A CONCEPTUAL FRAMEWORK FOR QUALITY ASSESSMENT OF SEMANTIC MAPPING BETWEEN SPATIAL ONTOLOGIES

Mohamed Bakillah[a], Mir Abolfazl Mostafavi[a], Yvan Bédard[a], Jean Brodeur[b]

[a] CRG, 0611 Pavillon Casault, Département des Sciences Géomatiques, Université Laval, Québec, Canada, G1K 7P4- Mohamed.bakillah.1@ulaval.ca, (Mir-Abolfazl.Mostafavi, Yvan.Bedard)@scg.ulaval.ca
[b] Centre d'information topographique de Sherbrooke, 2144 King, Canada- brodeur@nrcan.gc.ca

KEYWORDS: Data Quality, Ontology Mapping, Semantic Integration, Semantic Similarity Model, Mapping Conflict Predicates

ABSTRACT: Quality of integrated data depends on the quality of original data sources, but it can also be affected by the semantic mapping process between ontologies of different sources, thus influencing the quality of querying between multiple data sources. Being aware of semantic mapping quality could help interpreting mapping results in order to obtain better integration of heterogeneous data and would provide higher data quality to users. The question that is still unanswered is how semantic mapping quality can be defined and represented. In this paper, we propose a conceptual framework for characterising semantic mapping quality, which includes a metamodel showing relationship between semantic mappings and their quality aspects, as well as original definitions for characteristics of mapping quality. We define several Mapping Conflict Predicates that can be used to detect incoherence between mappings. We also propose a new semantic model of mapping that includes the different characteristics of mapping quality, which we called semantic model of quality mapping.

1. INTRODUCTION

Recent advances in spatial information technologies and the increasing number of spatial data sources available to users emphasize the importance of spatial data integration, which becomes even more important in the context of spatial decision making where a fast and effective processing of data is necessary. The quality of integrated data depends on approaches that are used to resolve heterogeneities between data coming from different sources. For semantic integration of heterogeneous geospatial data, many semantic models of mapping were proposed to establish semantic relationships between ontologies describing these sources (Noy and Musen, 2001; Maedche and Staab, 2002; Doan *et al.*, 2004; Mostafavi, 2006) or between schemas (Do and Rahm, 2001; Madhavan *et al.*, 2001; Berlin and Motro, 2002). Evaluation approaches used to determine the validity of these semantic models of mapping showed that they achieve a variable performance (Do *et al.*, 2003). In fact, they are adapted to specific situations, i.e. different structures of schema or representations of concepts in geospatial databases, etc. Nevertheless, the result of the semantic mapping process has a significant impact on decision making since it takes part in the query processing between multiple sources (Bouquet *et al.*, 2005). Consequently, it can affect the quality of data that will be provided to users. The user who is unaware of the quality of the semantic mapping process is unable to judge the quality of data which results from the semantic integration process. In this paper, we present a conceptual framework where we define and represent semantic mapping quality. This framework identifies and defines the multiple characteristics of mapping quality such as precision, coherence of mapping, etc. We also propose a semantic model of mapping that explicitly includes these characteristics, which we called semantic model of quality mapping. The content of this paper is structured as follow: section 2 gives the motivation of our research. Section 3 is a review of existing research related to ontology mapping and data quality. Section 4 presents our approach and a metamodel for mapping quality. In section 5 we propose the semantic model of quality mapping between ontologies. Section 6 presents the conceptual framework for mapping quality. Section 7 concludes this paper.

2. MOTIVATION

Ontology integration is the process of forming an ontology for a given subject by the re-use of several ontologies describing different subjects (Sofia Pinto and Martins, 2001). It generally involves a semantic mapping process, which consists in identifying a formal expression describing the semantic relationship between concepts of different ontologies (Bouquet *et al.*, 2005). It is known that the quality of data cannot be guaranteed following the integration, since it depends on each source (Wand and Wang, 1996). We also argue that the quality of data is affected during the integration process since mappings are used to rewrite a query on a first source for another source (Bouquet *et al.*, 2005).

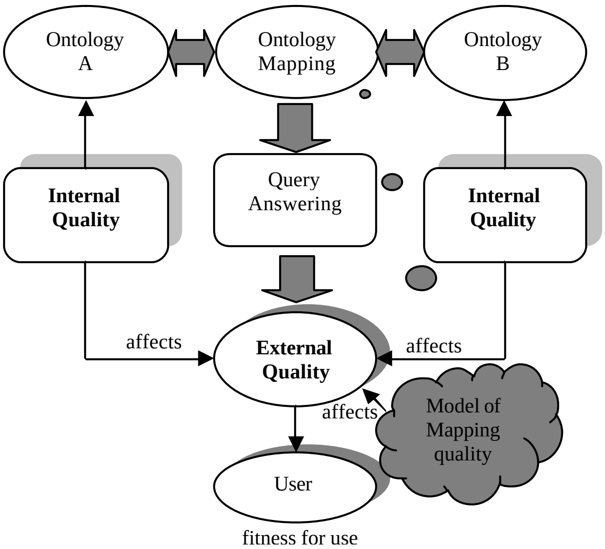

Figure 1: Impact of semantic model of mapping quality on external quality

Consequently, the external data quality (i.e. quality perceived by users, also called fitness for use) is affected not only by internal data quality (accuracy, consistency, actuality, etc.) but also by semantic mapping quality (Figure 1). Mapping quality can also play a significant role in interpreting results of the integration process. Suppose that a user wishes to identify the concept of an ontology *A* that is the most similar to concept *b* from ontology *B*. A semantic model of mapping identifies the semantic relationship between concepts of ontology *A* and concept *b* (Table 1). The semantic relationship can be quantitative, that is, it can be given by the degree of semantic similarity (second column of Table 1). A zero value means that concepts are completely dissimilar while a value of 1 indicates that they are equivalent. The semantic relationship can also be described qualitatively (third column of Table 1); in this case, it indicates the nature of the relationship between the compared concepts: equivalence, inclusion, disjunction, weak or strong overlaps, etc.

