RAPPORT DE RECHERCHE

L R I

SAME WORDS ? SAME WORLDS ? COMPARING ONTOLOGIES UNDERLYING GEOGRAPHIC DATA

MUSTIERE S / REYNAUD C / SAFAR B / ABADIE N

Unité Mixte de Recherche 8623
CNRS-Université Paris Sud – LRI

06/2009

Rapport de Recherche N° 1521

CNRS – Université de Paris Sud
Centre d'Orsay
LABORATOIRE DE RECHERCHE EN INFORMATIQUE
Bâtiment 490
91405 ORSAY Cedex (France)

Same words? Same worlds? Comparing ontologies underlying geographic data

Sébastien Mustière[1], Chantal Reynaud[2], Brigitte Safar[2], Nathalie Abadie[1]

[1] IGN/COGIT, 2 avenue Pasteur 94160 Saint-Mandé, France
[2] LRI, Université Paris-Sud, Bât. G, INRIA Saclay-Ile-de-France
2-4 rue Jacques Monod, F-91893 Orsay, France

sebastien.mustiere@ign.fr, chantal.reynaud@lri.fr, brigitte.safar@lri.fr,
nathalie-f.abadie@ign.fr.

Abstract. Assessing how much two geographic databases reflect the same point of view is a key issue for data integration. We argue that this task requires developing ontologies revealing the point of view of each piece of information, and neglecting the technical choices behind the organization of information in data schemas. These ontologies need then to be aligned and globally compared. In this paper, we describe techniques allowing doing that in the fields of ontology alignment and natural language processing. We illustrate those points through results from experiments made on actual data with a semi-automated analysis of their specifications.

Keywords: geographic database, data integration, semantics, ontology alignment, data specifications.

1 Assessing the point of view of geographic information

Geography is a scientific field that requires combining numerous information about human as well as environmental phenomena and processes. It is essentially a major field of application of multi-criteria analysis [17]. Hopefully, more and more geographic information are nowadays available, each one reflecting a special point of view about the same geographic "real world". This is a great opportunity for geographic analysis, and this amount of information could lead to significant knowledge. Unfortunately, those pieces of information are usually independently produced and managed. Many efforts are still necessary in the research area to help the integration of heterogeneous data, dealing with their diversities and complementarities but also their imprecision, uncertainty, incompleteness, redundancies and inconsistencies [5][8][22].

One of the main challenges for such integration is to be able to assess and compare the points of view behind pieces of geographic information to be integrated. Indeed, each geographic database reflects the particular conceptualization of the world of its producer [2]. As a consequence, a single phenomenon of the real world will have different interpretations and descriptions, reflecting the "semantic heterogeneity" of

geographic data [20]. Understanding the semantic heterogeneity of different geographic datasets is the key to assess their complementarities and redundancies. In other words, it is the key to answer the questions: is there any meaning of integrating those two pieces of information? If yes, how should it be done?

There is a common agreement of this issue: semantics of given information would gain a lot to be explicitly represented [20], and ontologies are tools to do that. In this paper we address two related questions: how to build ontologies underlying some given geographic datasets and, once this is done, how to compare those ontologies?

2 Discovering the ontology underlying geographic databases

2.1 The gap between schema and ontology

The first element that may explicit the semantics of a dataset is the data schema. Every database conceptual schema is made with a certain goal, and then is based on an underlying ontology [20]. A conceptual schema whose elements would correspond exactly to the concepts of the ontology would be called an ontological schema. However, due to technical and historical reasons, geographical schemas are far from being ontological schemas.

One explanation of that originates from the history of GIS softwares. Most, if not all, geographical schemas separate classes of objects holding surfacic, punctual and linear geometries. This originates from the fact that the main challenges for first GIS softwares was the management of geometric properties. Thus, geometric types have been defined and were the basis for defining the schema: an old and still current view is that geographical data are first of all geometries associated to some properties. A more ontological approach should be to rely only on geographic (not geometric) concepts to define the schema, and then to consider geometries as one property of features among others. This consideration explains some gaps between schemas and an ontology. For example, we find in some data schema the independent classes “surface building”, “point building” and “linear building”, while the underlying ontology would better contain only the concept of “building”, may be with some properties like “shape” or “height”.

Another explanation originates also from the nature of geographic data: most meaningful relations between data may be spatial relations like “to be near” / “to follow” / “to lead to”, which can be derived from the geometric properties of features, at least interactively when visualising the data. For this reason, geographic schema explicit very few relations between classes, if any, compared to most database schemas in other fields. In the opposite, ontologies intend to explicit those relations.

