A CONCEPTUAL FRAMEWORK FOR QUALITY ASSESSMENT OF SEMANTIC MAPPING BETWEEN ONTOLOGIES

Mohamed Bakillah*, Mir Abolfazl Mostafavi**, Yvan Bédard***, Jean Brodeur****

Centre de Recherche en Géomatique, 0611 Pavillon Casault
Département des Sciences Géomatiques
Université Laval, Québec, Canada, G1K 7P4
*Mohamed.bakillah.1@ulaval.ca
** Mir-Abolfazl.Mostafavi@scg.ulaval.ca
*** Yvan.Bedard@scg.ulaval.ca
**** brodeur@nrcan.gc.ca

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

ABSTRACT: The quality of semantic mapping between different ontologies may affect significantly the quality of the integrated information from different sources. Quality of integrated information generally depends on the quality of original data sources. However, it can also be affected by the mapping process. This can ultimately influence the quality of query processing between multiple data sources. Being aware of quality of mappings could help interpreting mapping results in order to obtain better integration of heterogeneous data sources and thus providing higher information quality to end-users. Actually, the question that is still unanswered is how quality of mapping between different ontologies can be defined and represented. In this article, we propose a conceptual framework for characterising the quality of the mapping, which include a metamodel for mapping quality showing relations between mapping process and quality aspects, and original definitions for characteristics of mapping quality. We define Mapping Conflict Predicates that can be used to detect incoherence between mappings. We also propose a new model of semantic quality mapping that includes the different characteristics of mapping quality. In future works we attempt to provide measures for these quality mapping characteristics and to test our approach with different models of mapping.

1. INTRODUCTION

Recent advances in the spatial information technology and the increase in the number of spatial data sources available for the users emphasize the importance of spatial data integration and fusion. The data integration becomes even more important in the context of spatial decision making where a fast and effective processing of the available data are very important for a successful and timely decision making. The quality of integrated data depends on different approches that we use to resolve semantic, geometric and structural heterogenities between data from different sources. Within approaches for the semantic integration of different heterogeneous geospatial data sources, many semantic mapping models were proposed in the literature to establish semantic relations between entities of multiple sources, for example between the concepts of different ontologies describing these sources or between classes of schemas of geospatial data sources. The majority of evaluations approaches used to determine the validity of these semantic mapping methods showed that in general, these models achieve an acceptable performance, but it is practically impossible to attempt the result to be perfect. In fact, the proposed semantic mapping models are not necessarily good or bad, but they are rather adapted to specific situations, i.e. different structures of schema, representations of concepts, etc. Nevertheless, result of the semantic mapping process has a significant impact since it take part in the query processing between multiple sources. Consequently, they can affect the quality of the information which will be provided to the user. The user who is unaware of the quality of the semantic mapping process is unable to judge the quality of the information which results from the process of semantic integration. Moreover, as it would be unrealistic to attempt integration process to be perfect, we consider that a framework for the evaluation of the quality of semantic mappings, such as the one we will present in this paper, will help to perform a better integration by selecting the most relevant mappings and the most appropriate semantic mapping model in terms of their quality. We also consider that a model for the assessment of the quality of semantic mapping methods between different data sources can indicate the multiple characteristics of quality that are affected by the mapping process.

The content of this article is structured as follow: section 2 gives motivation of our research. Section 3 is a review of existing research related to mapping models and quality of information. Section 4 presents our approach and a metamodel for quality mappings. In section 5 we propose a model of quality mapping that is composed of a semantic similarity model for the mapping between ontologies. Section 6 presents the conceptual framework for the quality of the mappings. Section 8 concludes this article and discusses future works.

2. MOTIVATION

2.1 Ontology Mapping

Semantic heterogeneity between geospatial data from different sources constitutes one of the major problems to the integration of these data. One of the solutions to this problem consists in developing ontologies to describe semantic of the geospatial data. Ontology forms a description of an abstract model and of the concepts which are used. However, ontologies are also heterogeneous, since they often differ according to their level of abstraction, their terminology, their structure, the definition of concepts, etc. In this case, the mapping or semantic integration of ontologies is a necessary condition to semantic interoperability (Klein, 2001). The integration of ontologies is the process of forming an ontology for a given subject by the

re-use of several ontologies describing different subjects (Sofia and Martins, 2001). The integration of ontologies can be carried out by the mapping, the alignment or the fusion of ontologies, these processes representing increasing degrees of integration of ontologies. In this article, we focus explicitly on mapping approaches. The mapping of ontologies consists in identifying a formal expression which describes the semantic relation between two concepts belonging to different ontologies (Bouquet and al., 2005). Consequently, the mapping of ontologies is closely related to the concept of semantic similarity. The majority of mapping methods use a mapping model which is based on a semantic similarity model to identify the semantic relations between the entities of two ontologies (Mostafavi, 2006; Do and Rahm, 2001; Madhavan and al., 2001; Maedche and Staab, 2002). The mapping model is thus at the heart of the process of integration. However, the quality of information cannot be guaranteed following the integration process, since the quality of the data is dependent on its source and data model (Wand and Wang, 1996). In this article, we also argue that the quality of the final information provided to the user cannot be guaranteed during the integration process because the quality of the mapping model has an impact on the information provided.