Concept of ontology *A*	**sim(*b*, a_i)**	**Nature of relationship**
a_1	0,66	strong overlap
a_2	0,24	weak overlap
a_3	0,85	a_3 includes *b*
a_4	0,88	*b* is included in a_4
...	...	...
a_n	0,00	a_n disjoint from *b*

Table 1: Example of semantic mapping results

According to the similarity results of Table 1, it could be assumed that concept a_4 is the most similar to *b*, since it has the higher semantic similarity value. But without information on the mapping quality, it can be arbitrary to conclude that a_4 is indeed the most similar to *b*. For example, it is possible that the mapping between a_4 and *b* involved a loss of precision or it may be based on incomplete data (for example, the definition of the concept a_4 is incomplete). This example illustrates that a model for the evaluation of mapping quality can help in the interpretation of results of the mapping process. However, evaluation methods for the quality of the semantic mapping process focus towards a global performance evaluation, generally using precision and recall metrics, and f-measure and overall-measure which are functions of the formers (Do *et al.*, 2003). These metrics are based on the comparison of the set of automatically computed mappings and the set of reference mappings, i.e. the real correspondences identified manually by experts of the domain (respectively set *A* and set *R* on Figure 2).

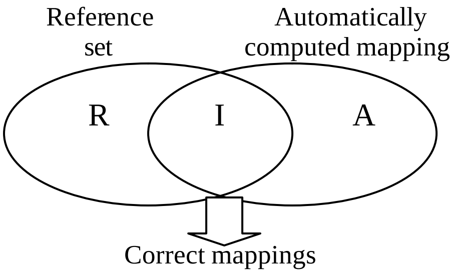

Figure 2: Sets used to evaluate quality of the mapping process.

Precision is the proportion of correct mappings in the set A (*card(I) / card(A)*) while recall is the proportion of reference mapping that were identified (*card(I) / card(R)*) (where *card* means the number of items in the set). However, these metrics do not tackle the question of the quality of the individual mapping. Also, they cannot be used when no reference is available. Actually, we do not know any method to evaluate the intrinsic quality of an individual mapping.

3. ONTOLOGY MAPPING AND DATA QUALITY

3.1 Research on Semantic Mapping between Ontologies

The problem of ontology mapping consists in determining the semantic correspondences between two ontologies. Mapping methods can differ according to the input data they use (for example, using instances or not), characteristics of the reconciliation process (for example, use of external resources such as a thesaurus) and results produced (for example, a quantitative or qualitative semantic relationship) (Bouquet *et al.*, 2005). Several approaches use a semantic similarity model to identify relationships between concepts of ontologies (Do and Rahm, 2001; Madhavan *et al.*, 2001; Maedche and Staab, 2002; Mostafavi, 2006). Semantic similarity expresses commonality between concepts. It supports the identification of concepts that have a similar meaning and refer to a similar entity of the reality. For example, concepts can be similar if they share common properties, or if they have common subsuming concepts in their respective hierarchy. Similarity models can be employed to compute the similarity between texts (i.e. between names or descriptions of concepts) with metrics such as *edit distance* (Giunchiglia and Yatskevich, 2004). In this case, similarity decreases with the number of operation required to transform one string into the other. Another technique often used is the graph-based technique, which consists in regarding the ontology as a graph and comparing positions of concepts in their respective graphs (Madhavan *et al.*, 2001) or finding similar relationships between concepts (Maedche and Staab, 2002). Concepts of taxonomy are considered as semantically similar if they are close to each other in the graph. The semantic similarity can also be evaluated by comparing common and exclusive properties of concepts (Rodriguez and Egenhofer, 2003). In this model, properties of concepts can be attributes, parts or functions. The more properties the concepts share, the more they are similar. Mapping approaches called models-based approaches aim at expressing relationships of equivalence, inclusion, intersection, disjunction, etc. between concepts of different ontologies (Bouquet *et al.*, 2003). Semantic relationships can also be established with geosemantic proximity (Brodeur and Bédard, 2001), which identify relationships between concepts by analogy with the topological model of Egenhofer (1993). Several approaches employ composite strategies, in opposition to single strategies which employ only one similarity model (Do and Rahm, 2001; Doan *et al.*, 2004). Learning techniques also imposed themselves for the automation of the mapping process since integration is often a repetitive task (Doan *et al.*, 2004). Mapping methods are also proposed to relate schemas of multidimensional geospatial databases, where different hierarchy levels of a spatial multidimensional database structure were considered (Bakillah *et al.*, 2006). Considering these semantic models of mapping, it arises that each of them will provide different results depending on their characteristics. Therefore, mapping quality may vary with different situations.

3.2 Research on Data Quality

Mapping quality is related to data quality since the mapping process uses input data for which quality is also variable. Several frameworks on data quality were proposed. Wand and Wang (1996) categorize quality dimensions according to internal quality view (which is related to the design, and includes correctness, completeness, precision, etc.) or according to external quality view (related to the use and the value of data, including relevance, utility, level of detail, accessibility, etc). A similar classification identifies intrinsic quality, contextual quality and reputational quality (Stvilia *et al.*, 2004). Metrics were proposed to evaluate some quality dimensions, for example, completeness can be measured by the number of incomplete items on the total number of items, one item being an attribute, a class, etc.(Pipino *et al.*, 2002). Another important aspect in data quality is the development of a cycle of management of the quality which includes definition of quality, definition and evaluation of a quality measure, analysis of results and the proposal of actions to improve quality of data (Wang, 1998). In a decisional context, one of the issues related to the management of data quality is its communication to users in order to avoid misuse of data. By comparing metadata provided by the producer and user's needs, indicators can be developed which describe the quality of geospatial data at various levels of detail (Devillers *et al.*, 2005). The integration of warning systems for SOLAP (Spatial On-Line Analytical Processing) applications allows informing users of elements that could be problematic in the analysis of geospatial data (Lévesque *et al.*, 2006). At the ontological level, definition of rules on the inconsistency of specifications allows to constitute a method to evaluate quality of spatial databases (Mostafavi *et al.*, 2003). Data quality can also be related to quality of sources according to various dimensions, for example comprehensibility, extent, availability, response time and cost of queries (Naumann, 1998). However, all these frameworks for quality do not consider how data quality can be modified when it undergoes semantic mapping process.