Another, and certainly deeper explanation, originates from the mapping history of geographic databases. Most geographic databases have been defined to produce maps, or at least by people and organisations with a strong mapping background. Classes of geographical schemas may then group objects based on cartographic habits rather than ontological considerations. In a caricatured manner, a class may contain all objects appearing in blue in a map. More reasonably, we encounter databases with a

class representing all together the concepts of “beach”, “summit”, valley”, “cave”, “peninsula” and “crest”. The main reason is that all these geographic concepts are considered in those databases as background information to be displayed only as a toponyms. The underlying ontology of this database would gain to contain all these disjoint concepts, associated with the property “geographical name”.

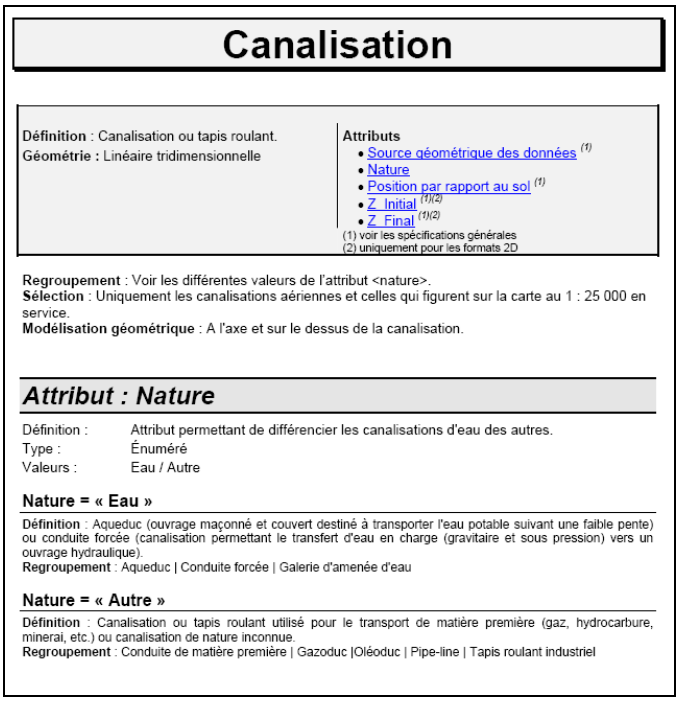
Finally, the design of geographic databases requires a complex conceptualization process where selection, aggregation and splitting operations are performed. The geographic space is far too complex to be represented entirely. Thus, two designers with the same goal may design two different database schema. If this can be said in any thematic domain, this is particularly true for geography because of the complexity of the geographic world. It is thus very difficult to separate differences between schemas originating from design choices from thus originating from actual differences in conceptualizations of the world.

However, we do not claim that geographical schemas should be redesigned with ontological principles because, first, we need to handle existing databases and, more importantly, these schemas are usually very adapted to most user needs among which mapping is important. Nevertheless, for all the reasons explained above, schemas are not rich enough and organised to be used as ontological references to assess the point of view of a database.

2.2 Textual specifications as sources of semantics

Most of the knowledge reflecting the database design and instantiation process is compiled in textual documents, known as database specifications. Specifications of topographic databases precisely describe the meaning of the database content through the description of the data capture process. They are paper documents, usually covering hundreds of pages, highly structured but containing a lot of informal information expressed in natural language. They are used as guidelines for data capture, as well as data description for data management and data usage.

If differences can appear between specifications from different data producers, all of them globally contain the same information (see example in Fig. 1). After some general information about the database, themes and feature classes are described one by one. After the name of the class, a definition explains the meaning of the represented concept in a few sentences. Then, selection criteria are usually detailed, i.e. the conditions that a real world object must observe to be part of the database. Attention is also paid to information related to the geometry of the feature class. Other attributes are finally listed and described, usually by textual definitions, the list of their possible values, and again a textual definition for each possible value.



Canalisation

Définition : Canalisation ou tapis roulant.
Géométrie : Linéaire tridimensionnelle

Attributs
- Source géométrique des données (1)
- Nature
- Position par rapport au sol (1)
- Z_Initial (1)(2)
- Z_Final (1)(2)

(1) voir les spécifications générales
(2) uniquement pour les formats 2D

Regroupement : Voir les différentes valeurs de l'attribut <nature>.
Sélection : Uniquement les canalisations aériennes et celles qui figurent sur la carte au 1 : 25 000 en service.
Modélisation géométrique : A l'axe et sur le dessus de la canalisation.

