A CONCEPTUAL FRAMEWORK FOR DEFINITION AND ASSESSMENT OF QUALITY OF SEMANTIC MAPPING

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

Chaire Industrielle CRSNG en bases de données géospatiales décisionnelles
Centre de Recherche en Géomatique, 0611 Pavillon Casault, Département des Sciences Géomatiques Faculté de Foresterie et de Géomatique
Université Laval, Québec, Canada, G1K 7P4
***Mohamed.bakillah.1@ulaval.ca**
**** Mir-Abolfazl Mostafavi @scg.ulaval.ca**
***** Yvan Bédard @scg.ulaval.ca**
******brodeur@nrcan.gc.ca**

KEYWORDS: Quality, Ontology Mapping, Semantic Integration, Semantic Similarity Model, Quality Measure.

ABSTRACT: Quality of information is one of the goals of semantic integration, which aim at achieving semantic agreement between heterogeneous and multiples data sources. Quality of information generally refers to quality of data, however, quality of information can also be affected by the mapping process and ultimately it will affect quality of query processing between multiples sources. In this article, we propose a conceptual framework for characterising quality of 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 also propose a model of quality mapping that will include the quantitative measurements of the different characteristics of mapping quality. Finally, we show results of our approach with two model of mapping.

1. INTRODUCTION

Within approaches for the semantic integration of heterogeneous data sources, many models of mapping were proposed in the literature to establish semantics relations between entities of multiples sources, for example between the concepts of different ontologies which describe these sources or between classes of schemas of the data sources. The majority of evaluations carried out to show the validity of these methods of mapping shows that in general, these models achieve an acceptable performance, but it is practically impossible to attempt the result to be perfect. In fact, models of mapping that are proposed 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 mapping process has a significant impact given that they take part in the query processing between multiple sources. Consequently, they can affect the quality of the information which will be ultimately provided to the user, who is unable to judge quality of the information which results from the process of semantic integration. Moreover, as it would unrealistic to attempt information to be perfect, we consider that a framework for the evaluation of the quality of mappings will help on the one hand to perform a better integration by selecting the most relevant mappings in term of quality. We also consider that a model for quality of mappings can indicate the multiple characteristics of quality that are affected by the mapping process.

1.1 Ontology Mapping

The semantic integration of data in a system of heterogeneous and multiples databases constitutes one of the major problems to consider to establish semantic interoperability in such a system. One of the solutions to this problem consists in developing ontologies to describe the context of use of the data. Ontology forms a description of an abstract model and of the terms 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 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 one or 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. The mapping of ontologies consists in identifying a formal expression which describes the semantic relation between two entities 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 methods of mapping uses a model of mapping which is based on a semantic similarity model to identify the semantic relation between the entities of two ontologies (for example, Do and Rahm, 2001; Madhavan and Al, 2001; Maedche and Staab, 2002). The model of mapping is thus at the heart of the process of integration. However, the quality of information cannot be guaranteed following the process of integration, because 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 process of integration because the quality of the model of mapping will have an impact on provided information.

1.2 Why Mapping Quality 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, a query submitted to the global schema of a federate 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 answers to queries, i.e. the quality of the information which results from the integration of the multiple sources. Actually, the external quality of information (quality perceived by the user) is affected at the same time by internal quality (data, definition of the concepts, relations of ontology, etc.) and the quality of the mapping process (figure 1).

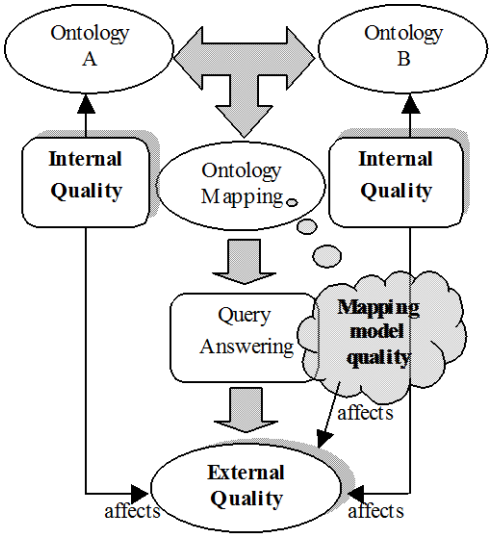

Figure 1 : Impact of quality of model of mapping on external quality

During the process of 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. The quality of mappings can also play a significant role in the interpretation of the results of the mapping process, as shows the following example. Let us suppose that the user wishes to identify, among the concepts of an ontology A, the concept which is the more 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 M that will be used, the identified semantic relations can be quantitative (sim(b, a_i) $\rightarrow$[0,1]) or qualitative (we illustrated these two cases in this example).

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 is based on incomplete data (for example, the definition of the concept a_4 is incomplete). This example illustrates that a model of quality of the mappings must help in the interpretation of results of the mapping process. For example, a model of quality of mappings can assist in the interpretation of the results in indicating, by means of relevant characteristics of quality, if a correspondence between two concepts must be retained or rejected. The model of quality can be complementary to applying threshold on semantic similarity measure, which, used alone, is not a very delicate method of selection. Consequently, we make the assumption that a model of quality of mappings can help to improve quality of the information which results from the semantic integration of heterogeneous data sources.

1.3 Current State of Ontology Mapping Evaluation

Traditionally, methods of evaluation of the quality of the mappings focus towards a global performance evaluation of the mapping process, generally by using the well-known precision and recall metrics, as well as f-measure and overall-measure (Do and al., 2003). These metrics are based on a set of reference mappings (the real correspondences) generally manually identified by experts in the field represented by ontologies (figure 2).

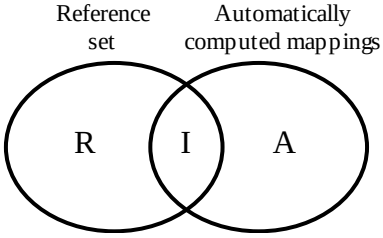

Figure 2: Sets of references mappings and automatically computed mappings.

$$precision = \frac{card(I)}{card(A)};\ recall = \frac{card(I)}{card(R)}; \qquad (1)$$

$$F-measure = \frac{precision\ recall}{(1-\alpha)\cdot precision + \alpha \cdot recall};$$
$$with\ \alpha\ a\ constant\ that\ weights\ precision\ and\ recall \qquad (2)$$
$$Overall = recall \cdot \left(2 - \frac{1}{precision}\right)$$

However, the existing evaluations do not tackle the question of the quality of the individual mapping. Therefore, they can be used only to validate a method of mapping, but not to indicate the quality of a mapping when no reference is available. Currently, we do not know any method to evaluate the intrinsic quality of the individual mapping that could refer to the quality of input and impact of the mapping process on the quality of information.