2.2 Why Quality of Semantic Mapping between ontologies is Important

Several models of mapping were developed for the semi-automatic integration of ontologies. These mappings can be used for various tasks: to identify the corresponding concepts between two ontologies, to transform a data source towards another, to create a set of axioms or rules between ontologies, or to rewrite a query on a first source for another source using a *query wrapper* (Bouquet and al., 2005). For example, in spatial databases, a query submitted to the global schema of a federated database can be translated, using the mapping, in a query on the local schema of the individual sources (approach Global-as-View) or, conversely, to translate the query submitted to a local schema into a query on the global schema (approach Local-as-View). Consequently, it appears that mappings will have an impact on the quality of responses to queries, i.e. the quality of the information which results from the integration of the multiple sources (Figure 1). Actually, the external quality of information (i.e. quality perceived by the user, also called fitness for use) is affected at the same time by internal quality of the data (semantic accuracy, geometric accuracy, genealogy, actuality, etc.) and the quality of the mapping process. During the integration of multiple sources, the quality of the mapping must be indicated so that the user can be informed of the quality of the information which results from this integration.

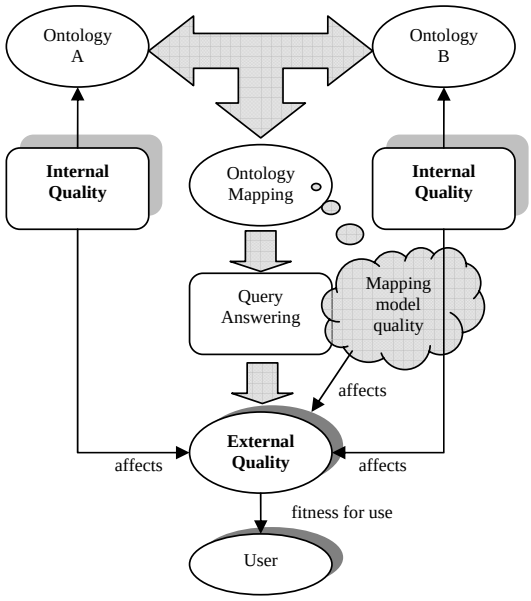

Figure 1: Impact of quality of model of mapping on external quality

The quality of mappings can also play a significant role in the interpretation of the results of the mapping process, as showed by the following example. Let us suppose that the user wishes to identify, among the concepts of an ontology A, the concept which is the most likely to correspond to the concept b coming from ontology B. The system identifies, using a model of mapping M, the semantic relations between the concepts $\{a_1, a_2\ldots, a_i\ldots, a_n\}$ of ontology A and the concept b (table 1). Dependently on the model that will be used, semantic relations can be quantitative ($sim(b, a_i) \rightarrow [\ 0,1\ ]$) or qualitative (nature of relationship).

Table 1: Example of mapping between a concept b of ontology B and concepts $a_1,\ldots a_n$ of ontology A.

Concept from Ontology A	sim (b, ai)	Nature of relationship
a1	0,66	strong overlap
a2	0,24	weak overlap
a3	0,85	a3 subsuming b
a4	0,88	b subsuming a4
...	...	...
ai	0,12	weak overlap
...	...	...
an	0,00	an disjunct from b

It could be estimated that concepts a_3 and a_4 seem to be semantically closer to concept b. But in each case, in the absence of information on the quality of the provided mappings, it can be arbitrary to conclude that a_4 is indeed the concept 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 quality of mappings must help in the interpretation of

results of the mapping process by indicating relevant quality characteristics. In the task of selecting the good mapping from a list such as shown on Table 1, the model for the quality of mappings can be use as a complementary tool to applying a threshold on semantic similarity measure, which, used alone, is not a very suitable method of selection because it force every mapping to respect a same criterion. Consequently, we make the assumption that a model of quality of mappings can help to improve the quality of the information which results from the semantic integration of heterogeneous geospatial databases.

2.3 Current State of Ontology Mapping Evaluation

Traditionally, evaluation methods for the quality of mapping process focus towards a global performance evaluation, generally using precision and recall metrics, as well as the f-measure and the overall-measure (Do and al., 2003). These metrics are based on the comparison between the set of automatically computed mappings and the set of reference mappings (the real correspondences) usually identified manually by experts of the domain (Figure 2). However, the existing evaluation methods do not tackle the question of the quality of the individual mapping.

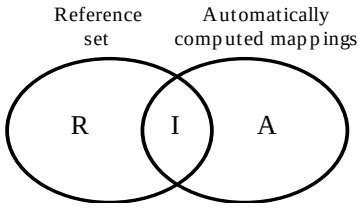

Figure 2: Sets of reference and automatically computed mappings used to evaluate the quality of the mapping process.

Therefore, they can only be used to validate a method of mapping, but not to indicate the quality of a mapping when no reference is available. Actually, we do not know any method to evaluate the intrinsic quality of an individual mapping.