4. THE APPROACH

4.1 Overview of the Proposed Approach

The proposed approach consists in developing a conceptual framework for mapping quality, including a semantic model of mapping between ontologies. This model will include characteristics of mapping quality so we called it a semantic model of quality mapping. This section presents a metamodel for mapping quality, which shows how aspects of quality are related to the semantic mapping process. Then we propose the new semantic model of quality mapping. Finally, we develop a conceptual framework in which we give original definitions of characteristics of mapping quality.

4.2 A Metamodel for Semantic Mapping Quality

Figure 3 shows the metamodel for mapping quality which clarifies relationships between the different entities and processes implied in mapping quality. The metamodel uses UML to define relations of aggregation, generalisation and association between classes. The *model of quality mapping* is composed of the *semantic model of mapping* and of the *mapping quality category* (Figure 3). The general class *mapping quality category* is specialized in three categories: *quality of input*, *quality of output* and *quality of mapping process*. *Quality of input* is defined by *quality of definition of concept*, which is a general class for three characteristics of quality: *informativeness*, *uncertainty* and *fuzziness*, to which we can add *accuracy* and *consistency* that characterise concepts of source and target ontologies. On the other hand, *quality of mapping process* is a general class for *precision* and *completeness* of the *semantic model of mapping*. Finally, *quality of output* is determined by *coherence* and *consistency* of *mappings*, which are automatically generated by the *semantic model of mapping*. The latter is composed by *semantic similarity model* (it could be composed of several similarity models) and by the *representation of concept*. For example, concepts may be represented as nodes in the ontology graph. Some characteristics of semantic mapping quality depicted in this metamodel will be defined in section 6. First, we will present the semantic model of quality mapping.

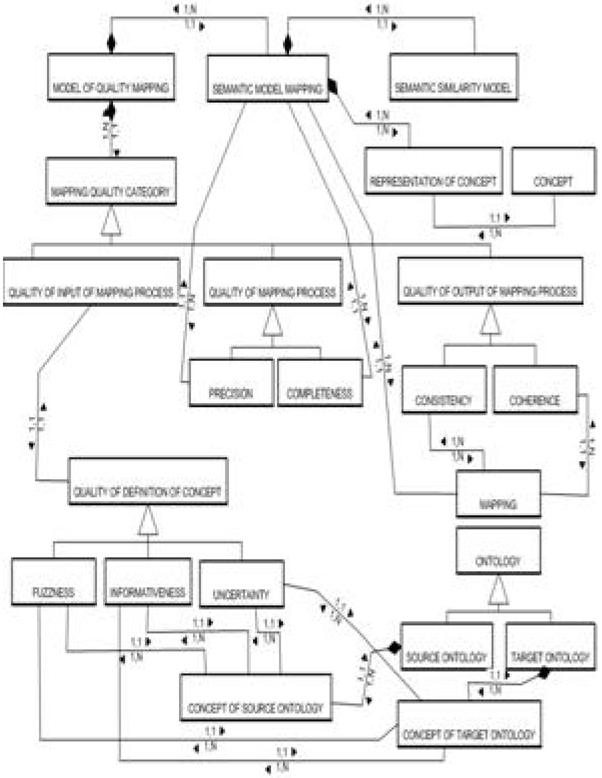
MODEL OF QUALITY MAPPING
SEMANTIC MODEL MAPPING
SEMANTIC SIMILARITY MODEL
MAPPING QUALITY CATEGORY
REPRESENTATION OF CONCEPT
CONCEPT
QUALITY OF INPUT OF MAPPING PROCESS
QUALITY OF MAPPING PROCESS
QUALITY OF OUTPUT OF MAPPING PROCESS
PRECISION
COMPLETENESS
CONSISTENCY
COHERENCE
MAPPING
ONTOLOGY
QUALITY OF DEFINITION OF CONCEPT
FUZZINESS
INFORMATIVENESS
UNCERTAINTY
SOURCE ONTOLOGY
TARGET ONTOLOGY
CONCEPT OF SOURCE ONTOLOGY
CONCEPT OF TARGET ONTOLOGY
1,1
1,N

Figure 3: Metamodel for semantic mapping quality

5. SEMANTIC MODEL OF QUALITY MAPPING

This new semantic model of quality mapping includes quality characteristics in the semantic model of mapping. Section 5.1 gives the definitions for the model. In section 5.2 we present the semantic similarity model.

5.1 Theoretical Framework

Ontology. Ontology is defined by $O=(C, G, R)$ where C is a set of concepts $C=\{c_i, i=1,2,\ldots\}$, $G=\{I(c_1), I(c_2),\ldots, I(c_i),\ldots\}$ is a set of instances for each concept and R is a set of relations $r=(c_i, c_j, type\_rel)$ between concepts. Concepts can be linked by generalization, specialization, inclusion or association relations.