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

for the quality of the mappings and section 6 proposes measures to evaluate the different characteristics of the quality of mappings. Section 7 presents an application of our approach. Section 8 concludes this article and discusses future works.

2. REVIEW OF RESEARCH

2.1 Research on Ontology Mapping

The problem of mapping (or matching) of ontologies consists in determining the semantic correspondences between entities of two ontologies. Techniques used for mapping of ontologies are very similar to those which are employed for matching schemas – which are developed in the field of the 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 will be developed in this article should be able to adapt easily to the problem of matching schemas of databases. However, some differences between schemas and ontologies must be noted (see Noy and Klein, 2003), for example the fact that ontologies represent semantics explicitly while schema usually does not.

Methods developed for the mapping can be different according to the data they use (for example, using instances or not), characteristics of the mapping process (for example, use of external resources, such as a thesaurus or Wordnet (Miller, 1995)) and 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 and to establish the mapping between entities of ontologies. Models of similarity can compute the similarity between texts (that is, between names of concepts or classes 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 the concepts of ontologies can be computed by a semantic similarity model which employs relationships with concepts of Wordnet (Rodriguez and Egenhofer, 2003; Giunchiglia and al., 2004). Another technique often used in methods of mapping is the graph-based technique, which consists in regarding the input (schema of database, ontology) as the structure of a graph and using a similarity measure which compare the positions in the 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).

Several approaches employ composite strategies of mapping (in opposition to single strategies, which employs for example only one model of similarity) –i.e. they combine several techniques adapted to various cases to compute mapping relations. (Do and Rahm, 2001; Doan and al., 2001; Bakillah et al, 2006). Learning techniques also quickly imposed themselves for the automation of the mapping process since integration often involves relating sources describing the same field and becomes, consequently, 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 schema (federator). The bayesian learner allows to compute probabilities of correspondence and mismatch for all attributes of the schema of an individual source, and the sum of the probability is maximized to determine mappings, which are of cardinality 1 :1 (i.e. an attribute can be associated to at most one other attribute). Learning techniques which use the mappings previously identified and decompose the problem of mapping between two ontologies into several small problems 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).

Finally, approaches of mapping called models-based approaches aim at expressing qualitative relations (equivalence, inclusion, intersection, disjunction…) between the concepts of various ontologies (Bouquet and al., 2003). Semantic relations can also be established by means of qualitative predicates of geosemantic proximity (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). These qualitative approaches of mapping, compared to the quantitative approaches which give as a result a single numerical value, have the advantage of providing a higher degree of expressivity by indicating the nature of the relation between the concepts. For example, the fact that two concepts are, according to a given model of semantic similarity, similar to 76,49% indicates little information to express in what these concepts are similar (which type of properties do they share ? Is a concept more general than another, or conversely?) However, they also present a disadvantage in the classification of the concepts according to their degree of similarity compared to a concept of reference. Actually, we consider that a method of mapping which gives a qualitative and quantitative result provides more complete information.

2.2 Research on Data Quality

Several frameworks on quality of information were proposed, for example the framework of Wang and al. (Wang and al., 1995), where definitions of quality dimensions are proposed This work has lead to several other subsequent work, for example 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 standard. 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; in particular we are interested to seek if methods of mapping can affect quality.

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 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).

In a context of interoperability between multiple heterogeneous data sources, 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 various dimensions of quality proposed.

3. THE APPROACH

3.1 Overview of the Proposed Approach

The suggested approach consists in developing a semantic model of quality mapping for the semantic reconciliation of two ontologies. Initially, it consists in developing a metamodel for quality of mappings which shows how the various aspects of quality of mapping are integrated in the semantic mapping process between two ontologies. Then we propose a model of quality mapping and describe how this model will integrate different characteristics of the quality of mappings to constitute a semantic model of quality mapping. Then, we propose a conceptual framework for the quality of the mappings in which original definitions of the characteristics of the quality of mappings are proposed. The last stage of the development of our approach consists in giving measures to evaluate quantitatively the characteristics of quality suggested within the conceptual framework. These measurements are finally integrated into the semantic model of quality mapping for the realisation of our final objective.

3.2 A Metamodel for Quality of Mapping

Figure 3 shows the metamodel for quality of mappings which clarifies the relations between the different entities and processes implied in global quality of mappings. It is seen there that the model of quality mapping is composed of the semantic model of mapping and of the representation of the quality of the mappings.

The metamodel further refines *Representation of Quality of Mappings* into three categories of quality, namely *Quality of Input in Mapping Process*, *Quality of Mapping Process* and *Quality of Output of Mapping Process*. *Quality of Input in Mapping Process* is defined by *Quality of Definition of Entities*, which is measured by two characteristics of quality, namely *Informativeness* and *Uncertainty* that characterise the source and target ontologies that are to be semantically mapped. On the other hand, *Quality of Mapping Process* is measured by *Precision* and *Completeness* of the *Semantic Model of Mapping*. Finally, *Quality of Output of Mapping Process* is determined by *Coherence* over the *Set of Mapping*, which is automatically generated by the *Semantic Model of Mapping*.

At last, the *Semantic Model of Mapping* is defined by a *Semantic Similarity Model* and also depends on the *Representation of Entities* that compose the ontologies that are to be mapped. For example, representation of entities may be representing concepts as nodes in the graph of the ontology, or representing concepts as sets of different categories of features.

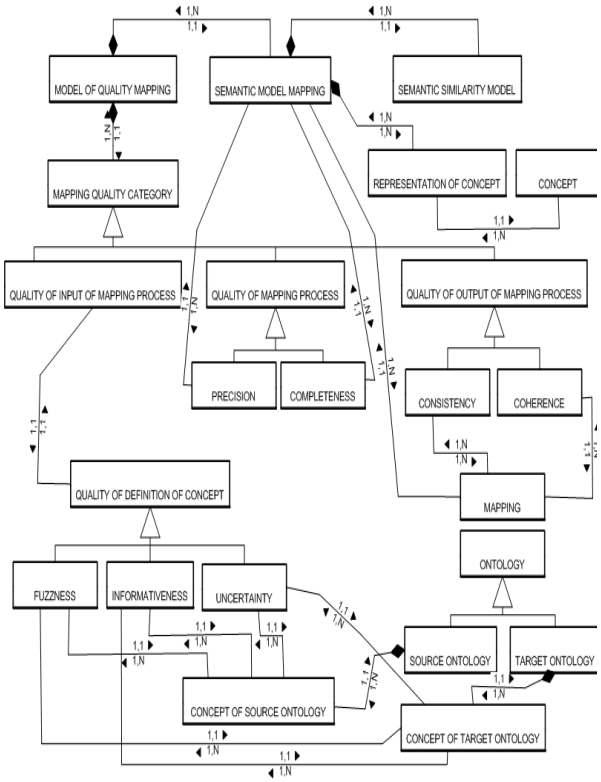

Figure 3: Metamodel for quality of mappings

The different characteristics of quality of mapping that are depicted in this metamodel will be further defined in the conceptual framework for quality of mapping presented in section 5. For now, we will present in the next section the semantic model of mapping.