3. MAPPING BETWEEN ONTOLOGIES AND DATA QUALITY ISSUES

3.1 Research on Ontology Mapping

The problem of ontology mapping consists in determining the semantic correspondences between entities of two ontologies. Techniques used for ontology mapping are very similar to those which are employed for matching schemas – which are developed in the field of the geospatial databases, but also used to map other types of models, for example XML schemas (Abels and al., 2005). Consequently, we consider that the conceptual framework which is presented in this article should be able to adapt easily to the problem of matching schemas of geospatial databases. However, some differences between schemas and ontologies must be noted (see Noy and Klein, 2003). Methods developed for the mapping can be different according to the data they use (for example, using instances or not), the characteristics of the mapping process (for example, use of external resources, such as a thesaurus or Wordnet (Miller, 1995)) and the results produced (for example, a quantitative or qualitative semantic relation) (Bouquet and al., 2005). Several approaches use a semantic similarity model to identify the semantic relations between entities of ontologies. Models of similarity can compute the similarity between texts (i.e. between names of concepts of two ontologies, or between their descriptions) with metrics such as *edit distance* (Giunchiglia and Yatskevich, 2004) or with the vector space model (Resnick, 1999). Relations between concepts of ontologies can be computed by a semantic similarity model which employs Wordnet (Rodriguez and Egenhofer, 2003; Giunchiglia and al., 2004). Another technique often used is the graph-based technique, which consists in regarding the input (schema of the ontology) as the structure of a graph and using a similarity measure which compare the positions of concepts in their respective graphs (Rada and al, 1989; Madhavan and al., 2001) or similar relations between entities (Maedche and Staab, 2002). The semantic similarity can also be evaluated by comparing the common and exclusive properties of concepts (Rodriguez and Egenhofer, 2003) according to the ratio model (Tversky, 1977). Mapping approaches called models-based approaches aim at expressing relations of equivalence, inclusion, intersection, disjunction, etc. between concepts of different ontologies (Bouquet and al., 2003). Semantic relations can also be established by means of geosemantic proximity predicates (Brodeur, 2004), in order to identify the nature of the relation between geo-concepts by analogy with the topological relations identified by Egenhofer (Egenhofer, 1991). Besides, several approaches employ composite strategies of mapping (in opposition to single strategies, which employs for example only one model of similarity) (Do and Rahm, 2001; Doan and al., 2001). Learning techniques also quickly imposed themselves for the automation of the mapping process since integration is often a repetitive task (Doan and al., 2001). For example, Automatch (Berlin and Motro, 2002) is a method of mapping where a Naïve Bayesian Learner uses the characteristics of instances to carry out the mapping between attributes of the relational schema of an individual source and the one of a global federating schema. Learning techniques which use the mappings previously identified are very promising to improve the effectiveness in the mapping between large ontologies (Aumüller and al., 2005). Methods of mapping are also proposed to relate schemas of multidimensional geospatial databases (Bakillah et al., 2006). In this last approach, we have proposed a semantic similarity model to cope with the different levels of hierarchy of multidimensional structure.

3.2 Research on Data Quality

Several frameworks on quality of information were proposed, the classification of Wand and al. (Wand and al., 1996) who categorizes dimensions of quality according to internal view (which is related to the design, for example correctness, the completeness, the precision, etc.) or according to an external view (which is related to the use and the value of information, for example relevance, contents, utility, level of detail, etc). A similar classification identifies intrinsic quality, contextual quality and reputational quality (Stvilia et al., 2004). Intrinsic quality depends little on the context and can be evaluated by measuring internal characteristics of information in relation to certain standards. Contextual quality measures the relation between information and certain aspects of its context of use, and consequently it is more context-dependant than intrinsic quality. Then reputational quality measures the position of information in an organisation and is often in relation to its origin. Some intuitive metrics were proposed to evaluate some dimensions of quality, for example the completeness can be indicated by the ratio of the number of items incomplete on the

total number of items (one item being for example an attribute, a class, etc.) (Pipino and al., 2002). These frameworks for the definition and the quality measurement, however, do not consider how the quality of information can be modified when it undergoes various processes such as semantic matching. Another important phase in the management of information quality is the development of a cycle of management of the quality which includes (1) the definition of quality, (2) the definition of a quality measure and its evaluation, (3) the analysis of the results and (4) the proposal and the implementation of actions to improve quality of information (Wang, 1998). One of the issues for the management of the quality of information in decisional context is its communication to users in order to avoid misuse of information. By comparing specifications (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 (Devilliers and al., 2002). The integration of systems of warnings for applications SOLAP (Spatial On-Line Analytical Processing) makes it possible to inform the users of certain elements which could be problematic in the analysis of the geospatial data (Lévesque et al., 2006). At the ontological level, definition of rules on the inconsistency of specifications allows to constitute a method for the evaluation of quality of the space databases (Mostafavi et al., 2003). The quality of information can also be related to quality of sources according to various dimensions, for example comprehensibility, extent, availability, response time to queries and the cost of queries (Naumann, 2005). Then several methods for decision-making can be employed to determine the source which has a higher quality, for example Data Envelopment Analysis. However, this approach does not propose method to measure the various dimensions of quality that were proposed.