Definition of concepts. For each concept c, there is a function $f{:}c?\ P(c)$ that associate it to a set of features $P(c)$ of different category: internal features (P_{int}), the set of relationships to the other concepts, called relational features (P_{rel}) and external features (P_{ext}), that is, features which characterize concepts in the neighbourhood of the concept c in the ontology. Internal features are the descriptive attributes of the concepts and their name, as well as domain values associated with the attributes. The neighbourhood of a concept c in an ontology graph is the set of concepts which are situated at a distance lower than a radius k from the concept c, the distance being determined by the number of relation between concepts. As an example, consider a spatial concept c=*sanitary region* taken from an ontology of environmental health, where we consider a radius k=1:

f: sanitary region ? {P_{int}=<code, case of intoxication, death-rate, water quality, geometry>; P_{rel}=<specialises (region)>; P_{ext}=<pesticide sales,rate of toxical emanations, geometry>}.

Relation matrix. The relation matrix *M* defined by eq.(1) relates concepts c_1 and c_2 and determine the nature of their qualitative relationship, labelled *r*, as it will be shown later.

$$M(c_1,c_2)_{ij} = S_{ij}(c_1,c_2) \qquad (1)$$

Quality Mapping. Quality mapping *m* is a mapping that includes quality characteristics. It is given by a 5-tuple that relates concepts c_1 and c_2 with a global semantic similarity value *s*, a qualitative semantic relationship *r* and a set of semantic mapping quality characteristics *Q*:

$$m = (c_1,c_2,s,r,Q) \text{ with } Q = (q_1,q_2,...q_n) \qquad (2)$$

The semantic similarity model presented in the following section computes *s* and *r*; in section 6, we will define characteristics that form the quality tuple *Q*.

5.2 Semantic Similarity Model

The semantic similarity model gives two outputs; a qualitative relationship between concepts and a global value of semantic similarity. The semantic similarity is a function of the intersection of the set of features of each concept, given for each category of feature. Two kinds of terms can be distinguished in the semantic similarity measure: on the one hand, non-mixed terms S_i comparing features of the same category i, and on the other hand, mixed terms S_{ij} comparing features of different categories i and j. Non-mixed terms cannot be regarded as being equivalent to mixed terms since they indicate a higher similarity. For example, two concepts *risk factor* and *sanitary region* can have the same feature *water quality*, but this feature is a relational feature for the first concept (as *water quality* is a specialized concept of *risk factor*) whereas it is an internal feature (attribute) for the *sanitary region*. The quantity of information shared by the concepts is less than if this common feature was part of the same category (for example an internal feature) for both concepts. Moreover, each term considered in the similarity model is balanced by a weight *?$_{ij}$* which gives the importance of the categories of features *i* and *j* being compared by the similarity terms S_i and S_{ij}. The similarity between c_1 and c_2 is given by:

$$S(c_1,c_2) = \sum_i w_i S_i + \sum_i \sum_{j \neq i} w_{ij} S_{ij} \qquad (3)$$

Semantic similarity ranges between 0 (indicating completely disjoined concepts) and 1 (indicating identical concepts). We developed an approach for computing weights based on the concept of importance of information. The method for computing **non-mixed weights** considers that the weight given to a non-mixed term (i.e. first member in eq.(3)) depends on the importance of information carried by features of the category P_i, labelled *?(P_i)*. Importance of information of a feature *p_i(c)* is high if the frequency of this feature as a feature of category P_i, *freq(p_i(C) Î P_i)*, is high according to its total frequency *freq(p_i(C))*. The importance of information also depends, according to a logarithmic function, on the number of occurrence *N(p_i)* of feature p_i within *N* concepts of ontology, by considering that the more this feature is rare in the ontology, the more the importance of information is large, because it distinguish the concept *c* from other concepts. Importance of information *? (p_i(c))* for non-mixed terms is given by:

$$y\,(p_i(C))_{non\_mixed} = \frac{freq(\,p_i(C)\in P_i\,)}{freq(\,p_i(C))} \times \log\left[\frac{N}{N(p_i)}\right] \qquad (4)$$

For example, consider an ontology of 70 concepts where a feature *medical treatment* appears in 12 concepts as an internal feature and 26 times in total, the importance of information of this feature is (12/26)log(70/26)=0,1985. Thus, importance of information of feature treatment is not very high since maximal importance of information for a feature of this ontology is 1,85 (considering a feature that always appears as a feature of the same category and appears only once in the ontology). This is because *medical treatment* is a feature that characterizes a lot of concepts in the ontology, so it must not be considered as a great importance to affirm that two concepts are similar. This concept of importance of information is similar to the principle of variability presented in Rodriguez and Egenhofer (2003) but in addition, it takes into account the different categories of features. The method for computing **mixed weights** considers that the weight given to a mixed term depends on the frequencies with which the feature p_i is regarded as a feature of category P_i or of category $P_{j.}$ Importance of information *? (p_i(c))* for mixed terms is given by

$$y\,(p_i(C))_{mixed} = \frac{freq(\,p_i(C)\in P_i\,)}{freq(\,p_i(C)\in P_j\,)} \times \log\left[\frac{N}{N(p_i)}\right] \qquad (5)$$

Total importance of the information carried by a term comparing categories of features P_i and P_j of two concepts *c_1* and c_2 is given by the weighted sum of the importance of information for each feature of *c_1* and c_2 (i=j indicate non-mixed terms):

$$y\,(P_{ij}) = \frac{1}{card(P_i(c_1) \cup P_j(c_2))} \sum_i y\,(p_i(C)) \qquad (6)$$

Weights are proportional to the importance of information of terms and are balanced by the importance of the information carried by all the terms:

$$w_{ij} = \frac{y\,(P_{ij})}{\sum_{i=1}^{n} y\,(P_{ii}) + \sum_{i=1}^{n} \sum_{i \neq j} y\,(P_{ij})} \qquad (7)$$