4. MODEL OF QUALITY MAPPING

The model of quality mapping integrates the representation of the quality of mappings with the semantic model of mapping. On the other hand, the proposed semantic model of mapping, as shown on the metamodel for quality of mapping, 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 in the theoretical framework the representation of entities we have adopted. 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 mappings in a generic way so that characteristics of quality that will be defined and measured in the subsequent sections can be included in it. Then in section 4.2 we present the semantic similarity model that composes the model of quality mapping.

4.1 Theoretical Framework

This section presents the basic definitions on which our approach is based, that is the definition of ontology, the concepts and the relation matrix, and finally the definition of a quality mapping.

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 close to the concept c in ontology. Internal features are the descriptive attributes of the concepts and their name, as well as domain values associated with the attributes. In order to identify the set of the external features of a concept c, we define neighbourhood V(k,c) of a concept c:

$$V(k,c)=\{c_i \in C \mid d(c_i,c)\le k\} \quad with\ d(c_i,c)\ \text{the number of links between}\ c_i\ \text{and}\ c \tag{3}$$

The neighbourhood of a concept c contains 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 $S(P_i(c_1), P_j(c_2))$ between sets of features $P_i(c_1)$ and $P_j(c_2)$ (further noted $S(P_i(c_1), P_j(c_2)) = S_{ij}(c_1,c_2)$):

$$M(c_1,c_2)=\begin{pmatrix} S_{11}(c_1,c_2) & S_{12}(c_1,c_2) & ... & S_{1n}(c_1,c_2) \\ S_{21}(c_1,c_2) & S_{22}(c_1,c_2) & ... & S_{2n}(c_1,c_2) \\ ... & ... & ... & ... \\ S_{n1}(c_1,c_2) & S_{n2}(c_1,c_2) & ... & S_{nn}(c_1,c_2) \end{pmatrix} \tag{2}$$

Within the framework of a qualitative model, this matrix will be used to 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:

$$m=(c_1,c_2,s,r,Q) \tag{3}$$

Q is also a tuple which includes the various characteristics of quality $q_1 \ldots q_n$:

$$Q=(q_1,q_2,....q_n) \tag{4}$$

The semantic similarity model presented in the following section compute the values of the similarity s and the qualitative relation r; in section 5 and 6, we will establish how the set of quality mapping characteristics Q can be determined.

4.2 Semantic Similarity Model

The proposed semantic similarity model considers qualitative and quantitative levels to establish the semantic relations between a concept c1 from ontology Oi and concepts of the set Cj from a second ontology Oj.. The semantic similarity model gives two output, that is a qualitative relation between the concepts and a global value of semantic similarity.

A concept is characterized by a set of features categorized according to n different categories:

$$c_1=\{P_1(c_1)\cup P_2(c_1)\cup ...\cup P_n(c_1)\} \tag{5}$$

The semantic relation between c_1 and c_2 is a function of the intersection of the features of concepts, given for each category of feature, but also it must be balanced compared to the set of features of the referent concept c_1:

$$S(c_1,c_2)=S\left[\frac{P_i(c_1)\cap P_i(c_2)}{P_i(c_1)}\right] \tag{6}$$

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, 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. This distinction is made in order to consider the fact that the non-mixed terms cannot be regarded as being equivalent under 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 of the second. In this case, the quantity of information shared by the concepts compared is less than if this common feature of the same category for both concepts. Moreover, each type of term considered in the model of similarity must itself be balanced by a weight ω_{ij} which gives the importance of the categories of features i and j.
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:

$$S(c_1,c_2)=\omega_{\text{int}}S_{\text{int}}(c_1,c_2)+\omega_{rel}S_{rel}(c_1,c_2)+\omega_{ext}S_{ext}(c_1,c_2)+\sum poids\_mixtes\cdot termes\_mixtes \tag{7}$$

When expressed in a general manner, the global expression of the similarity between two concepts is defined by considering non-mixed terms, i.e. terms which indicate the comparison between the same categories of features (for example the comparison between the internal features of c_1 and c_2) and then considering mixed terms, i.e. terms which consider the comparison between features of distinct categories (for example

comparison between external features of c_1 and the relational features of c_2):

$$S(c_1,c_2) = \sum_i \omega_i S_i + \sum_i \sum_{j \neq i} \omega_{ij} S_{ij} \quad (8)$$

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). The model of similarity provides an asymmetrical result, i.e. $S(c_1, c_2) = S(c_2, c_1)$ is not a condition which must be necessarily observed. The function of similarity must be asymmetrical since in general the qualitative relations which link concepts (such as inclusion, for example) are asymmetrical. We incorporate in the model of similarity the concept of Bayes conditional probability of which enables us to consider that**:**

$$S_{ij}(c_1,c_2) = \frac{P_i(c_1) \cap P_j(c_2)}{P_i(c_1)} = P\left[\frac{P_i(c_1) \cap P_j(c_2)}{P_i(c_1)}\right] = P(P_i(c_1) \mid P_j(c_2)) \quad (9)$$

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

$$P(P_i(c_1) \mid P_j(c_2)) = \frac{\sum_{s=1}^{k}\sum_{t=1}^{m} P(p_{si}(c_1))P(p_{tj}(c_2))}{\sum_{s=1}^{k} P(p_{si}(c_1))} \quad (10)$$

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*:

$$P(p(c)) = \frac{card(I_p(c))}{card(I(c))} \quad (11)$$

4.3 Computing Weight for the Mapping Model

Weight ω_{ij} in equation (6) indicate the importance of each term in the global similarity between the concepts c_1 and c_2. The semantic similarity model distinguishes various types of features P_i in the representation of concepts. However, in a first concept c_1, a feature p_i can belong to a category of feature P_i, but for a second concept c_2, it can belong to a distinct category P_j, so the mixed terms of equation 7 are often non null. Weights computation defines a weight for each combination of feature categories and by distinguishing the non-mixed terms from the mixed terms (figure 4).