4. THE APPROACH

4.1 Overview of the Proposed Approach

The proposed approach consists in developing a conceptual framework for quality of mapping, including a semantic model of quality mapping for the semantic reconciliation of two ontologies. This section first presents our metamodel for quality mappings which shows how the various aspects of quality of mapping are integrated in the semantic mapping process. Then we propose a new model of quality mapping and we propose a conceptual framework for the quality of the mappings in which we give original definitions of the characteristics of the quality of mappings between ontologies.

4.2 A Metamodel for Assessment of the Quality of Mapping between ontologies

Figure 3 shows the metamodel for quality mapping which clarifies the relations between the different entities and processes implied in global quality of mapping. The metamodel uses UML to define relations of agregation, generalisation and association between classes.

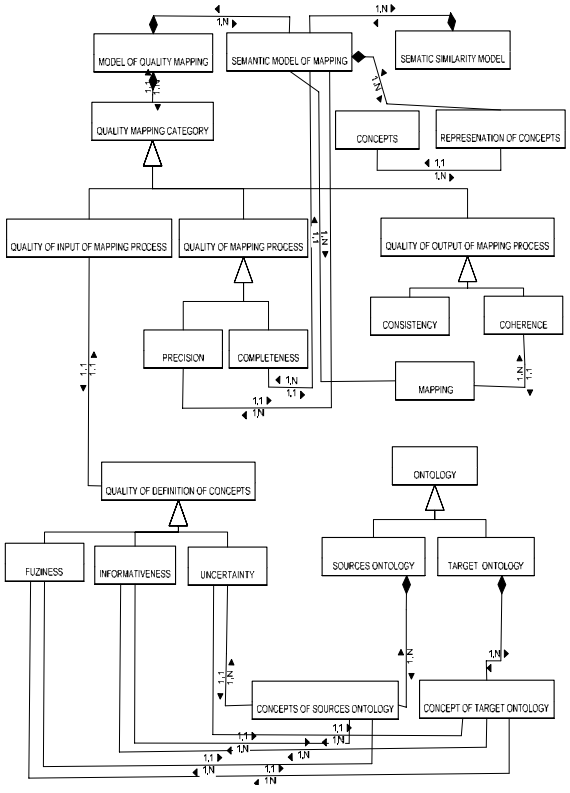

Figure 3: Metamodel for quality mappings

It is seen there that the model of quality mapping is composed of the *semantic model of mapping* and of the *quality mapping category*. The metamodel further the general class *quality mapping category* into three specialized categories of quality, namely *Quality of Input of Mapping Process*, *Quality of Mapping Process* and *Quality of Output of Mapping Process*. *Quality of Input of Mapping Process* is defined by *Quality of Definition of Concepts*, which is a general class for three characteristics of quality, namely *Informativeness*, *Uncertainty* and *Fuziness*, to which we can add *Accuracy* and *Consistency* that characterise concepts of the source and target ontologies that are to be semantically mapped. 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 of Mapping Process* is determined by *Coherence* and *Consistency* of the *Mappings*. *Mappings* are automatically generated by the *Semantic Model of Mapping*. At last, the *Semantic Model of Mapping* is composed by *Semantic Similarity Model* (in general, it could be composed of several similarity models as indicated by the cardinality of the relation) and by the *Representation of Concepts* that compose the ontologies that are to be mapped. For example, representation of concepts may be representing concepts as nodes in the graph of the ontology, or representing concepts as sets of different categories of features.

Some characteristics of quality mapping that are depicted in this metamodel will be further defined in the conceptual framework for quality mapping presented in section 6. For now, we will present in the next section the semantic model of quality mapping.

5. MODEL OF QUALITY MAPPING

The new model of quality mapping integrates the representation of the quality of mapping with the semantic model of mapping. The proposed semantic model of mapping, as shown by the metamodel, is defined by a semantic similarity model and a representation of entities of the ontologies. So in this section we will start in 4.1 by defining the representation of entities we have adopted in the theoretical framework. Since the model is designed to map different ontologies, we consider by now entities to be concepts of the ontologies. In this theoretical framework we will also define the quality of mapping in a generic way so that characteristics of quality that will be defined in the next section can be included in it. Then in section 4.2 we present the semantic similarity model we have developed and that composes the model of quality mapping.

5.1 Theoretical Framework

Ontology. Ontology O^v is defined by $O^v=(C, \Gamma, R)$ where C is the set of concepts $C=\{c_i, i=1,2,...\}$ of the ontology, $\Gamma=\{I(c_1), I(c_2),..., I(c_i),...\}$ is the set of instances for each concept c_i and R is the set of relations $r=(c_i, c_j, type\_rel)$ relating concepts of the ontology. Concepts of ontology can be linked by relations of generalization or specialisation.

Definition of the concepts. For each concept c, there is a function $f:c\rightarrow P(c)$ that associate this concept to a set of features $P(c)=\{P_1(c), P_2(c), ..., Pi(c),..., P_n(c)\}$ where $P_i(c)$ is a set of features of category i. In our model, concepts will be defined by the following sets of features, that is: internal features of the concepts (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 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 the concepts. Finally, a concept c is associated to a set of instances $\Gamma(c)=\{I_1(c), I_2(c), ..., I_i(c),...\}$ where each instance is defined by a particular association of features of c.