Similarity is computed with Bayes conditional probability in order to take into account the probability that a feature can be part of a concept**:**

$$S_{ij}(c_1,c_2) = P\left[\frac{P_i(c_1) \cap P_j(c_2)}{P_i(c_1)}\right] = P(P_i(c_1) \mid P_j(c_2)) \qquad (8)$$

where *P($_i$(c_1)|P_j (c_2))* is the conditional probability of *P_i (c_1)* knowing *P_j (c_2)*. Considering that *P_i (c_1) ={p_{1i} (c_1), p $_ì$ (c_1)...., p $_{ki}$ (c_1)}* and *P_j(c_2) = {p_{1j} (c_2), p_{2j} (c_2)...., p_{mj} (c_2)}*, and using the Bayes theorem:

$$P(P_i(c_1)\,|\,P_j(c_2)) = \frac{\sum_{s=1}^{k}\sum_{t=1}^{m} P(p_{si}(c_1))P(p_{tj}(c_2))}{\sum_{s=1}^{k} P(p_{si}(c_1))} \qquad (9)$$

The *P(p(c))* probability that a feature *p(c)* is part of a concept *c* is evaluated by considering the set I_p *(c)* of instances of the concept *c* which has the characteristic *p(c)* compared to the *I(c)* set of instances of the concept *c*:

$$P(p(c)) = \frac{card(I_p(c))}{card(I(c))} \qquad (10)$$

For example, if concept *parasitic disease* has 40 instances out of 50 which have as internal feature *case of transmission*, the probability of feature *case of transmission* is 0,80. Terms of eq. 3 constitute the matrix of relations defined in eq.1. Terms of the diagonal correspond to the non-mixed terms and the terms out of diagonal correspond to the mixed terms:

$$M(c_1,c_2) = \begin{pmatrix} S_{\text{int\_int}}(c_1,c_2) & S_{\text{int\_}rel}(c_1,c_2) & S_{\text{int\_}ext}(c_1,c_2) \\ S_{rel\_\text{int}}(c_1,c_2) & S_{rel\_rel}(c_1,c_2) & S_{rel\_ext}(c_1,c_2) \\ S_{ext\_\text{int}}(c_1,c_2) & S_{ext\_rel}(c_1,c_2) & S_{ext\_ext}(c_1,c_2) \end{pmatrix} \qquad (11)$$

The relationship *r* between concepts c_1 and c_2 can be identified by examining the state of this matrix. Concepts are equivalent if *M* (c_1, c_2)= *M* (c_2, c_1) and *M* (c_1, c_2) is the identity matrix, since the non-mixed terms correspond perfectly and the mixed terms all are null. A concept c_1 is included in a concept c_2 if *M*(c_1,c_2) is the identity matrix but *M* (c_2, c_1) is a diagonal matrix with at least one value of diagonal inferior to 1. If this state is verified, it implies that the reciprocal relation holds, i.e. c_2 includes c_1. Two concepts c_1 and c_2 overlap if *M*(c_1,c_2) and *M* (c_2, c_1) have both at least one value different from 0 but none of them is the identity matrix. Finally, c_1 and c_2 are disjoint if *M*(c_1,c_2) and *M* (c_2, c_1) are both zero matrices. In the next section, we present the conceptual framework for the quality of the mapping, which will make it possible to determine *Q*.

6. A FRAMEWORK FOR QUALITY OF MAPPING

The development of a framework to evaluate mapping quality initially requires providing an adequate definition of it. This definition should be in agreement with the ISO 9000 standards (2000), which indicate that quality is "*the totality of the properties and characteristic of a product or service which influence its ability to satisfy explicit or implicit needs*". Starting from it, we can propose the following definition:

Definition 1: Semantic mapping quality. Semantic mapping quality indicates the totality of behaviours and characteristics of a mapping which influence its skill to satisfy its explicit or implicit objectives, that is, to identify the semantic relationships between entities and consequently to provide adequate information on the relationship between these entities.

Furthermore, the majority of the approaches defines quality according to a set of characteristics which constitute recognizable properties of a product (Bansiya and Davis, 2002). We adopted characteristics recognized for data quality and adapted them for semantic mapping quality mapping. Thus, our approach is anchored as much as possible within framework of existing work. We conceive that mapping quality must integrate the characteristics indicated on Figure 4.

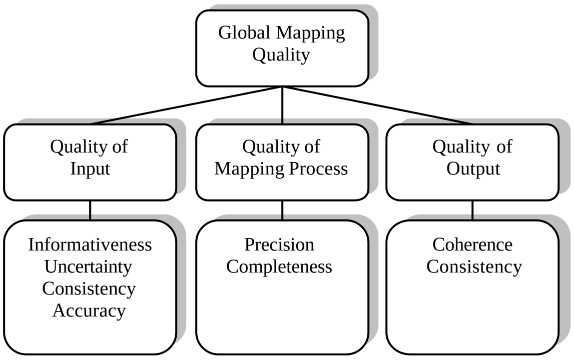

Figure 4: Structure of characteristics of mapping quality

The first level of the diagram of Figure 4 indicates the categories of characteristics of mapping quality. In the first category, mapping quality can be described by the quality of the data exploited by the mapping (*quality of input*), which can be generally related to the quality of the definition of concepts. Characteristics of that are part of the *quality of input* category are known to affect internal quality, that is, intrinsic properties resulting from data production methods (Devillers *et al.*, 2005). Mapping quality is also determined by the quality of the process, which is related to the *precision* and the *completeness* of the mapping process. It is seen that some characteristic appears in more than one category (*consistency* appears in *quality of input* and *quality of output*) since some characteristic can affect both input and output of mapping in different way. Finally, in the third category, mapping quality is affected by *quality of output*, which contains coherence of a mapping with the existing relationships in ontologies, and mapping consistency. For each one of these characteristics, we propose the following definitions.

