterogeneities focused in Geographic Information Systems. In that work the authors specify the ontological issues, present a little set of concepts in Description Logic (DL) and the concept's features that must be considered when solving heterogeneities: names, relations (attributes) and taxonomic relationships.

Uitermark, van Oosterom, Mars and Molenaar (1999) present a framework to aid the geographic data integration (not schemas, as in this paper). The proposal uses a domain ontology specific for topography data. The target was to enable queries in a distributed environment with heterogeneous data.

Stoimenov and Djordjevic-Kajan (2003) proposed the GeoNis framework to reach the semantic interoperability between GIS data. The use of ontologies was proposed in the context of serving as a knowledge base to solve semantic conflicts as homonyms, synonyms and taxonomic heterogeneities. The need of some type of syntactic integration is also pointed as an important task in order to unify the possibly different data models.

Even though most of the works are focused in data and this work is focused in conceptual schemas, the idea behind is practically the same. However, none of the cited works specifies an algorithm or methodology to search the ontology given a concept. Furthermore, no matching methods were described.

3. THE INTEGRATION ARCHITECTURE

Most of the Geographic Information Systems (GIS) support a proprietary data model, specific for a particular system architecture. This design scenario leads to non-reusable as well as implementation-dependent geographic database projects. Because of the missing of a standard model to design geographic conceptual schemas, such as the Extended Entity-

Relationship for conventional, non-spatial databases, some preliminary work has to be done to enable the treatment of semantic heterogeneities. In other words, the GDB conceptual schemas often have syntactic heterogeneities between them.

Since there is not a modeling standard for GDB, a variety of conceptual data models and meta-models (e.g., the OGC's GML (OpenGIS, 2003)) as well as modeling frameworks (e.g., UML-GeoFrame (Lisboa and Iochpe, 1999; Rocha, Edelweiss and Iochpe, 2001), MADS (Parent et. al., 1999) and OMT-G (Borges, 1997)) have been proposed. The core of most of them is equivalent, and a complete comparative study is presented in (Bassalo, Iochpe and Bigolin, 2002). Anyone of them could be used as the canonical model in the preliminary (also called preparation) phase of the semantic heterogeneities resolution.

Putting the syntactic and semantic integrators in the same software architecture (illustrated in Figure 1) a conceptual schema is primarily converted into a Canonical Syntactic Format (SCF) that is, only in the syntactic level and then to a Canonical Semantic and Syntactic Format (CSSF).

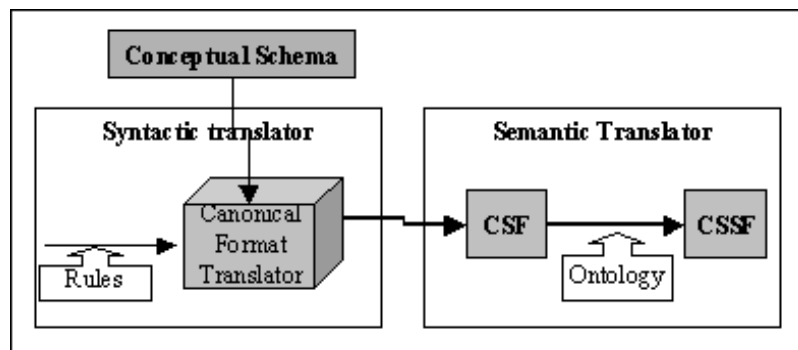


Figure 1. The architecture for the integration

To make the integration of geographic application possible, three requisites must be satisfied (Batini, Lenzerini and Navathe, 1986):

- The conceptual schemas of each source must be available;
- There must be semantic information in the schema;
- A canonical data model must exist. This standard model has to have enough expressiveness power to describe all the models to be integrated;

Since the target of the integration proposed in this paper is of conceptual schemas, the first requisite is automatically satisfied. The other requisites are filled by the matching of the constructors of the different data models with the same meaning and by choosing one to be the standard one.

The Geographic Markup Language (GML) (OpenGIS, 2003) is used in our methodology as the canonical data model, since it is a standard for storing and exchanging geographical data. Even knowing the GML is not capable to represent all of the constructors from all the data models it was adopted as it covers a significant set of elements used in the GDB modeling. In addition, GML may be extended to handle the missing constructors.