Attribut : Nature

Définition : Attribut permettant de différencier les canalisations d'eau des autres.
Type : Énuméré
Valeurs : Eau / Autre

Nature = « Eau »

Définition : Aqueduc (ouvrage maçonné et couvert destiné à transporter l'eau potable suivant une faible pente) ou conduite forcée (canalisation permettant le transfert d'eau en charge (gravitaire et sous pression) vers un ouvrage hydraulique).
Regroupement : Aqueduc | Conduite forcée | Galerie d'amenée d'eau

Nature = « Autre »

Définition : Canalisation ou tapis roulant utilisé pour le transport de matière première (gaz, hydrocarbure, minerai, etc.) ou canalisation de nature inconnue.
Regroupement : Conduite de matière première | Gazoduc |Oléoduc | Pipe-line | Tapis roulant industriel

Fig. 1. Excerpt from textual specifications of IGN-France.

All this information, be it formal or textual, is very rich. It is the best available source of knowledge describing the point of view followed when designing the database. We may thus reasonably think that these documents could be sources of knowledge to explicit the ontology underlying a geographical database, and even gaps between this ontology and the schema [10].

In order to experiment this idea, we analysed two different specifications documents from IGN, the French mapping agency. By means of natural language processing tools including morpho-syntactic parsing, we searched for geographic concepts expressed in the textual parts of the specification, like in definitions of classes. The overall process was to identify all nominal groups in sentences, then filter them with some external corpuses of non geographic documents, and organise the retained concepts in a hierarchy according to their location in the document [1]. Based on results of this first automated step, an interactive filtering and reorganisation has been definitely necessary to complete the work. For each database and its related specification, this process led us to a taxonomy of concepts encountered in the specification, a first step toward a more formal ontology[1] (see extract in Fig. 2). Insights from these experiments were manifold.

[1] We use the word 'ontology' in this paper in order to designate a formal conceptualization of the world, be it a simple taxonomy, without any reference to a particular modeling of it, like for example in OWL format with the formal concepts of properties, relations, definitions…

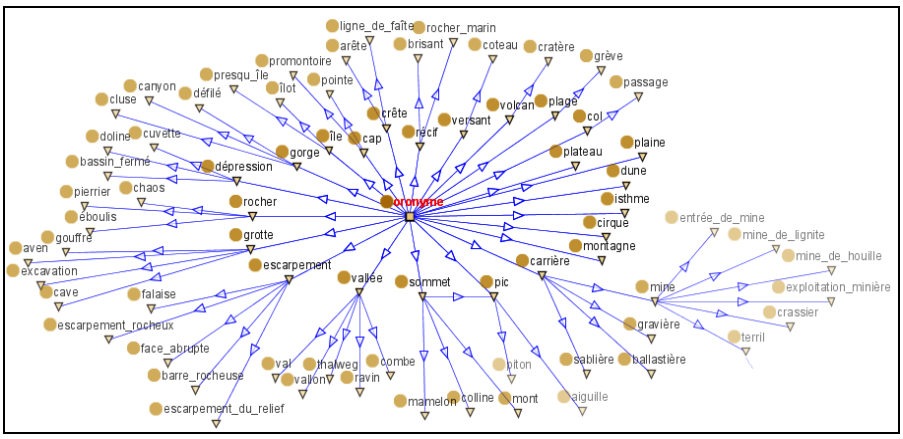


Fig. 2. Excerpt from the obtained taxonomy (in French)

First, our experiments have shown that a highly assisted process to identify relevant geographic concepts from textual specifications is possible. If this is not fully automated, a limited amount of interactive work is enough to reach reasonable results. We believe that this approach should then be extended to enhance the process and to take one step forward to a more complete ontology with properties, relations, definitions…

Second, if schemas of the studied databases contain a few tenths of classes (respectively 40 and 35), their respective taxonomies contain hundreds of concepts (respectively 580 and 500), which clearly shows the richness of the specifications in terms of semantics: each class is described by one term in the schema (e.g. 'river'[2]), while it actually refers to more than ten different geographic concepts in average (e.g. 'river', 'waterway', 'stream', 'canal', 'watercourse, 'rivulet'…).

Third, a more specific insight is that we discovered many geographic concepts in all classes related to toponymy. Without generalizing too much, this may be due to the fact that what is named is in general well known and then well described.