6.1 Quality of Input

Characteristic 1: Uncertainty of mapping input. A mapping input is uncertain when it is based on an uncertain definition of the compared concepts. Uncertainty in the definition of concepts can be thematic, that is referring to a vague definition of properties of concepts, or it can be related to uncertainty on domain values of the concept's attributes. As such, uncertainty of a mapping input can be also be, in addition to thematic uncertainty, spatial or temporal uncertainty. Features are mapping input that can be uncertain because they have uncertain values. For example, the feature *cause of mortality=infectious disease* of the concept *patient* can be uncertain because of the competitive effects of the multiple possible cause of mortality.

Characteristic 2: Informativeness of mapping input. A mapping input is informative when the definition of the concepts is complete, that is there is no missing value in the definition of concepts. An incomplete definition of the concepts implied in the mapping indicates that the degree of information exploited, and thus carried by the mapping, is less.

Characteristic 3: Consistency of mapping input. A mapping input is consistent when it does not create conflict with the integrity constraints defined in the ontology. Integrity constraints give some conditions that must be verified in the ontology in order to preserve its consistency. If some of the integrity constraints were not respected when defining concepts and relations in one of the ontologies, the mapping involving these concepts and relations will be less consistent.

Characteristic 4: Accuracy of mapping input. Accuracy is related to the difference between an observed value and the real value. Consequently, accuracy of mapping input is low when the difference between the definition that is given to a concept and the reality that this concept must represent is important, for example when the value of an attribute differ from its real value. Like uncertainty of a mapping input, accuracy of a mapping input can be thematic, spatial or temporal.

6.2 Quality of the mapping process

The following characteristics are related to the quality of the mapping process. They are related to the adequacy between the properties of the semantic model of mapping and the properties of the concepts compared by the model.

Characteristic 1: Precision of a mapping. A mapping preserves the precision of the concepts when it uses their finer level of definition. For example, suppose an attribute of a concept is associated to a domain value [b1, b2]; the model of mapping is imprecise if it only evaluates the correspondence between the attributes. Consider another example where a concept is related to other concepts of the ontology by is-a and part-of relations. The mapping does not preserve the precision if it considers all the relations as being equal.

Characteristic 2: Completeness of a mapping. A mapping preserves completeness of concepts when it takes into account all the aspects of the definition of concepts. For example, consider a concept associated with a set of instances. A mapping does not preserve the completeness if it does not consider instances of concepts. Another example can be the fact that a model of mapping does not take into account relationships between concepts, but only their attributes.

6.3 Quality of Output

The next quality characteristics are related to the quality of the output of the mapping process. We consider that mapping quality must be considered from the point of view of its coherence with mappings established between other concepts. Incoherence between mappings happens when two mappings generate a conflict in the logical organisation of ontologies. Concepts of an ontology are structured by relationships (such as relations of generalisation and specialisation in a taxonomy, or relations of inclusion in a part-of hierarchy). For the purpose of this discussion, let us call these relationships between concepts of a single ontology "internal relationships". Just like internal relationships describe a logic in classification of concepts, mappings describe logical relationships between concepts from different ontologies. It may happen that two automatically generated mappings express relationships that are contradictory with the logic of the internal relationships. For example, suppose that we have concepts $\{a_0, a_1, a_2\}$ and $\{b_0, b_1, b_2\}$ respectively from ontologies of part-of relations *A* and *B* with the following internal relationships:

$$a_0 \supseteq a_1 \supseteq a_2 \text{ and } b_0 \supseteq b_1 \supseteq b_2 \qquad (12)$$

If, moreover, the following mappings were computed:

$$m = (a_1, b_1, r = \subseteq) \text{ and } m' = (a_1, b_0, r = \subseteq) \qquad (13)$$

m' would be conflicting with m since it state that b_0 is included in a_1, and thus expressing that b_0 is included in b_1, which is contradictory with internal relationships of ontology *B* expressed in equation 12, which states the reverse. Before defining coherence, we define the significance of neighbour mapping and hierarchical conflict.

Definition 2: Neighbour Mapping. Consider a mapping *m* which relates two concepts a_1 and b_1 with relation r_1, and a mapping *m'* which relates concepts a_2 and b_2 with relation r_2:

$$m = (a_1, b_1, s_1, r_1) \text{ and } m' = (a_2, b_2, s_2, r_2) \qquad (14)$$

And finally consider *dist (c,c')* the number of relations that separate concepts *c* et *c'* in the ontology graph. *m* and *m'* are neighbour mappings if *dist* (a_1,a_2)=1 or if *dist* (b_1, b_2)=1.

Definition 3: Hierarchical conflict. Consider m and m' two neighbour mappings. These mappings cause hierarchical conflict if the relationship they establish is in contradiction with the internal relationships of an ontology. We have established a set of conditions for expressing hierarchical conflict between concepts considering their internal relationships. We consider for these conditions two portions of ontologies *A* and *B* with the concepts $\{a_0, a_1, a_2\}$ and $\{b_0, b_1, b_2\}$ which respect the following internal relationships :

$$a_0 \supseteq a_1 \supseteq a_2 \text{ and } \mathrm{b}_0 \supseteq b_1 \supseteq b_2 . \qquad (15)$$

We study five cases of conflicts for each possible semantic relationship between concepts, leading to the definition of five categories of Mapping Conflict Predicates.