According to the data model in which the schema is based a specific set of rules is applied. Some translation rules can be applied for the constructors of more than one data model once they are common to most of them (e.g. classes, attributes, simple associations) and some ones are specific for the data model, depending on its purposes (e.g. net topology, temporality). In (Hess and Lochpe, 2003) the rules for the syntactic integration are specified, and all the syntactic translator is detailed. At the moment, the GeoFrame, MADS and OMT-G have been mapped to GML 3.

Once having the schemas described in the same data model (or data format) the semantic heterogeneities are able to be handled. The proposed system's semantic translator is responsible for that task, aided by an ontology. The semantic converter module is responsible for solving all sort of semantic conflicts and heterogeneities. Each one of the elements of this GML file is compared against the ontology's concepts to find a match for it. After processing all the CSF's elements, the result is a canonic syntactic and semantic format (CSSCF). This CSSF file is also in GML and free of heterogeneities and ambiguities.

It is a consensus that to perform the semantic integration in a semi-automate way, allowing the reuse of the modeled elements, the use of a Knowledge Organization System (KOS), such as an ontology (Guarino, 1998), is mandatory (Hodge, 2000).

The role of the ontology is similar to the role of the global conceptual schema proposed in the works of Batini, Lenzerini and Navathe (1986) and Hayne and Ram (1990). It also may be seen as a mediator as proposed by Fonseca et al. (2002). Each one of the conceptual schemas to be integrated is compared with the ontology, and for each conflict found the system calculates a similarity measurement between the ontology's concept and the input conceptual schema's element.

4. THE ONTOLOGY'S ROLE

Before explaining how does the ontology works in our propose, it is important to clarify the heterogeneities issues it must handle. There are two basic types of ontological heterogeneity: conceptualization and explanation (Visser, Jones, Bench-Capon and Shave, 1997). The first happens when two

or more conceptualizations differ in terms of concepts covered, that is, they are not from the same domain, or in concepts relationships, which means that the concepts are the same but in a different context. An explanation mismatch occurs when two schemas have distinct definitions but their terms, meanings or descriptions are the same.

To formalize the different semantic conflicts we can separate them in four types: equality, dissimilarity, intersection and contain (Stoimenov and Djordjevic-Kajan, 2003).

Semantic equality (similarity) $SEqu(c1,c2)$ occurs when there is a 1:1 mapping between concepts from different schemas (S) or ontologies (O) in terms of meaning and structure (attributes and relationships). The concepts of this type of semantic heterogeneity are called synonyms, and it can be defined as:

$$SEqu(c1,c2) = \{(c1,c2) | c1 \in S \wedge c2 \in O \wedge E(c1) = E(c2) \wedge S(c1) = S(c2)\}$$

Where $E(ci)$ is the meaning of the concept and $S(ci)$ is the concept's structure.

Semantic dissimilarity $SNEqu(c1,c2)$ occurs when there is no mapping between the description of a concept $c1$ from a schema S and a concept $c2$ from the ontology O . In addition, if the $Name(c1)$ is equal to $Name(c2)$, the semantic heterogeneity is also called homonym. The semantic dissimilarity can be defined as:

$$SNEqu(c1,c2) = \{(c1,c2) | c1 \in S \wedge c2 \in O \wedge E(c1) \neq E(c2) \wedge Name(c1) = Name(c2)\}.$$

Semantic intersection $SIntersec(c1,c2)$ occurs when there is a partial 1:1 mapping between the concept $c1$ from the schema S and the concept $c2$ from the ontology O , in terms of structure and in terms of meaning. The later case may happen when a same concept is being modeled for different applications and therefore some attributes and relationships are the same and some are not the same. This semantic heterogeneity is defined as:

$$SIntersec(c1,c2) = \{(c1,c2) | c1 \in S \wedge c2 \in O \wedge S(c1) \cap S(c2) \wedge S(c1) \not\subset S(c2) \wedge S(c2) \not\subset S(c1)\}.$$