Relation matrix. The relation matrix relates concepts c_1 and c_2 that belong respectively to ontologies O^1 et O^2 and indicates the similarity relation $S_{ij}(c_1,c_2)$ between sets of features $P_i(c_1)$ and $P_j(c_2)$:

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

Within the framework of a qualitative model, this matrix can determine the nature of the relation between the concepts of two ontologies.

Quality Mapping. Quality mapping is a 5-tuple that relates concepts c_1 and c_2 with a global semantic similarity value s, a qualitative semantic relation r and a set of mapping quality characteristics Q:

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

The semantic similarity model presented in the following section compute the values of the similarity s and the qualitative relation r; in section 5, we will define characteristics that form the quality tuple Q.

5.2 Semantic Similarity Model

The proposed semantic similarity model establishes the semantic relations between a concept c_1 from ontology O^i and concepts c_2 from a second ontology O^j. The semantic similarity model gives two outputs; a qualitative relation between the concepts and a global value of semantic similarity. The semantic relation between c_1 and c_2 is a function of the intersection of the set of features of each concept, given for each category of feature, but also it must be balanced compared to the set of features of the referent concept c_1, so semantic similarity is asymmetric. Two cases can be distinguished concerning the sets of features considered: on the one hand, the case where the sets of compared features are the same, i.e. when $j=i$, in this case they indicate a non-mixed term S_i, and on the other hand, when the sets of compared features are different, i.e. the complementary case where $i\neq j$, in this case, they will indicate a mixed term S_{ij}. Non-mixed terms cannot be regarded as being equivalent to mixed terms since they indicate a stronger similarity between the concepts c_1 and c_2. For example, two concepts can have the same feature but this feature is a relational one for the first concept whereas it is a descriptive attribute for the second. In this case, the quantity of information shared by the concepts compared is less than if this common feature is part of the same category for both concepts. Moreover, each type term considered in the model of similarity must be balanced by a weight ω_{ij} which gives the importance of the categories of features i and j being compared. By considering the representation of the concepts adopted in this approach and proposed in the definition of concepts, the similarity between the concepts c_1 and c_2 must comprise the following terms:

$$S(c_1,c_2) = \sum_i \omega_i S_i + \sum_i \sum_{j\neq i} \omega_{ij} S_{ij} \quad (3)$$

The result of the similarity assessment gives a value ranging between 0 (indicating that concepts are completely disjoined) and 1 (indicating that concepts are identical). 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. which compares the features of the same categories P_i for concepts c_1 and c_2) in equation (8) depends on the importance of the information carried by the features of the category P_i which is designated by $\psi(P_i)$. Importance of the information of a feature $p_i(c)$ is high if this feature is regarded most of the time as a feature of category P_i among the neighbour concepts of c. Moreover, 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 specific (i.e. it appears rarely in ontology) more the importance of the information it has is large, because it distinguish the concept c from other concepts. Importance of information $\psi(p_i(c))$ for non-mixed terms is given by:

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

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 $\psi(p_i(c))$ for mixed terms is given by

$$\psi(p_i(C))_{mixte} = \frac{freq(p_i(C) \in P_i)}{freq(p_i(C) \in P_j)} \times \log\left[\frac{N}{N(p_i)}\right] \quad (5)$$

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):

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

Weights ω_{ii} for non-mixed terms and ω_{ij} for the mixed terms are proportional to the importance of information of the considered term which is balanced by the importance of the information carried by all the terms:

$$\omega_{ij} = \frac{\psi(P_{ij})}{\sum_{i=1}^{n} \psi(P_{ii}) + \sum_{i=1}^{n} \sum_{i \neq j} \psi(P_{ij})} \quad (7)$$

It should be noted that the weights are different for each couple of concept thus distinguishing how their features are categorized in another ontology. Incorporate in the similarity model the concept of Bayes conditional probability enables to consider that**:**

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

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 $P_i(c_1) = \{p_{1i}(c_1), p_i(c_1)\ldots., p_{ki}(c_1)\}$ and $P_j(c_2) = \{p_{1j}(c_2), p_{2j}(c_2)\ldots., p_{mj}(c_2)\}$, and using the Bayes theorem:

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

The P(p(c)) probability of a feature *p(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*:

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

Terms of equation 3 constitute the matrix of relations defined in equation 1, which comprises nine terms, where the terms of the diagonal correspond to the non-mixed terms and the terms out of diagonal correspond to the mixed terms:

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

This matrix of relations allows identifying the qualitative relation r which relates the concepts c_1 and c_2. by examining the state of this matrix, that is which cells are empty or non- empty. The most obvious case is that of equivalence where one obtains that M (c_1, c_2) is the identity matrix since the non-mixed terms correspond perfectly and the mixed terms all are null. Otherwise, the matrix can express inclusion between concepts, overlap or disjoint concepts. 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 the quality of the mapping initially requires providing an adequate definition of the quality of the mapping. We then identify the relevant characteristics to describe the quality of the mapping.

6.1 What is Semantic Mapping Quality

A discussion on the quality of the semantic mapping must necessarily begin with a definition of what is quality of mapping. This definition should also be in agreement with the definition established by the standards ISO 9000, 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" (ISO Standards 9000, 2000). Starting from this definition, we can propose the following definition:

Definition 1: Quality of semantic mapping between different ontologies. Quality of the semantic mappings indicates the totality of the behaviours and of the characteristics of a mapping which influence its skill to satisfy its explicit or implicit objectives, that is, to identify the semantic relations between entities and consequently to provide adequate information on the relation between these entities.