Category 1 (Mapping Conflict Predicates): Consider m=$(a_1,b_1,r$ = *equals*) and *m'* two neighbour mappings ; *m* and *m'* are in hierarchical conflict if one of the following conditions is checked:

$$m = (a_1, b_1, r = equals) \wedge \begin{cases} m' = (a_0, b_0, r = \{\perp\}) \\ m' = (a_0, b_1, r = \{\equiv, \subseteq, \perp, \cap\}) \\ m' = (a_0, b_2, r = \{\equiv, \subseteq, \perp, \cap\}) \\ m' = (a_1, b_0, r = \{\equiv, \supseteq, \perp, \cap\}) \\ m' = (a_1, b_2, r = \{\equiv, \subseteq, \perp, \cap\}) \\ m' = (a_2, b_0, r = \{\equiv, \supseteq, \perp, \cap\}) \\ m' = (a_2, b_1, r = \{\equiv, \subseteq, \perp, \cap\}) \end{cases} \qquad (16)$$

Category 2 (Mapping Conflict Predicates): Consider $m = (a_1, b_1, r = \subseteq)$ and *m'* two neighbour mappings; *m* and *m'* are in hierarchical conflict if one of the following conditions is checked:

$$m = (a_1, b_1, r = \subseteq) \wedge \begin{cases} m' = (a_0, b_0, r = \{\perp\}) \\ m' = (a_0, b_1, r = \{\perp\}) \\ m' = (a_0, b_2, r = \{\perp\}) \\ m' = (a_1, b_0, r = \{\equiv, \supseteq, \perp, \cap\}) \\ m' = (a_2, b_0, r = \{\equiv, \supseteq, \perp, \cap\}) \\ m' = (a_2, b_1, r = \{\equiv, \supseteq, \perp, \cap\}) \end{cases} \qquad (17)$$

Category 3 (Mapping Conflict Predicates): Consider $m=(a_1,b_1,r=\supseteq)$ and *m'* two neighbour mappings; *m* and *m'* are in hierarchical conflict if one of the following conditions is checked:

$$m=(a_1,b_1,r=\supseteq)\wedge\begin{cases} m'=(a_0,b_0,r=\{\perp\}) \\ m'=(a_0,b_1,r=\{\equiv,\subseteq,\perp,\cap\}) \\ m'=(a_0,b_2,r=\{\equiv,\subseteq,\perp,\cap\}) \\ m'=(a_1,b_0,r=\{\perp\}) \\ m'=(a_1,b_2,r=\{\equiv,\subseteq,\perp,\cap\}) \end{cases} \tag{18}$$

Category 4 (Mapping Conflict Predicates): Consider $m=(a_1,b_1,r=\cap)$ and *m'* two neighbour mappings; *m* and *m'* are in hierarchical conflict if one of the following conditions is checked:

$$m=(a_1,b_1,r=\cap)\wedge\begin{cases} m'=(a_0,b_0,r=\{\perp\}) \\ m'=(a_0,b_1,r=\{\equiv,\subseteq,\perp\}) \\ m'=(a_0,b2,r=\{\equiv,\subseteq\}) \\ m'=(a_1,b_0,r=\{\equiv,\supseteq,\perp\}) \\ m'=(a_1,b_2,r=\{\equiv,\subseteq\}) \\ m'=(a_2,b_0,r=\{\equiv,\supseteq\}) \\ m'=(a_2,b_1,r=\{\equiv,\supseteq\}) \end{cases} \tag{19}$$

Category 5 (Mapping Conflict Predicates): Consider $m=(a_1,b_1,r=\perp)$ and *m'* two neighbour mappings; *m* and *m'* are in hierarchical conflict if one of the following conditions is checked:

$$m=(a_1,b_1,r=\perp)\wedge\begin{cases} m'=(a_0,b_1,r=\{\equiv,\subseteq\}) \\ m'=(a_0,b_2,r=\{\equiv,\subseteq\}) \\ m'=(a_1,b_0,r=\{\equiv,\supseteq\}) \\ m'=(a_1,b_2,r=\{\equiv,\supseteq,\subseteq,\cap\}) \\ m'=(a_2,b_0,r=\{\equiv,\supseteq\}) \\ m'=(a_2,b_1,r=\{\equiv,\supseteq,\subseteq,\cap\}) \\ m'=(a_2,b_2,r=\{\equiv,\supseteq,\subseteq,\cap\}) \end{cases} \tag{20}$$

Characteristic 1: Coherence of a mapping

A mapping preserves coherence when it does not create hierarchical conflict with the neighbour mappings, in other words when it does not verify any of the predicates from category 1 to 5. For example, consider two spatial ontologies from the environmental health domain. The first ontology has concepts *region* $\supseteq$ *sanitary region* $\supseteq$ *medical territory* and the second ontology has concepts *region* $\supseteq$ *sanitary zone* $\supseteq$ *local medical territory.* If the mapping *m=(sanitary region, sanitary zone, r=Ê)* is computed, the mapping *m=(sanitary region, local medical territory, r=Í)* would create a hierarchical conflict since it verifies the last condition of category 3 of mapping conflict predicates.

Finally, the last characteristic we define in our conceptual framework is the consistency of a mapping.

Characteristic 2: Consistency of a mapping

A mapping preserves consistency when it does not create conflict with the integrity constraints defined in ontologies. Since a mapping establishes a relation between two concepts of different ontologies, this relation can be in contradiction with integrity constraints of one of the ontologies.

Now that we have defined characteristics of mapping quality, we can finally define completely the quality mapping *m=(c$_1$,c$_2$,s,r,Q)* of eq.2 by defining the quality tuple: *Q* = (Uncertainty of input, Completeness of input, Consistency of input, Accuracy of input, Precision, Completeness, Coherence, Consistency). We have proposed in this paper a conceptual framework that will help to define mapping quality. When combine with a semantic model mapping, we can obtain better information on the meaning of mappings and enhance the quality of the mapping process. The quality tuple can be used to determine quality of input of the mapping process, and thus it gives quality of data coming from multiples sources. The quality tuple can also be used to indicate quality of mapping process, in that case it can indicate if the model of mapping is enough precise and complete for the concepts being compared. Finally, the semantic model of quality mapping can be used to verify if the new relationships that are established between concepts of different ontologies are coherent and consistent.