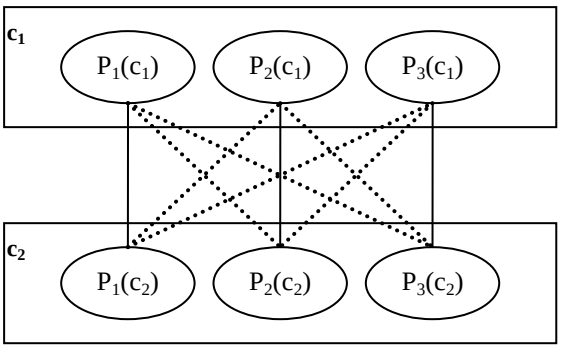

Figure 4: Mixed weights and non-mixed weights

In this section, we define the method of computation of weights which is based on the principle of the importance of information, that we define beforehand for the non-mixed terms and then for the mixed terms before giving the final expression of the weights.

The method of computation of **the 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 $\psi(P_i)$. The latter is evaluated by considering the importance of the information carried by each feature p_i *of* category P_i belonging to the concepts c_1 and c_2. Importance of the information carried by a feature $p_i(c)$ belonging to a concept c, noted $\psi(p_i(c))$, depends on the frequency with which this feature characterizes the concepts parents of the concept *c* as a feature of category P_i, evaluated compared to the total frequency of the feature p_i among the parents of the concept *c*. In other words, importance of the information of a feature $p_i(c)$ is high if this feature is also regarded most of the time as a feature of category P_i among the concept parents of *c*.
We define the parents of a concept c by the set of concepts of ontology which form one of the path *t* of the set of paths $T = \{t_i, i=1,2,3...\}$, defined by the following condition:

$$T = \{t \mid d(t) = root, sub(e(t)) = \varnothing, C \in t\} \quad (12)$$

That is *t* is a path for which the starting point *d(t)* is located at the root of ontology and the point of arrival *e(t)* is a concept which does not have an other subordinated concept, and which passes by the concept *c*. Moreover, the importance of information also depends, but in a less significant way and 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. The logarithmic function makes reduce the influence of the frequency of the features in whole ontology compared to their frequency among the parents concepts since we consider that the parents of a concept are more significant to define it. Importance of information $\psi(p_i(c))$ for mixed terms is given by:

$$\psi(p_i(C))_{non\_mixte} = \frac{freq(p_i(C) \in P_i)}{freq(p_i(C))} \times \log\left[\frac{N}{N(p_i)}\right] \quad (13)$$

where

$$freq(p_i(C) \in P_i) \quad (14)$$

give the frequency of p_i in T as a feature of category P_i and

$$freq(p_i(C)) \quad (15)$$

give the total frequency of the feature p_i in T. Total importance of the information carried by a category of features P_i for the comparison of two concepts c_1 and c_2 is given by the weighted sum of the importance of information for each feature:

$$\psi(P_i)_{non\_mixte} = \frac{1}{card(P_i(c_1) \cup P_i(c_2))} \sum_i \psi(p_i(C)) \quad (16)$$
$$p_i \in P_i(c_1) \cup P_i(c_2)$$

The method for computing **mixed weights** considers that the weight given to a mixed term (i.e. which compares the features of different categories P_i et P_j of concepts c_1 and c_2) of equation (6) depends on the frequency with which the feature p_i in T as a feature of category P_i and of the frequency of the feature p_i in T as a feature of category P_j. Importance of information $\psi(p_i(c))$ for mixed terms is given by

$$\psi(p_i(C))_{mixte} = \frac{freq(p_i(C) \in P_i)}{freq(p_i(C) \in P_j)} \times \log\left[\frac{N}{N(p_i)}\right] \quad (17)$$

Total importance of the information carried by the features of categories P_i compared to the features of category P_j for the comparison of the mixed terms of two concepts c_1 et c_2 is given by the balanced sum of the importance of information for each feature:

$$\psi(P_i, P_j)_{mixte} = \frac{1}{card(P_i(c_1) \cup P_j(c_2))} \sum_i \psi(p_i(C)) \quad (18)$$
$$p_i \in P_i(c_1) \cup P_j(c_2)$$

Les poids ω_{ii} pour les termes non mixtes et ω_{ij} pour les termes mixtes sont proportionnels à l'importance de l'information correspondante à chaque terme laquelle est pondérée par l'importance de l'information portée par tous les termes :

$$\omega_{ii} = \frac{\psi(P_i)}{\sum_{i=1}^{n} \psi(P_i) + \sum_{i=1}^{n} \sum_{i \neq j} \psi(P_i, P_j)} \quad (19)$$

$$\omega_{ij} = \frac{\psi(P_i, P_j)}{\sum_{i=1}^{n} \psi(P_i) + \sum_{i=1}^{n} \sum_{i \neq j} \psi(P_i, P_j)} \quad (20)$$

It should be noted that the weights are different for each couple of concept thus giving the distinction of our approach, since in other approaches does not consider these various points of view when computing a total weight, which does not distinguish where the concepts are in the ontology nor how their features are categorized in another ontology.

The model of similarity expressed by the equation (8) give a single value to the similarity between two concepts c_1 and c_2, which is the value s of the mapping of equation 3. The constitution of this expression in form of a sum of terms related to the comparison of the categories of features (internal, relational and external) makes it possible to consider these terms in an individual way to give the nature of the qualitative relation between the concepts c_1 and c_2. The various terms of equation 8 make it possible to constitute the matrix of relations definite in section 4.1 (equation 2), since each term corresponds in fact to the comparison between two categories of features of the two concepts. By considering the definition of the concepts in terms of the union of the sets of the three categories of features, the quantitative matrix of the relations comprises 9 terms, where the terms of the diagonal correspond to the non-mixed terms and the terms out of diagonal correspond to the mixed terms:

$$M(c_1, c_2) = \begin{pmatrix} S_{\text{int\_int}}(c_1,c_2) & S_{\text{int\_}rel}(c_1,c_2) & S_{\text{int\_}ext}(c_1,c_2) \\ S_{rel\_\text{int}}(c_1,c_2) & S_{rel\_rel}(c_1,c_2) & S_{rel\_ext}(c_1,c_2) \\ S_{ext\_\text{int}}(c_1,c_2) & S_{ext\_rel}(c_1,c_2) & S_{ext\_ext}(c_1,c_2) \end{pmatrix} \quad (21)$$

This matrix of relations makes allows to identify the qualitative relation R which bonds the concepts c_1 and c_2 by considering the particular cases of global equation (8). 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. Table 2 presents some cases of qualitative relations associated in possible states of the matrix of relations.