This definition suggests that a model of quality of the mapping will be able to improve the quality of the final information which will be provided to the user (external quality). However, we also conceive that mappings carry a degree of subjectivity which makes it impossible to identify the nature of a perfect mapping, since definition of concept is also subjective. Within this conceptual framework, we consider two ontologies A and B for which the respective structure can be described by a graph where the nodes are concepts and the arcs are relations between these concepts. We note $\{a_i\}$ the set of concepts of ontology A and $\{b_i\}$ the whole of the concepts of ontology B. Finally we consider a set of relations of mapping R={(a_i, b_j, s, r)} that indicates that concepts a_i and b_j are concepts related by the relation r and a value of similarity s. We consider the following relations:

$$\left\{\begin{matrix} \text{int}\,er\,\text{sec}\,ts(\cap), include\,\text{in}(\subseteq), includes(\supseteq), \\ equals(\equiv), disjo\,\text{int}\,ness(\perp) \end{matrix}\right\} \quad (12)$$

Mappings can be of simple cardinality (1:1; 1: n; n: 1) or multiple (n:m). For the moment, the mapping does not include the tuple quality of mapping Q since this one remains to be determined.

6.2 Defining Quality Characteristics for Mapping

The majority of the approaches for the evaluation of quality define the concept of quality according to a set of characteristics of quality which constitute recognizable properties of a product or a service (Bansiya and Davis, 2002). In order to determine the relevant characteristics for the description of the quality of the mapping, we analyzed the possible impacts of the process of mapping on the various characteristics of the quality of information recognized in the field of information quality. Thus, our approach tries to be anchored as much as possible within framework of existing work. As a starting point, we conceive that the quality of the mapping must integrate the characteristics indicated on Figure 4. The first level of the diagram of Figure 4 indicates the categories of characteristics of quality. In the first category, the quality of a mapping can be described by the quality of the information exploited by the mapping (*quality of input*), which can be generally related to the quality of the definition of concepts being compared by the mapping model. Characteristics of quality that are part of the *quality of input* category are characteristic which are know to affect internal quality, that is intrinsic properties resulting from data production methods (Devilliers, 2004). These characteritic need to be considered in our framework since quality of input in the mapping process will affect directly the quality of mapping.

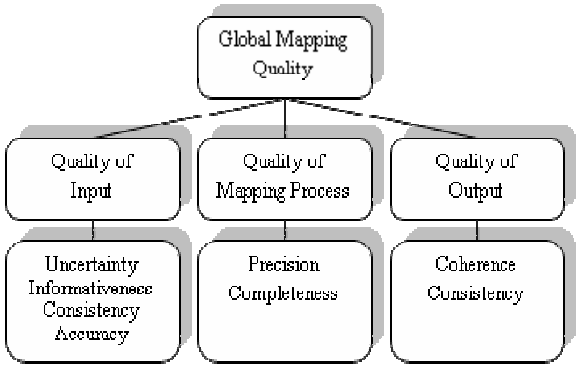

Figure 4: Structure of characteristics of the quality of mapping

Then the quality of the mapping 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 (for example, *consistency* appears in *quality of input* et *quality of output*) since, as it will be shown in our framework, some characteristic can affect both input of mapping and output of mapping in a different way. Thus, these characteristics have different definition according to the category they belong. Finally, in the third category, the quality of a mapping is measured with the *quality of output*, which contains coherence of a mapping with the existing relations in ontologies, and consistency of the mapping. For each one of these characteristics, we propose the following definitions.

6.2.1 Quality of Input

Characteristic 1: Uncertainty of a 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.

Characteristic 2: Informativeness of mapping input. A mapping input is informative when the definition of the concepts is complete, that is ther is no missing values 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 a mapping input. A mapping input is consistent when it does not create conflict with the axioms and 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 a 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.

6.2.2 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 exploited model of mapping and the properties of the entities compared by the model.

Characteristic 5: Precision of a mapping. A mapping preserves the precision of the concepts when it uses their finer level of definition.

Example1. 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. Another example is a concept is related to the 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.

The next characteristic for the quality of the mapping is the completeness, which is defined in general as the ability to represent any state of a real system (Wand et Wang, 1996). In the case of the mappings, the completeness will thus be defined by the ability of the model of mapping to represent all the states of the entities which are compared.

Characteristic 6: Completeness of mapping. A mapping preserves completeness of concepts when it takes account of all the aspects of the definition of the concepts.

Example 2. Consider a concept associated with a set of instances. A mapping does not preserve the completeness if it does not consider instances of the concepts. Another example can be the fact that a model of mapping does not take into account relations between concepts, but only their attributes.