At last, the fourth type of semantic conflict is the contain $SContain(c1,c2)$ which occurs when the structure of a concept $c1$ from a schema S is contained in the structure of the concept $c2$ from the schema, or the contrary. This case happens when one concept is a specialization or generalization of the other. This conflict is defined as:

$$S\text{Contain}(c1,c2) = \{(c1,c2) \mid c1 \in S \wedge c2 \in O \wedge S(c1) \subset S(c2) \wedge S(c2) \not\subset S(c1)\}.$$

4.1 The similarity measurement

In our proposal, we combine the use of an ontology as a sort denominator between conceptual schemas, in order to achieve integration, with some similarity matching calculus. We decided to use the similarity measurement because the similarity of two concepts can not be binary, that is, has only the values 1 to equivalent and 0 to not equivalent. To reach a more accurate level of similarity among the concepts, in this section we describe the mathematical formulas used during the comparison algorithm described in the next section. As result, the similarity probability may assume any value between 0 (totally different) and 1 (equivalent).

To measure the similarity between two concepts, one from the input conceptual schema and the other from the ontology, we combine syntactic matching between strings and semantic matching (Hess and Iochpe, 2004).

In the syntactic matching, a distance function is applied over a pair of strings, to determine the dissimilarity between them. The smaller this dissimilarity (measured by an integer value) is, the more similar are the strings (Cohen, 1998). In this work we adopted the Levenshtein distance. It is applied to the calculus of similarity between concept names ($\text{SimName}(Cc,Co)$) and attributes names.

The techniques to calculate the distance between two strings may be applied to acronyms and typing error cases, but no semantic issues are considered by these functions. Therefore, a correct semantic unification of concepts must be accomplished by a complementary technique that is capable of both detecting synonyms and considering the context in which those concepts exist.

Our approach considers two semantic techniques to compare a pair of concepts. The first one is the nearest neighbor (Holt, 2000), which is used to calculate the similarity in terms of the attributes each concept presents, and is given by the formula:

$$\text{SimAt}(Cc,Co) = \sum_{i=1}^n f(Cc_i,Co_i) \times W_{ati} \quad (1)$$

where Cc and Co are, respectively, the conceptual schema's and the ontology's concept, n is the number of attributes considered, i is the index of the attribute being processed, $f(Cc_i,Co_i)$ is the distance function between the

attributes of the compared concepts (Levenshtein) and Wat_i is the weight of the i^{th} attribute in the ontology.

The weight of an attribute is given by an adapted TF-IDF (Cohen, 1998) formula:

$$W_{at} = 1 - (C_a/C) \quad (2)$$

where C_a is the number of concepts that have the attribute, and C is the total number of concepts. The more concepts have the same attribute, the less significant this attribute is.

Three types of relationships are considered for the similarity measurement of a pair of concepts. The first one is the taxonomic (IS-A) associations, and the others two are the aggregation and composition ones. The similarity in terms of the place in the hierarchy where each concepts is located is obtained by the formula:

$$\text{SimHier}(Cc, Co) = \frac{(\sum(\text{Hier}(Cc, Pc) \cdot \text{Wt}(c, p))}{\text{Nhier}(Cc, Pc)} \quad (3)$$

where $\text{Hier}(Cc, Pc)$ is each one of the taxonomic relationships existing in both the conceptual schema and in the ontology. $\text{Wt}(c, p)$ is the weight of the hierarchical relationship arc and $\text{Nhier}(Cc, Pc)$ is the number of IS-A associations in both the ontology and the conceptual schema.

The weight $\text{Wt}(c, p)$ of a taxonomic arc is given by the following formula (Jiang and Conrath, 1997):

$$\text{Wt}(c, p) = \frac{(E) \cdot (d(p)+1) \cdot (IC(c) - IC(p))}{E(p) \cdot d(p)} \quad (4)$$

where $d(p)$ is the depth of the parent node (p) of the node corresponding to the concept being compared. E is the density of the whole ontology's hierarchy, that is, the number of nodes it has. $E(p)$ is the density of the taxonomy considering the node p as the root concept, that is, the number of direct and indirect children it has. Finally, IC (Information Content) represents the amount of information the node has (Resnik, 1998), and its value is given by:

$$IC(c) = -\log((\sum(1/\text{sup}(c))).1/N) \quad (5)$$

where $\text{sup}(c)$ is the number of super classes (direct or indirect) the class c has, and N is the total number of concepts of the ontology. The more specialized a concept is, the more information it intrinsically possesses.