Table 2: Qualitative relations between the concepts, compared to the states of the matrix of relations.

States of the qualitative relation	State of the matrix of relation
c_1 equivalent to c_2 ($c_1 \equiv c_2$)	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
c_1 overlap c_2	$\begin{pmatrix}]0,1[&]0,1] &]0,1] \\ ]0,1] &]0,1] &]0,1] \\ ]0,1] &]0,1] &]0,1[\end{pmatrix}$
c_1 is included in c_2	$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ ]0,1] &]0,1] &]0,1[\end{pmatrix}$

This section concludes the presentation of the model of mapping of quality which makes it possible to establish mappings expressing at the same time qualitative and

quantitative information. In the next section, we present the conceptual framework for the quality of the mappings, which will make it possible to determine Q.

5. A FRAMEWORK FOR QUALITY OF MAPPING

The development of a framework for evaluation of the quality of the mapping initially requires to provide an adequate definition of the quality of the mapping. We identify the relevant characteristics to describe the quality of the mapping. These characteristics will be then defined within the framework of the mapping between two ontologies. In section 5.3, we develop the measures to evaluate the characteristics of the quality of the mappings.

5.1 What is Mapping Quality

A discussion on the quality of the mapping must necessarily begin with a definition from quality from the 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 the mappings
The quality of the 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 mappings will be able to help to improve 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 that it impossible to identify the nature of a perfect mapping, for same the reasons as the development, for example, of an ontology cannot be described as "perfect". For example, from the definition we gave for quality of a mapping, it can be observed that the latter depends on the quality of the definition of the entities and thus of the quality of the development of ontology.

Before identifying and defining the characteristics of the quality of the mapping, we situate the conceptual framework of our approach in the following section.

5.2 Conceptual Setting for Mapping Quality

Within this conceptual framework, we consider two ontologies A and B of 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:

$$\left\{\begin{array}{l} intersects(\cap), includein(\subseteq), includes(\supseteq), \\ equals(\equiv), disjointness(\perp) \end{array}\right\} \quad (20)$$

For example, the mapping $m=(a_1,b_1,s,r=\subseteq)$ indicates that. $a_1 \subseteq b_1$. The 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.
We consider that the set of mappings between ontologies A and B forms a matrix of mapping which indicates the relations between the concepts, for example for mapping of cardinality n: m with qualitative relations, one would have a matrix of the form:

$$M = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ ... \\ a_n \end{array} \begin{array}{c} \begin{array}{ccccc} b_1 & b_2 & b_3 & ... & b_5 \end{array} \\ \begin{pmatrix} \equiv & \subseteq & \subseteq & ... & \perp \\ \cap & \cap & \supseteq & ... & \perp \\ \perp & \perp & \perp & ... & \cap \\ ... & ... & ... & ... & ... \\ \cap & \subseteq & \subseteq & ... & \perp \end{pmatrix} \end{array} \quad (21)$$

For a model of mapping which provides mappings of cardinality 1:1 by establishing a threshold on the results of a semantic similarity measurement s, there would be rather a matrix of the form:

$$M = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ ... \\ a_n \end{array} \begin{array}{c} \begin{array}{ccccc} b_1 & b_2 & b_3 & ... & b_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ 0 & 0 & 0 & ... & 1 \\ ... & ... & ... & ... & ... \\ 0 & 1 & 0 & ... & 0 \end{pmatrix} \end{array} \quad (22)$$

where 1 indicates that the concepts are regarded as similar and 0 indicate that the concepts are regarded as dissimilar.

5.3 Defining Quality Characteristics for Mapping

The majority of the approaches for the evaluation of quality defines 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 mappings, we analyzed the possible impacts of the process of mapping on the various characteristics of the quality of information recognized in the field of the quality of information. Thus, our approach tries to be anchored as much as possible within framework of existing work.
For our conceptual framework, we conceive that the quality of the mapping must integrate the characteristics indicated on figure 5.

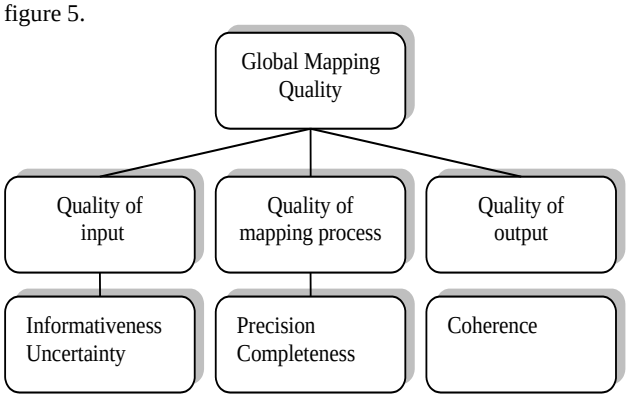

Figure 5 : Structure of characteristics of the quality of the mappings

The first level of the diagram of figure 5 indicates the categories of characteristics of quality. The quality of a mapping can be described by the quality of the information exploited by the mapping (quality of input), which is generally related to the quality of the definition of the entities compared by the model of mapping. Then the quality of the mapping is also determined by the quality of the process, which is related on the precision and the completeness of the mapping process. Finally, the quality of a mapping is measured with the quality of the result, which we will describe compared to the coherence of a mapping with the existing relations in ontologies.

For each one of these characteristics, we propose definitions which account for the possible impact of the mappings on these characteristics of quality. The first characteristics of the quality of the mapping relate to the quality of information that the mapping exploits to establish the relation between the concepts. These characteristics are thus related to the quality of the definition of the concepts.

Characteristic 1: Informativeness of mapping
A mapping is informative when the definition of the concepts is complete. 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 2: Uncertainty of a mapping
A mapping is uncertain when it is based on a uncertain definition of the compared concepts.

The following characteristics relate to the quality of the process of mapping. 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 3: 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.

Example2.
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 second dimension of the quality of the mappings 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 4 : Completeness of mapping
A mapping preserves completeness of concepts when it takes account of all the aspects of the definition of the concepts.

Example 3.
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.

The next characteristic of quality relates to the third category, that is the one that assess the quality of the result of the mapping process. The quality of a mapping, considered individually, must also be considered from the point of view of its coherence with the mappings established between other concepts. This coherence must be evaluated considering that concepts of an ontology are structured by relations (hierarchical relations, taxonomy, part-of relations, etc). Just like a hierarchy describes a set of logic relations between the concepts, the set of mappings describes logical relations between concepts from different hierarchies. Consequently, we will describe the coherence of a mapping compared to its coherence with the relations which already exist between the concepts of the same ontology. Previously, we define the significance of neighbour mappings and hierarchical conflict between mapping.