6.2.3 Quality of Output

The next characteristic of quality are related to the third category, which is the one that assess the quality of the output of the mapping process. We consider that the quality of a mapping must be considered from the point of view of its coherence with the mappings established between other concepts. Incoherence between mappings happens when two mappings generate a conflict in the logical organisation of concepts in the ontologies. Concepts of an ontology are structured by relations (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 relations between concepts of a single ontology "internal relations". Just like internal relations describe a logic in classification of concepts, the set of mappings relations describes logical relations between concepts from different ontologies. It may happen that two automatically generated mappings express relations that are contradictory with the logic of the internal relation. 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. Let us also represent concepts by sets (Figure 5). We see on this figure that if the following internal relations hold:

$$a_0 \supseteq a_1 \supseteq a_2 \text{ and } b_0 \supseteq b_1 \supseteq b_2 \quad (13)$$

and if moreover the following mapping was computed:

$$m = (a_1, b_1, r = \subseteq) \quad (14)$$

meaning that a_1 is included in b_1, the following mapping

$$m = (a_1, b_0, r = \subseteq) \quad (15)$$

would be conflicting 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 relations of ontology B expressed in equation 13 that state that is included in b_0.

Before defining coherence, we define the significance of neighbour mapping and hierarchical conflict between mappings, giving five categories of Mapping Conflicts Predicates.

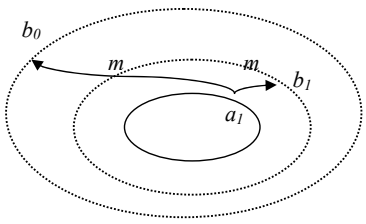

Figure 5: Incoherent mapping m and m' (equations 14 and 15) according to internal relation (equation 13).

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:

$$m = (a_1, b_1, s, r_1) \text{ and } m' = (a_2, b_2, s, r_2) \quad (16)$$

And finally consider *dist (c,c')* the number of relations (of arcs) between the concepts c et c' in the graph of an ontology. Then 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 two mappings m and m' cause hierarchical conflict if the relation they establish is in contradiction with the internal relations of an ontology, i.e. relations existing between concepts of the same ontology.

We have established a set of conditions for expressing hierarchical conflict between concepts considering their internal relations. 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 relations :

$$a_0 \supseteq a_1 \supseteq a_2 \text{ and } b_0 \supseteq b_1 \supseteq b_2 . \quad (17)$$

We study five cases of conflicts for each possible semantic relation (equation 12) between the concepts a_1 and b_1. In each case, we identify the semantic relations between neighbour concepts which are in logical conflict with the relation between a_1 and b_1 and with the conditions of equation (17). Thus we define five categories of Mapping Conflict Predicates. The Mapping Conflicts Predicates that could be detected in a set of mapping between two ontologies contribute to reduce the coherence of the mappings carried out automatically by the model of mapping.

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:

$$m = (a_1, b_1, r = equals) \wedge \begin{cases} m' = (a_0, b_0, r = \{\perp\}) \\ m' = (a_0, b_1, r = \{\equiv, \subseteq, \perp, \cap\}) \\ m' = (a_0, b_2, r = \{\equiv, \subseteq, \perp, \cap\}) \\ m' = (a_1, b_0, r = \{\equiv, \supseteq, \perp, \cap\}) \\ m' = (a_1, b_2, r = \{\equiv, \subseteq, \perp, \cap\}) \\ m' = (a_2, b_0, r = \{\equiv, \supseteq, \perp, \cap\}) \\ m' = (a_2, b_1, r = \{\equiv, \subseteq, \perp, \cap\}) \end{cases} \quad (18)$$

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:

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

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:

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

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:

$$m = (a_1, b_1, r = \cap) \wedge \begin{cases} m' = (a_0, b_0, r = \{\perp\}) \\ m' = (a_0, b_1, r = \{\equiv, \subseteq, \perp\}) \\ m' = (a_0, b2, r = \{\equiv, \subseteq\}) \\ m' = (a_1, b_0, r = \{\equiv, \supseteq, \perp\}) \\ m' = (a_1, b_2, r = \{\equiv, \subseteq\}) \\ m' = (a_2, b_0, r = \{\equiv, \supseteq\}) \\ m' = (a_2, b_1, r = \{\equiv, \supseteq\}) \end{cases} \quad (21)$$

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:

$$m = (a_1, b_1, r = \perp) \wedge \begin{cases} m' = (a_0, b_1, r = \{\equiv, \subseteq\}) \\ m' = (a_0, b_2, r = \{\equiv, \subseteq\}) \\ m' = (a_1, b_0, r = \{\equiv, \supseteq\}) \\ m' = (a_1, b_2, r = \{\equiv, \supseteq, \subseteq, \cap\}) \\ m' = (a_2, b_0, r = \{\equiv, \supseteq\}) \\ m' = (a_2, b_1, r = \{\equiv, \supseteq, \subseteq, \cap\}) \\ m' = (a_2, b_2, r = \{\equiv, \supseteq, \subseteq, \cap\}) \end{cases} \quad (22)$$

Characteristic 7: 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.

Finally, the last characteristic we define in our framework for mapping quality is the consistency of a mapping. The consistency of a mapping is related to integrity constraints that already exist in both ontologies. Integrity constraints give some conditions that must be verified in the ontology in order to preserve its consistency, for example an integrity constraint can be a relation that cannot be verified between two concepts.

Characteristic 8: Consistency of a mapping
A mapping preserves consistency when it does not create conflict with the integrity constraints define in the 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 ontology.