Finally, the aggregation and composition links are considered to calculate the similarity of two concepts, by the simple formula:

$$\text{SimRel}(Cc,Co) = (\sum(\text{Rel}(Cc,Co))/\text{Rel}(Cc)) \quad (6)$$

where $\text{Rel}(Cc,Co)$ is each composition/aggregation link existing both in the ontology and in the conceptual schema and $\text{Rel}(Cc)$ is the ones present only in the conceptual schema.

The final value of similarity is given by a weighed sum of the similarities:

$$\text{Sim}(Cc,Co) = \text{WN}.\text{SimName}(Cc,Co) + \text{WA}.\text{SimAt}(Cc,Co) + \text{WH}.\text{SimHier}(Cc,Co) + \text{WR}.\text{SimRel}(Cc,Co) \quad (7)$$

where WN, WA, WH and WR are the weights for names, attributes, hierarchies and relationships similarities, respectively.

The $\text{Sim}(Cc, Co)$ value is calculated for every concept in the conceptual schema against each concept present in the ontology. The higher the $\text{Sim}(Cc,Co)$ value is, the more similar the concepts are.

During this process of similarity measurement, the original conceptual schema is kept unaltered and a new, equivalent one is generated relying on the canonical semantic model expressed by the ontology.

The ontology can also be dynamically updated depending on the similarity measurements carried out with every new GDB schema. Attributes and relationships can be added to existing concepts, and even new concepts may be inserted.

4.2 The ontology algorithm

The algorithm described next and illustrated in Figure 2 details in a high abstraction level, the steps sequence to search and update the ontology. The similarity matching formulas presented in the preview section are applied in some of the algorithm steps.

To minimize the need of the expert intervention two parameters have to be set at the beginning of the algorithm execution: the analysis threshold and the acceptance threshold. Only the concepts having similarity probability higher than the specified analysis threshold are considered. If no candidate reaches the threshold, the input concept is considered as not existing in the ontology and added as a new concept. If one or more of the ontology candidates have similarity probability higher than the acceptance threshold

specified by the user, the one with the higher value is considered as equivalent (a synonym) of the input concept.

As the similarities probabilities are rough, a third parameter has to be defined to increase the algorithm's confidence. It is a "confidence value" *delta* that indicates the accepted error in the similarity measurement. The actual minimum limit is given by analysis threshold minus *delta*. In the same way, if a candidate has its similarity probability higher than the acceptance threshold it is not automatically consider as the one equivalent to the input concept. All the candidates with the similarity value higher than the maximum similarity obtained minus *delta* are considered and presented to the expert.