Definition 2: Mapping neighbours
Consider a mapping m which relates two concepts a_1 and b_1 with relation r_1:

$$m = (a_1, b_1, s, r_1) \tag{23}$$

and consider a mapping m' which relates concepts a_2 and b_2 with relation r_2 :

$$m' = (a_2, b_2, s, r_2) \tag{24}$$

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 neighbours 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 neighbours 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.

For the description of the hierarchical conflicts between the mappings, we consider 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 relations :

$$a_0 \supseteq a_1 \supseteq a_2 \text{ and } b_0 \supseteq b_1 \supseteq b_2 \tag{25}$$

as illustrated on figure 5. We consider five cases of conflicts that is each possible semantic relation between the concepts a_1 and b_1 (equation 20). In each case, we identify the semantic relations between neighbours concepts which are in logical conflict with the relation between a_1 and b_1 and with the conditions of equation (25). These contradictions define the hierarchical conflicts between the mappings. The hierarchical conflicts which 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.

1) 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:

$$m=(a_1,b_1,r=equals)\wedge\begin{cases} m'=(a_0,b_0,r=\{\perp\}) \\ m'=(a_0,b_1,r=\{\equiv,\subseteq,\perp,\cap\}) \\ m'=(a_0,b_2,r=\{\equiv,\subseteq,\perp,\cap\}) \\ m'=(a_1,b_0,r=\{\equiv,\supseteq,\perp,\cap\}) \\ m'=(a_1,b_2,r=\{\equiv,\subseteq,\perp,\cap\}) \\ m'=(a_2,b_0,r=\{\equiv,\supseteq,\perp,\cap\}) \\ m'=(a_2,b_1,r=\{\equiv,\subseteq,\perp,\cap\}) \end{cases} \quad (26)$$

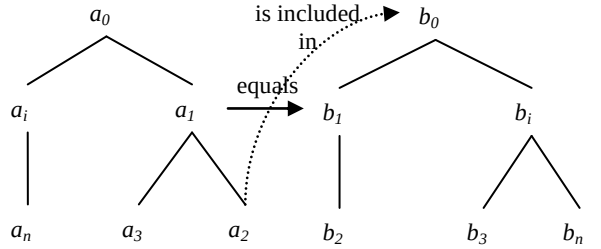

Figure 5: Example of hierarchical conflict between mappings m=(a_1,b_1,r = *equals*) and m=(a_2,b_0,$r = \supseteq$).

2) 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:

$$m=(a_1,b_1,r=\subseteq)\wedge\begin{cases} m'=(a_0,b_0,r=\{\perp\}) \\ m'=(a_0,b_1,r=\{\perp\}) \\ m'=(a_0,b_2,r=\{\perp\}) \\ m'=(a_1,b_0,r=\{\equiv,\supseteq,\perp,\cap\}) \\ m'=(a_2,b_0,r=\{\equiv,\supseteq,\perp,\cap\}) \\ m'=(a_2,b_1,r=\{\equiv,\supseteq,\perp,\cap\}) \end{cases} \quad (27)$$

3) 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:

$$m=(a_1,b_1,r=\supseteq)\wedge\begin{cases} m'=(a_0,b_0,r=\{\perp\}) \\ m'=(a_0,b_1,r=\{\equiv,\subseteq,\perp,\cap\}) \\ m'=(a_0,b_2,r=\{\equiv,\subseteq,\perp,\cap\}) \\ m'=(a_1,b_0,r=\{\perp\}) \\ m'=(a_1,b_2,r=\{\equiv,\subseteq,\perp,\cap\}) \end{cases} \quad (28)$$

4) 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:

$$m=(a_1,b_1,r=\cap)\wedge\begin{cases} m'=(a_0,b_0,r=\{\perp\}) \\ m'=(a_0,b_1,r=\{\equiv,\subseteq,\perp\}) \\ m'=(a_0,b2,r=\{\equiv,\subseteq\}) \\ m'=(a_1,b_0,r=\{\equiv,\supseteq,\perp\}) \\ m'=(a_1,b_2,r=\{\equiv,\subseteq\}) \\ m'=(a_2,b_0,r=\{\equiv,\supseteq\}) \\ m'=(a_2,b_1,r=\{\equiv,\supseteq\}) \end{cases} \quad (29)$$

5) 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:

$$m=(a_1,b_1,r=\perp)\wedge\begin{cases} m'=(a_0,b_1,r=\{\equiv,\subseteq\}) \\ m'=(a_0,b_2,r=\{\equiv,\subseteq\}) \\ m'=(a_1,b_0,r=\{\equiv,\supseteq\}) \\ m'=(a_1,b_2,r=\{\equiv,\supseteq,\subseteq,\cap\}) \\ m'=(a_2,b_0,r=\{\equiv,\supseteq\}) \\ m'=(a_2,b_1,r=\{\equiv,\supseteq,\subseteq,\cap\}) \\ m'=(a_2,b_2,r=\{\equiv,\supseteq,\subseteq,\cap\}) \end{cases} \quad (30)$$

Characteristic 3: 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 conditions presented in definition 3.

Starting from these definitions which we established to form the conceptual framework of the quality of the mappings, we develop in the following section the measures for the different quality mapping characteristics.

6. DEVELOPMENT OF QUALITY MESURES FOR MAPPING

In this section, we propose measures to evaluate quantitatively the characteristics of quality which were defined in the preceding section.

Measuring informativeness of mapping

An incomplete definition of the concepts related by a mapping decreases the degree of information of this mapping. The measure of informativeness of a mapping indicates if a mapping was computed while being based on missing values or information in the definition of the concepts. Consider d the expected degree of information for a concept. The expected degree of information is equivalent to the number of attributes which a concept should possess, considering the general definition of the concepts in the ontology. Consider d' the actual number of attributes for the concept. Then the informativeness of a mapping is given by

$$Informativeness=\frac{1}{2}\left(\frac{d_1'}{d_1}+\frac{d_2'}{d_2}\right) \quad (31)$$

where the subscripts indicate the degree of information for the two compared concepts in the mapping. If the entities which are compared by the model of mapping are instances, d corresponds

to the number of attributes for which a value must be given, and d' to the actual number of attributes for which a value is effectively given.

The informativeness measure yields values between 0 and 1, 0 meaning that informativeness is null and 1 meaning that informativeness is maximum, i.e. mapping is based one has full information.