Now that we have defined characteristics of the quality of mapping between ontologies, we can finally define completely the quality mapping m=(c_1,c_2,s,r,Q) of equation 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 quality of mapping. When combining quality of mapping with model of 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 either 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 mapping model is enough precise and complete for the concepts being compared. Finally, the model of quality mapping can be used to verify if the new relations that are established between concepts of different ontologies are coherent and consistent.

7. CONCLUSION AND FUTURE WORK

In this article, we have argued that quality of mapping is an important feature because it has an impact on the quality of query answering over multiple geospatial data sources. We have proposed a model of quality mapping, and then we have presented a conceptual framework for quality mapping, giving original definitions for quality characteristics of mappings. We believe that this approach can lead to better results of mapping and can also indicate to users the quality of information resulting of semantic integration of multiple sources. In the future work, we attempt to define more characteristic that can affect quality of mapping and we will provide quantitative measurements for characteristics of quality mapping and make a comparative study of our quality model in a concrete application domain. We will also extend this model so it can take into account temporal mapping in the context of mapping maintenance in a dynamic network of heterogeneous geospatial data sources. We will finally explore how the quality of mapping can be related to a measure of semantic interoperability.

8. REFERENCES

Abels, S., Haak, L., Hahn, A., 2005. Identification of Common Methods Used for Ontology Integration Tasks. *IHIS'05*, Bremen, Germany.

Aumüller, D., Do, H.H., Massmann, S., Rahm, E., 2005. Schema and Ontology Matching with COMA++. In *Proceedings on International Conference on Management of Data, Software Demonstration.*

Bakillah, M., Mostafavi, M.A., Bedard,Y. (2006). A Semantic Similarity Model for Mapping between evolving Geospatial Data Cubes, *OTM Workshops 2006*, LNCS 4278, pp.1658 – 1669.

Berlin, J., Motro, A., 2002. Database Schema Matching Using Machine Learning with Feature Selection. *CAiSE 2002*.

Bouquet, P., Serafini, L., Zanobini, S., 2003. Semantic Coordination: A New Approach and an Application. In *Proceedings of International Semantic Web Conference*, 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. (2004) Interopérabilité des Données Géospatiales : Élaboration du Concept de Proximité Géosémantique. Thèse de doctorat, Université Laval.

Devilliers, R. (2004) Conception d'un Système Multidimensionnel d'Information sur la Qualité des Données Géospatiales. Thèse de Doctorat, Université Laval & Université de Marne-La-Vallée, 167 pages.

Do, H.H., Melnik, S., Rahm, E., 2003. Comparison of Schema Matching Evaluation. A.B. Chaudhri et al. (Eds.): *Web Databases and Web Services 2002*, LNCS 2593, pp.221-237.

Do, H.H., Rahm, E., 2001. COMA- A System for Flexible Combination of Schema Matching Approaches. In *Proceedings of Very Large Data Bases Conference*, p.610-621.

Doan, A., Domingos, P., Halevy. A.Y., 2001. Reconciling Schemas of Disparate Data Sources: A Machine Learning Approach. In *Proceedings of the CAN SIGMOD Conference 2001*.

Giunchiglia, F., Yatskevich, M., 2004. Element Level Semantic Matching. In *Proceedings of Meaning Coordination and Negotiation Workshop at International Semantic Web Conference*.

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

Klein, M., 2001. Combining and Relating Ontologies: An Analysis of Problems and Solutions. In: Gomez-Perez, A., Gruninger, M., Stuckenschmidt, H., Uschold, M. (Eds.): *Workshop on Ontologies and Information Sharing*, Seattle, USA.

Levesque, M-A., Y. Bédard, M. Gervais & R. Devillers. (2006) *Développement d'un système d'avertissements automatiques pour diminuer les risques de mauvais usages de la donnée géospatiale décisionnelle*. Colloque Géomatique 2006 – Au coeur des processus, 25-26 octobre, Montréal, Canada.

Madhavan, J., Bernstein, P., Rahm, E., 2001. Generic Schema Matching with Cupid. In *Proceedings of Very Large Data Bases Conference*, pp. 49-58.

Maedche, A., Stabb, S., 2002. Measuring Similarity between Ontologies. In *Proceedings of International Conference on Knowledge Engineering and Knowledge Management*, pp.251-263.

Miller, A.G., 1995. Wordnet: A Lexical Database for English. *Communications of the ACM*, 38(11), pp. 39-41.

Mostafavi, M.A., Edwards, G., Jeansoulin, R. (2003) An Ontology-Based Method for Quality Assesment of Spatial Data Bases.

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

Noy, N.F., Klein, M., 2003. Ontology Evolution: Not the Same as Schema Evolution. Knowledge and Information Systems, Vol. 5.

Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1), pp.17-30.

Resnick, P., 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* Vol.11, pp.95-130.

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), pp. 442-456.

Shannon, C.E., 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, Vol.21, pp. 379-423.

Sofia, H., Matrins, J.P., 2001. A Methodology for Ontology Integration. In: *Proceedings of the International Conference on Knowledge Capture*, ACM SIGART.

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.

Tversky, A. (1977) Features of Similarity. *Psychological Review* 84(4): 327-352.

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