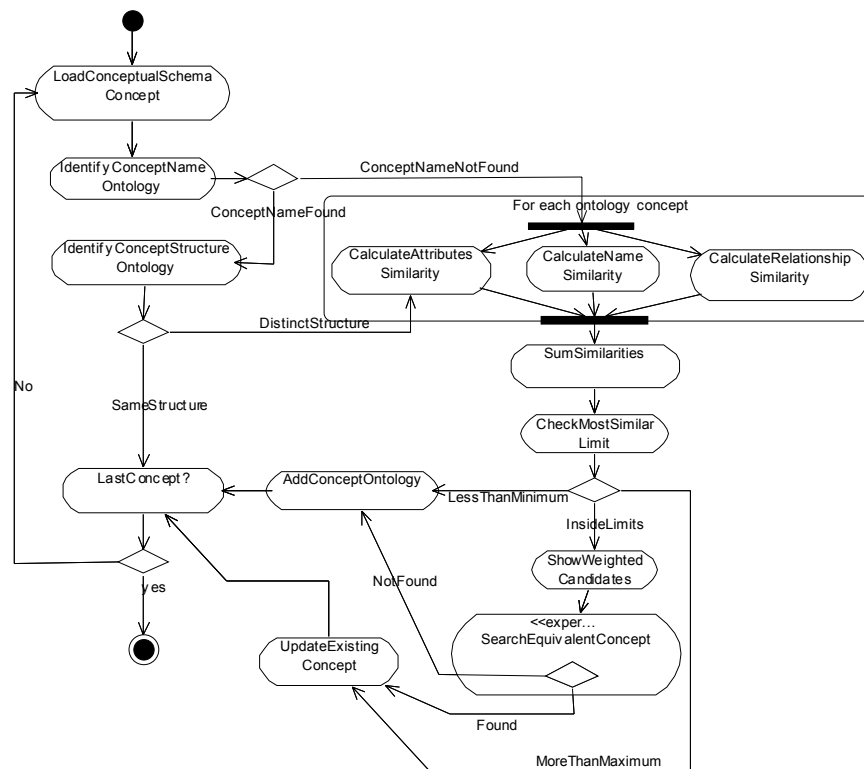


Figure 2. The ontology search algorithm

To ensure the correct operation of the algorithm, it is necessary that every input conceptual schema has an associated metadata, specifying in which language the modelling is based.

Step 0 – Schema translation to the ontology's language: If the ontology's language is not the same of the one indicated by the conceptual schema's metadata, this has to be translated, aided by a dictionary.

Step 1 – Search concept’s name in the ontology: If the concept’s name or one of its synonyms or acronyms is found in the ontology, go to step 2. Else go to steps 4, 5 and 6, in parallel.

Step 2 – Search concept’s structure in the ontology: Once the term which nominates the concept is found in the ontology, its structure is compared against the ontology, attribute by attribute. The algorithm verifies if there is an one to one correspondence between each input concept’s attribute and the ontology concept’s attribute. If the equivalence is complete go to step 3. If there are differences in at least one of the attributes, go to step 5.

Step 3 – Tests if it is the last concept: If the current concept is the last one of the schema, go to the end. Else go back to step 1 to processes of the next concept.

Step 4 – Calculate the similarity of the term that nominates the concept: The similarity of the input concept’s name is compared against the name of each ontology’s concept. Go to step 7

Step 5 – Calculate concept’s structural similarity: The input concept’s structural similarity is calculated, in terms of its attributes against each one of the ontology’s concept. Go to step 7.

Step 6 – Calculate relationship similarity: The input concept’s relationship similarity is calculated, in terms of aggregation and composition associations, and also in terms of taxonomic (IS-A) relations against each one of the ontology’s concept. Go to step 7.

Step 7 – Sum of the similarities: Based on some method of balance, the structural similarity, the name similarity and the relationship similarity of the input concept are summed, resulting in the similarity probability. This calculation is performed for each ontology’s concept. Go to step 8.

Step 8 – Verify threshold: Check if the similarity value of the concept with the highest similarity probability. If it is lower than the analysis threshold minus *delta* go to step 12. If the similarity value is greater than the acceptance threshold and there are no other candidates with similarity probability higher than the highest similarity value minus the *delta* parameter go to step 11. Otherwise, go to step 9.

Step 9 – Show candidates: Present each found candidates, with its balanced similarity probability. If there are candidates with similarity value higher than the maximum threshold only the candidates within the *delta* parameter are shown. They are displayed ordered, with the ones with higher similarity first. Go to step 10.

Step 10 – Term selection: At this point the domain expert intervention is necessary. He selects the concept he judges as the most equivalent to the input schema’s concept. If an ontology’s existing concept is selected, go to step 11. If the expert decides that the input concept has not an equivalent in the ontology go to step 12.

Step 11 – Update an existing concept: Depending on which step called this step, a distinct action is performed to update the ontology. This action can be the addition of a new synonym or acronym to an existing term, the addition of a new attribute to an existing concept's structure, or the creation of a new relationship between two existing concepts. Go back to step 3.

Step 12 – Addition of a new concept to the ontology: A new concept is added in the on the ontology, with all its attributes. Go back to step 3.

The use of ontology by itself does not provide a complete solution to the semantic integration problem. It is impossible to a single ontology to contemplate all the alternatives to express a real world phenomenon. This happens because of the inherent restrictions to the ontology and because of the differences derived by the individual process of interpretation of the reality (Resnik, 1998).

The human intervention in the resolution of the conflict is practically mandatory in the identification of correspondences process between different schemas. At most, what can be reached is that the ontology suggests the best solutions based on similarity and probability calculus.