Measuring uncertainty of mappings

As described previously, the uncertainty of a mapping is determined by uncertainty in the definition of the concepts. The uncertainty of the definition of the concepts is related to uncertainty on the domain values of the attributes of concepts. If the comparison of the concepts, during the process of mapping, is based on the comparison of their attributes and their domain values, uncertainty on these domain values will be reflected on the quality of mapping.

We suppose that a concept a_i is defined by an attribute p whose domain value is given by an interval [q0, q1]. Moreover, let us suppose that the limits of this interval are uncertain, and that probability that q0=x or q1=x is given by a density probability function P(x), which verify that:

$$\int_{-\infty}^{\infty} P(x)dx = 1 \tag{32}$$

In this case, we will use the principle of the entropy to determine the uncertainty of the domain values. In the information theory, entropy is a concept which indicates the average uncertainty of a source of information (Shannon, 1948). Entropy of a distribution of density of probability P(x) is given by:

$$H(x) = -\int_{-\infty}^{\infty} P(x)\ln P(x)dx \tag{33}$$

Noting that entropy is additive, i.e. entropy of two independent variables x1 and x2 is given by:

$$H(x1,x2) = H(x1) + H(x2) \tag{34}$$

it is possible to combine entropy of several independent variables, i.e. entropy of the various attributes of a concept p_1, p_2, …p_n. Consider P(x1), P(x2),…P(xn) density probabilities functions for attributes p_1, p_2, …p_n. Then the total entropy for a concept is given by:

$$H(x1,..xn) = \sum_{i=1}^{n} H(xi) \tag{35}$$

For example, the density probability function for the domain of value of an attribute can be a normal distribution, which is often the case for spatial attributes (for example, density of forest cover). Finally, average relative uncertainty on the definition of a concept is given by the radius of the interval of uncertainty of entropy, normalized by size of domain values of attributes l_i:

$$\Delta c = \frac{1}{2\sum_i l_i}\exp(H(x1,..xn)) \tag{36}$$

Measuring precision of mappings

The precision of mapping must indicate the degree of agreement between the level of detail taken into account by the model of mapping and the level of detail in the actual representation of concepts to be compared. Let L be the level of detail of the concept. For example, if the concept (level 1) is successively defined by attributes (level 2), attributes are defined by domain values (level 3) and domain values are described by metadata (level 4), we would say that L=4. Then let L' be the level of detail considered by the model of mapping (for example, the model of mapping can compare up to the domain values of the attributes). Then we can measure precision of mapping by

$$\mathrm{Pr}ecision = \frac{1}{2}\left(\frac{L_1'}{L_1} + \frac{L_2'}{L_2}\right) \tag{37}$$

We see that precision is maximal when L'=L, which means that the model of mapping uses the finer level of definition possible for comparing concepts. As the level of detail considered by the model becomes lower, precision goes down toward 0.

Measuring completeness of mappings

Completeness of mapping is similar to precision as it should measure the degree of agreement between aspects of the definition of concepts considered in the model of mapping and the actual aspects of the definition of concept in ontologies. Let C be the number of aspects used to define concepts of an ontology, and let C' be the number of these aspects actually been used by the model of mapping. Then if we consider these quantities for both concept been compared c_1 and c_2, completeness of a mapping may be defined as:

$$Completeness = \frac{1}{2}\left(\frac{C_1'}{C_1} + \frac{C_2'}{C_2}\right) \tag{38}$$

As informativeness and precision, mapping preserves entirely completeness when its value is 1 and preserves no completeness when its value is 0.

Measuring coherence of mappings

The definition proposed for the coherence of mappings suggests that for each neighbour concept of a concept a_i related to concept b_j by the mapping $m = (a_i, b_j, r)$, coherent and incoherent mappings can be identified by checking each condition of coherence given in definition 3. Consequently, a ratio coefficient can be used to compare the number of coherent mapping to the number of incoherent mapping, and this for each neighbour concept of a_i.

Consider $m = (a_1, b_1, r_{11})$ a mapping for which we want to compute coherence with neighbour mappings. Let us consider

$$\begin{aligned} V(a_1) &= \left\{v_i^a \mid dist(a_1, v_i^a) = 1\right\} \\ V(b_1) &= \left\{v_j^b \mid dist(b_1, v_j^b) = 1\right\} \end{aligned} \tag{39}$$

the set of neighbour concepts of concept a_i, and the set of neighbour concepts of b_j, respectively. These concepts are related by a set of mapping

$$M = \left\{ m_{ij} \mid m_{ij} = (v_i^a, v_j^b, r_{ij}) \right\} \tag{40}$$

where the state of each mapping m_{ij} compared to the mapping m is said to be coherent or incoherent, according to conditions' stated within definition 3 of section 4.3. Consider nc_i the set of coherent mapping that relates the concept v_i^a to a concept $v_j^b \in V(b_1)$ and consider nn_i the set of incoherent mapping that relate concept v_i^a to a concept $v_j^b \in V(b_1)$. The global coherence of the mapping m is measured by:

$$cohérence = \frac{1}{n} \sum_{i=1}^{n} \frac{nc_i}{nc_i + nn_i} \tag{41}$$

where n is the number of neighbour concepts of a_i : n = card($V(a_1)$). It is not necessary to consider mappings from the point of view of neighbour concepts of b_j since these mappings have each one a reciprocal mapping established from the point of view of neighbours of a_i. In other words, if one determines the number of mapping incoherent with mapping $m = (a_i, b_j, r)$, one detects by the way mappings that are incoherent with the reciprocal mapping $m = (b_j, a_i, r)$.

Now that we have defined and proposed measures for the different characteristics of quality of mapping, 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:

$$Q = (uncertaint y, \inf ormativeness, precision, completeness, coherence) = (\Delta, I, P, C, \theta)$$

7. EXAMPLE OF APPLICATION

In order to show the usefulness of our approach, we applied the model of quality mapping with two models of mapping, namely the model of mapping developed in this article and the Matching Distance model that was designed to relate geospatial classes entities from different ontologies (Rodriguez and Egenhofer, 2003). We use part of the ontologies of the NTDB (National Topographic DataBase of Canada) and of the BDTQ (Base de données topographique du Québec). For concept having attributes being associated to domain values, we suppose uncertainty of the values was determined by a Gaussian distribution, meaning:

$$P(x_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{x^2}{2\sigma^2} \right\}; \sigma = 1.$$

For example, the concept Water Disturbance from NTDB is define by the following set of features:

P_{int} = {site, zone, movement, disturbed water, surface, slope};

P_{rel} = {divides river, divides obstacle to navigation, connects river, connects lakes, connects obstacle to navigation};

P_{ext} = {P(watercourse)), P(obstacle to navigation), P(lake)}, where P(watercourse) expresses the set of the properties of the concept watercourse.