7. CONCLUSION AND FUTURE WORK

Mapping quality is important because it has an impact on the quality of querying between multiple geospatial data sources. We have proposed a semantic model of quality mapping, and then we have presented a conceptual framework for mapping quality, giving original definitions for quality characteristics. We believe that this approach can eventually help to indicate to users the quality of data resulting from the semantic integration of multiple sources. In future work, we attempt to define more characteristic that can affect mapping quality and we will provide quantitative measurements for the characteristics. We will finally explore how mapping quality is related to a semantic interoperability measure.

8. REFERENCES

Bakillah, M., Mostafavi, M.A., Bédard, Y., 2006. A Semantic Similarity Model for Mapping Between Evolving Geospatial Data Cubes. In R. Meersman, Z. Tari, P. Herrero *et al.* (Editors). Lecture Notes in Computer Science 4278. Springer-Verlag, Berlin, Heidelberg, pp. 1658-1669.

Bansiya, J., Davis, C.G., 2002. A Hierarchical Model for Object-oriented Design Quality Assessment. IEEE Transactions on Software Engineering 28(1), 4-17.

Berlin, J., Motro, A., 2002. Database Schema Matching Using Machine Learning with Feature Selection. 14th International Conference on Advanced Information Systems Engineering, Toronto, Canada, pp. 452.

Bouquet, P., Serafini, L., Zanobini, S., 2003. Semantic Coordination: A New Approach and an Application. Proceedings of International Semantic Web Conference, Florida, USA, pp.130-145.

Bouquet, P., Mikalai, Y., Zanobini, S., 2005. Critical Analysis of Mapping Languages and Mapping Techniques. Technical Report DIT-05-052, University of Trento, Italy.

Brodeur, J., Bédard, Y., 2001. Geosemantic Proximity, a Component of Spatial Data Interoperability. International Workshop on Semantic of Entreprise Integration, ACM Conference on OOPSLA, Tampa Bay, Florida, pp. 14-18.

Devillers, R., Bédard, Y., Jeansoulin, R., 2005. Multidimensional Management of Geospatial Data Quality Information for its Dynamic Use within Geographical Information Systems. American Society for Photogrammetry and Remote Sensing 71(2), 205-215.

Do, H.H., Melnik, S., Rahm, E., 2003. Comparison of Schema Matching Evaluation. In A.B. Chaudhri *et al.* (Editors). LNCS 2593, Springer-Verlag, Berlin, Heidelberg, pp.221-237.

Do, H.H., Rahm, E., 2001. COMA- A System for Flexible Combination of Schema Matching Approaches. Proceedings of the 28th Conference on Very Large Data Bases Conference, Hong Kong, China, pp.610-621.

Doan, A., Madhavan, J., Domingos, P., Halevy. A.Y., 2004. Ontology Matching: A Machine Learning Approach. In Steffen Staab and Rudi Studer (Editors). Handbook of Ontologies, International Handbooks on Information Systems, Springer Verlag, Berlin, pp. 385-404.

Egenhofer, M., 1993. A Model for Detailed Binary Topological Relationships. Geomatica 47(3&4), 261-273.

Giunchiglia, F., Yatskevich, M., 2004. Element Level Semantic Matching. Proceedings of Meaning Coordination and Negotiation Workshop at International Semantic Web Conference, Hiroshima, Japan, pp. 37-48.

ISO Standard 9000-2000, 2000. Quality Management Systems: Fundamentals and Vocabulary. International Standards Organisation.

Lévesque, M-A., Bédard, Y., Gervais, M., Devillers, R., 2006. Développement d'un système d'avertissements automatiques pour diminuer les risques de mauvais usages de la donnée géospatiale décisionnelle. Colloque Géomatique 2006 – Au cœur des processus, Montréal, Canada.

Madhavan, J., Bernstein, P., Rahm, E., 2001. Generic Schema Matching with Cupid. Proceedings of the 28th Conference on Very Large Data Bases, Hong Kong, China, pp. 49-58.

Maedche, A., Staab, S., 2002. Measuring Similarity between Ontologies. Proceedings of International Conference on Knowledge Engineering and Knowledge Management, Siguenza, Spain, pp. 251-263.

Mostafavi, M.A., Edwards, G., Jeansoulin, R., 2003. An Ontology-Based Method for Quality Assesment of Spatial Data Bases. ISSDQ'04 Proceedings, Bruck am der Leitha, Austria, pp.49-66.

Mostafavi, M. A., 2006. Semantic Similarity Assesment in Support of Geospatial Data Integration. The 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Lisbon, Portugal, pp. 685-693.

Naumann, F., 1998. Data Fusion and Data Quality. In Seminar on New Techniques and Technologies for Statistics, Eurostat.

Noy, N.F., Musen, M.A., 2001. Anchor-Prompt: Using Non-Local Context for Semantic Matching. Proceedings of workshop on Ontologies and Information Sharing at International Joint Conference on Artificial Intelligence, Seattle, USA, pp.63-70.

Pipino, L.L., Lee, Y.W., Wang, R.Y., 2002. Data Quality Assessment. Communication of the ACM 45(4), 211-218.

Rodriguez, M.A., Egenhofer, M.J., 2003. Determining Semantic Similarity among Entity Classes from Different Ontologies. IEEE Transactions on Knowledge and Data Engineering 15(2), 442-456.

Sofia Pinto, H., Martins, J.P., 2001. A Methodology for Ontology Integration. 1st International Conference on Knowledge Capture, Victoria, Canada, pp.131-138.

Stvilia, B., Gasser, L., Twidale, M., Shreeves, S., Cole, T., 2004. Metadata Quality for Federated Collections. Proceedings of the International Conference on Information Quality, Cambridge, MA, pp. 111-125.

Wand, Y., Wang, R., 1996. Anchoring Data Quality Dimensions in Ontological Foundation. Communications of the ACM 39(11), 86-95.

Wang, R., 1998. A Product Perspective on Total Data Quality Management. Communication of the ACM 41(2), 58-65.