5. CASE STUDY

In order to test the algorithm as well as the similarity matching formulas proposed, a geographic ontology called ontoGeo was built. Since it is a domain ontology, any other geographic application may use the concepts stored in this ontology. The concepts represented in ontoGeo belong to the domain of the physical, natural phenomena (basic cartography) such as hydrology, relief, and vegetation. There are also some concepts related to infra-structure, especially those of the transportation theme, and some from the locality domain.

ontoGeo was built using the semi-structured language RDF schema. The concepts described in ontoGeo were taken from a set conceptual GDB schemas relying on different data models and modeled by different designers working for distinct organizations. They are all GDB schemas for real applications that have been actually implemented.

Since ontoGeo is based on geographic databases that were implemented to support geographic information systems of brazilian organizations, it supports only concepts written in Portuguese in its current state of development.

At the beginning of the case study, ontoGeo had 156 classes (concepts) and 104 slots (attributes and relationships). It is worth noting that ontoGeo is not intended to be a complete geographic ontology. It serves as a starting point to integrate GDB conceptual schemas. In case a concept belonging to

an input schema is not found in the ontology, ontoGeo is updated with the insertion of the new concept, as predicted in the algorithm presented in section 3.

5.1 The algorithm execution

Since most of the schemas evaluated in the case study model hydrology, only the subset of ontoGeo's concepts related to this theme has been used.

We ran the experiments considering three different types of conceptual schemas. The first type represents those schemas that model only classes and taxonomic relationships. Schemas that model classes with attributes and taxonomic relationships were considered as belonging to a second type. The last type represents schemas that can have both classes with and without attributes, besides hierarchies. Furthermore, as a GDB schema of the third type was processed by the similarity measurement algorithm ontoGeo had already been updated on the basis of schemas of the other two types.

As shown in Table 1, for each schema the similarity matching algorithm was executed twice, each time with a different set of values for the weights WN, WA, WH, and WR. In the following, each one of these sets is called a different case study scenario.

For all three GDB schemas processed, in the first scenario all weights were given the value of 0.25. In the second scenario for each schema the WN, WA, WH, and WR assumed different values depending on the characteristics of the input conceptual schema.

Table 1. The case study scenarios

Input Conceptual Schema	Test	WN	WA	WH	WR
Classes and Hierarchies	1	0.25	0.25	0.25	0.25
Classes and Hierarchies	2	0.70	0.00	0.30	0.00
Classes, Attributes and Hierarchies	1	0.25	0.25	0.25	0.25
Classes, Attributes and Hierarchies	2	0.45	0.35	0.20	0.00
Classes, Attributes, Hierarchies, Ontology Updated.	1	0.25	0.25	0.25	0.25
Classes, Attributes, Hierarchies, Ontology Updated.	2	0.50	0.25	0.25	0.00

In this first case study we report here, the main goal was to determine a good scenario where the weights for the different types of similarity formulas express their relative importance within the global similarity formula (7). A next step should be to investigate the best set of values for the set of weights concerning both the similarity matching algorithm and the ontology ontoGeo. It is possible, though, that one comes up with a set of good scenarios, each one of which being the best fit for a certain type of GDB schema.

The acceptance threshold was fixed in 0.75 (75%), while the analysis threshold was set to 0.4 (40%), and the *delta* value kept by 0.1 (10%).

With the above values assigned to the respective parameters, no equivalent concept in ontoGeo was found for any of the processed concepts of the input schema. In this case, SimAt and SimRel were always zero, and thus the similarity had at most the value of 0.5. As we increased the values of both WN and WH, and decreased the values of WR and WA down to zero, the algorithm showed very good results, with 100% of accuracy in automatically matching equivalent concepts with more than 75% of similarity, and 60% of precision in finding equivalent concepts in ontoGeo within 40% and 70% of similarity.

The second conceptual schema that was evaluated against ontoGeo contains taxonomies, classes and attributes. In a first execution of the algorithm, the results were 100% correct in the cases where similarity fell within the threshold limits. Again, no match was indicated as having more than 75% of similarity because no aggregation or composition associations were modeled. In a second execution, by increasing the values of WN and WA as well as decreasing the value of WH a little and setting WR to zero, the results were the same and a correct automatic matching was obtained once again.