Instances = {Unknown or generic water disturbance, small waterfalls, large waterfalls}.

Tables 3 and 4 show the results obtained for the query asking what are the relation between concept watercourse from ontology of NTDB and concepts of a small part of the ontology of BDTQ.

Table 3: Result obtained with the proposed model

Concepts of NTDB compared to concept : *watercourse*	Results with the proposed model
Watercourse	$m = (c_1, c_2, s = 0{,}6768, r = \subseteq)$ $Q = (\Delta = 6{,}89\%; I = 0{,}78; P = 1; C = 1; \theta = 0{,}90)$
Water disturbance	$m = (c_1, c_2, s = 0{,}6278, r = \cap)$ $Q = (\Delta = 6{,}89\%; I = 0{,}55; P = 1; C = 1; \theta = 0{,}85)$
Water Channel	$m = (c_1, c_2, s = 0{,}5791, r = \cap)$ $Q = (\Delta = 7{,}51\%; I = 0{,}90; P = 1; C = 1; \theta = 0{,}90)$
Water source	$m = (c_1, c_2, s = 0{,}2548, r = \cap)$ $Q = (\Delta = 6{,}89\%; I = 0{,}78; P = 1; C = 1; \theta = 0{,}90)$
Liquid Deposit	$m = (c_1, c_2, s = 0{,}1441, r = \cap)$ $Q = (I = 1; P = 1; C = 1; \theta = 0{,}90)$
Dam	$m = (c_1, c_2, s = 0{,}1156, r = \cap)$ $Q = (\Delta = 3{,}25\%; I = 0{,}78; P = 1; C = 1; \theta = 0{,}90)$
Quay	$m = (c_1, c_2, s = 0{,}0604, r = \cap)$ $Q = (I = 1; P = 1; C = 1; \theta = 0{,}74)$
Pond separator	$m = (c_1, c_2, s = 0{,}0000, r = \perp)$ $Q = (I = 0{,}80; P = 1; C = 1; \theta = 0{,}88)$

Table 4: Result obtained with the MD model

Concepts of NTDB compared to concept : *watercourse*	Results with Matching Distance Model
Watercourse	$m = (c_1, c_2, s = 0{,}6148, r = equals)$ $Q = (\Delta = 6{,}89\%; I = 0{,}78; P = 0{,}66; C = 0{,}5; \theta = 0{,}40)$
Water disturbance	$m = (c_1, c_2, s = 0{,}5424, r = equals)$

	$Q = (\Delta = 6{,}89\%;\ I = 0{,}55;$ $P = 0{,}66;\ C = 0{,}5;\ \theta = 0{,}42)$
Water Channel	$m = (c_1, c_2, s = 0{,}5424, r = equals)$ $Q = (\Delta = 7{,}51\%;\ I = 0{,}90;$ $P = 0{,}66;\ C = 0{,}5;\ \theta = 0{,}68)$
Water source	$m = (c_1, c_2, s = 0{,}1066, r = \perp)$ $Q = (\Delta = 6{,}89\%;\ I = 0{,}78;$ $P = 0{,}66;\ C = 0{,}5;\ \theta = 0{,}40)$
Liquid Deposit	$m = (c_1, c_2, s = 0{,}0463, r = \perp)$ $Q = (I = 1;$ $P = 1;\ C = 0{,}5;\ \theta = 0{,}42)$
Dam	$m = (c_1, c_2, s = 0{,}0355, r = \perp)$ $Q = (\Delta = 3{,}25\%;\ I = 0{,}78;$ $P = 0{,}66;\ C = 0{,}5;\ \theta = 0{,}55)$
Quay	$m = (c_1, c_2, s = 0{,}0202, r = \perp)$ $Q = (I = 1;$ $P = 1;\ C = 0{,}5;\ \theta = 0{,}74)$
Pond separator	$m = (c_1, c_2, s = 0{,}0000, r = \perp)$ $Q = (I = 0{,}80;$ $P = 1;\ C = 0{,}5;\ \theta = 0{,}88)$

As the output of the MD model is only a global semantic similarity value but no qualitative relation between concepts, we had to consider that over a threshold (chosen to be 0,5) concepts were matched (i.e. considered to be equal) and under this threshold they were considered to be disjoint. This had for consequence that coherence is generally lower than the first model,and it shows that a model of mapping that can not indicate the nature of the relationship between concept is insufficient to complete the matching task since applying a threshold leads to less coherent mappings.

We can see in this example that informativeness (I) and uncertainty (Δ) are independent of the model of mapping that is used. This is because these characteristics of quality are related to the quality of input of the model of mapping, therefore they indicate quality of mapping but quality of the model of mapping. On the other hand, precision (P) depends on the model of mapping and we can see that in general precision of MD model was lower, this is because it was not designed to evaluate similarity between domain values. When precision of MD model was 1, it was because attributes has no domain values. No domain values also cause uncertainty to be absent. Completeness (C) also depends on the model that was used and it was systematically lower using MD model because, compare to the model of mapping developed in this article, it does not consider instances to compute semantic similarity. We consider that for each user some characteristic may be of relative interest depending on their context.

This example will also show that quality of mapping, as we expected, depends on the agreement between model of mapping and the mapping task to be executed, that is, what are the "things" to be mapped, because precision for example become smaller for a given model as definition of concept becomes more precise as expected by the model of mapping. It will also show that quality of mapping influences quality of information provided to the end-user since the characteristics of quality can help to select mappings that we can thrust comparing to other, thus giving better final information.

8. CONCLUSION AND FUTURE WORKS

In this article, we have argued that quality of mapping is an important feature because it may have impact on the quality of query answering over multiples data sources. Since quality of mapping has been addressed from the point of view of the global performance but not for individual mapping, existing evaluation of quality of mapping are of no help to evaluate actual quality of the individual mapping that will be used for query processing. In this article, we presented a conceptual framework for quality of mapping, giving original definitions for quality characteristics of mappings, measures to evaluate these characteristic and a model of quality mapping. We believe that this approach can lead to better results of mapping and can also indicate to user the quality of information resulting of semantic integration of multiple sources.

In the future works, we attempt to define a framework for quality mapping that can be spatial, semantic or temporal mappings. We also project to define in more details the concept of semantic mapping quality, and its properties and behaviors, in the semantic interoperability process.

9. 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.

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.

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.

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.

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.