The third conceptual schema contains taxonomies, classes and attributes. It was processed by the similarity matching algorithm after the ontology had been updated. In its new state, it presented 170 classes and 106 slots. As most of the constructs of this third schema were already stored as concepts in the ontology, the results of the two executions (i.e., one for each different scenario) were almost the same.

As already mentioned above, for each schema, in the second execution of the algorithm, we used a customized combination of weights to the parameters WN, WA, WH and WR. This was due to the different characteristics of each input schema. In the first schema, as only classes and taxonomies were modeled, only the WR and WA parameters mattered. Since almost all classes were specialized from the same super class, the hierarchy was less significant than the name to differentiate the classes. For that reason WN was much higher than WH.

In the second schema, as there were classes as well as attributes, and taxonomies, but the hierarchies were again not so deep, WH was given a low value. Since a considerable number of attributes were present in more than one class, the name similarity was more significant than the attribute similarity. Thus, WN was a little higher than WA.

Finally, in the third conceptual schema, the main difference was that a number of classes did not have attributes. For that reason WA was given a lower value than in the experiment with the second schema. As aggregation and composition relationships were not present in any of the schemas, WR was always set to zero.

Even if the absolute values of the parameters WN, WA, WH and WR were not entirely correct, from the experiments above it is possible to conclude that we achieved a reasonable set of ratios among them for the input schemas that were tested and, therefore, have obtained a very good similarity measurement between concepts.

6. CONCLUSIONS AND FUTURE WORKS

This work aimed to address an important issue for the interoperability of geographic data and conceptual schemas, which is the integration of different schemas. Input data must be syntactically as well as semantically unified as to produce good results. Especially for GDB conceptual schemas, the integration plays a very important role since it handles heterogeneities in terms of data models (syntax) and in terms of concepts (semantics). It was made clear that the use of an ontology associated to a canonical data model helps enhance the schema unification process. The technique can also improve the process of understanding as well as avoiding conflicts such as those originated from heterogeneities and redundancy of concepts.

On the basis of the software architecture proposed as well as relying on the ontology created, and the similarity matching algorithm developed, we were able to test the proposed metrics and also start tuning the weights related to each of the different similarity measurement formulas proposed here.

The results obtained by the case study were very satisfactory with a high rate of correct matching. By the execution of the algorithm a supposition was confirmed: the more detailed the conceptual schema is, the more precise is the algorithm output. If only classes are modeled, probably a considerable number of candidate matches may be found in the ontology for each one of the concepts of the input GDB schema, as only the concept name will differ. On the other hand, if the elements are modeled in a complete taxonomy and with a number of attributes are defined, the ontology's concepts would have to match more requisites, producing less, but more accurate, candidate matches.

On the other hand, the case study tends to show that there is not a perfect combination of values for the set of weights composed by WN, WA, WR and WH. The relations among them depend on the characteristics of the input conceptual schema being processed. In this work we did not test all different types of schema that may exist. Though, the case study showed that might be important to be able to tune those parameters for each different type of schema, in order to obtain the best combination each time.

A second important point to be considered is the fact the algorithm is not capable to solve heterogeneities in terms of different data model constructors used to model the same concept of the real world. For example, if a concept in one schema is represented as a class and in another as an attribute, the current algorithm does not match one to another. This problem should be dealt with in the future. We also plan, as future work, to extend the algorithm to take into account the spatial features of the schemas to be compared. Spatial attributes as well as spatial relations are meant to be addressed.

7. REFERENCES

- Bassalo, G., Iochpe, C., and Bigolin, N. Representing schemas in the attribute-value format for the inference of analysis patterns. In Proc. of the IV Brazilian Symposium on GeoInformatics (GeoInfo). Caxambu, Brazil, 2002 (in portuguese).
- Batini, C., Lenzerini, M., and Navathe, S. A Comparative Analysis Of Methodologies For Database Schema Integration. In *ACM Computing Surveys*, v.18, n.4, p.323-364, 1986.
- Borges, K.V.A. Geographic Data Modeling: An Extension to the OMT Model for Geographic Applications. In Escola do governo de MG/FJP. Belo Horizonte, Brazil, 1997 (in Portuguese).
- Burrough, A.; McDonnell, R. A. Principles of Geographical Information Systems. Great Britain: Oxford university Press, 1997.
- Cohen, W. Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Text Similarity. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data. 1998.
- Fonseca, F. T., Egenhofer, M. J., Agouris, P. and Câmara, G. Using Ontologies for Integrated Geographic Information Systems. In: *Transactions in GIS* 6(3), 2002.
- Fonseca, F. T., Davis, C., Câmara, G. Bridging Ontologies and Conceptual Schemas in Geographic Information Integration. *GeoInformatica*. v.7, n.4, p.355-378. Kluwer Academic Publishers, 2003.
- Gotthard, W., Lockemann, P. C., and Neufeld, A. System-Guided View Integration for Object-Oriented Databases. *Ieee Transactions on knowledge and data engineering*, p.1-22, Vol.4, n° 1, Feb.1992.
- Guarino, N. Formal Ontology and Information Systems. In Proc. of In International Conference on Formal Ontology in Information Systems (FOIS'98).Italy, june, 1998.
- Hakimpour, F.; Timpf, S. Using Ontologies for Resolution of Semantic Heterogeneity in GIS. In *4th AGILE Conference on Geographic Information Science*. Brno, Czech Republic. April, 2001.
- Hayne S.; Ram, S. Multi-User View Integration System (MUVIS): An Expert System for View Integration. In *proc. Of the 6th International Conference on Data Engineering*, p. 402-409, 1990.
- Hess, G., and Iochpe, C. Using the GML for the Identification of GDB Analysis Patterns Candidates. In proc. Of the V Brazilian Symposium on GeoInformatics (GeoInfo). Campos do Jordão, Brazil, 2003 (in portuguese).
- Hess, G.; Iochpe, C. Applying Ontologies in the KDD Pre-Processing Phase. In Proc. of 16th International Conference on Software Engineering and Knowledge Engineering (SEKE'04). Banff, Canada. June 2004.

- Hodge, G. Knowledge Organization Systems: An Overview. In System of knowledge Organization for Digital Libraries: Beyond Traditional authority files. April, 2000.
- Holt, A. Understanding environment and geographical complexities through similarity matching. In Complexity International, number 7, 2000.
- Jiang, J., and Conrath, D. Semantic Similarity Based in Corpus Statistics and Lexical Taxonomy. In Proc of International Conference Research in Computational Linguistics (ROCLING X). Taiwan, 1997.
- Lisboa F. J., and Iochpe, C. Specifying analysis patterns for geographic databases on the basis of a conceptual framework. In Proc. 7th ACM GIS, Kansas City, USA, 1999.
- OpenGIS Consortium. Geography Markup Language (GML) 3.0. Open GIS Implementation Specification, 2003. Available in <http://www.opengis.net>. Last access in december 2003.
- Parent, C. et al. Spatio-temporal conceptual models: data structures + space + time. In Proc. 7th ACM GIS, Kansas City, USA, 1999.
- Resnik, P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, number 11, p. 95-130, 1998.
- Rocha, L. V., Edelweiss, N., and Iochpe, C. GeoFrame-T: A Temporal Conceptual Framework for Data Modeling. In: ACM Symposium on Advances in GIS. Atlanta, USA, 2001.
- Stoimenov, L.; Djordjevic-Kajan, S. Realization of GIS Semantic Interoperability in Local Community Environment. In Proc. of the 6th AGILE Conference on Geographic Information Science. Lyon, France. April, 2003.
- Sugumaran, V., Srorey, V. C. Ontologies for Conceptual Modeling: their creation, use and management. Data & Knowledge Engineering. Elsevier, april, 2002.
- Uitermark, H. T., Van Oosterom, P. J. M and Molenaar, M. Ontology-Based Geographic Data Set Integration. International Workshop on Spatio-Temporal Database Management (STDBM'99). Edinburg, Scotland. September, 1999.
- Visser, V., Jones, D., Bench-Capon, T, and Shave, M. An Analysis of Ontology Mismatches Heterogeneity versus Interoperability. In proc. Of the AAAI 1997 Spring Symposium on Ontological Engineering, 1997.