Forth, difficulties encountered may be interesting for an extension of such a work. Some difficulties where linked to the detection of concepts. The most difficult question we had to solve was: when do we consider that a nominal group is a geographic concept in its entirety (e.g. 'terraced house') or is a geographic concept qualified by an attribute (e.g. 'big house')? The answer is difficult and a matter of philosophy. For example, if we encounter the nominal group 'protestant church', we may choose to introduce in our ontology the concept of 'church', with an associated property (religion). However, if we rather encounter the term 'temple', we may

[2] All examples in this paper originate from actual databases and specifications from IGN-France. For the sake of clarity, even if the work has been done on texts in French, most examples in this paper are given in English, except in section 3.1 describing label comparison techniques, where a translation would have been meaningless. We are aware of imprecision due to this translation or adaptation; however we consider it as sufficiently insignificant for illustrative examples.

introduce it in its entirety in the ontology, with the synonym 'protestant church'. We believe there are no universal answer to that question, and that the only reasonable one in our context is to follow the principles followed by the developers of the database (did they separate or not 'protestant church' from 'catholic church' in the data?). Some other difficulties were linked to the hierarchisation of concepts. For example, some definitions of classes encountered in specifications are real definitions (e.g. "cape: prominent part of the shoreline…"), while some other definitions are more selection principles like (e.g. "hamlet: hamlet with a name'). As another example, some concepts encountered in definitions are real subtypes of the general concept corresponding to the name of the class, while some are not. For example in 'river = watercourse even in town', 'watercourse' can be thought of as a subtype of 'river', while 'town' of course not.

3 Comparing ontologies for assessing differences and commonalities

Once ontologies underlying two geographic databases have been determined, from the analysis of specifications as explained above or by any other mean, an important question needs to be answered: in what extent do these ontologies reflect the same conceptualization of the world? In other words, the question is: are the differences between the databases somehow artificial and only due to technical or terminological choices; or conversely, are they deeper differences that actually reflect different conceptualizations of the world? The answer is a key to assess how tractable, useful and meaningful is the integration of those databases, and how to do it.

Visser et al. [26] make a very useful distinction between ontology mismatches, identifying two types of basic mismatches: conceptualization mismatches which are mismatches between two conceptualizations of a domain, and explication mismatches which are mismatches in the way a conceptualization is specified. Considerable efforts have been devoted to the development of algorithms and tools that attempt to identify and resolve ontology mismatches in the field of ontology matching in general [6], and in the geographic context in particular [5]. Comparing vocabularies related to geography from different cultures also received attention in the literature [18]. These works usually focus on the analysis of differences between concepts, terms and definitions. However, as far as we know, few works exist on the global comparison of ontologies for gaining a compiled overview of differences and commonalities between them. In this section we first introduce works related to ontology matching, a necessary first step before a global comparison. Based on actual experimentations, we then exemplify which insights could be discovered from ontology matching. We finally introduce initial ideas towards a more global comparison of ontologies.

3.1 Ontology Matching

Due to the development of an ever-growing number of ontologies, ontology matching algorithms or techniques occupy a key role in facilitating the design of ontology-

based applications. The matching process aims at finding an alignment between two ontologies which express correspondences between their entities found according to a particular matching algorithm. Matching algorithms primarily provide *equivalence* relationships (*isEq*) meaning that the matched objects are the same or are equivalent. It is also possible to obtain *more specific* relationships (*isA*) meaning that a class is a sub class of another, *disjointness* when two classes are supposed to be disjoint or *semantically related* relations (*isClose*) for a link between two classes considered as related but without a specific typing of the relationship.

More formally, the matching process can be seen as a function *f* which, from a pair of ontologies to match *O* and *O*', an input alignment A which can be completed, a set of parameters p (e.g. weights, thresholds), and a set of external resources r, returns an alignment A' between these ontologies [6]: A' = f (*O*, *O*', *A*, *p*, *r*). All the parameters (*A, p, r*) are optional. Their use depends on the matching techniques performed by matcher tools. These techniques can be performed according to different approaches. We can distinguish individual algorithms [11][23] and combinations of the individual algorithms, either hybrid [15] when several individual algorithms are synthesized into a new one or composite solutions [4] allowing an increased user interaction.

Whatever the approach is, elementary techniques or algorithms that are used for solving the ontology matching problem exploit various types of ontology information, e.g. element names, data types, structural properties as well as characteristics of data instances. Based on the classification according to the kind of input described in [6], the following techniques can be distinguished: terminological, structural, extensional and semantic. Terminological techniques work on strings. Terms can be either considered as sequences of characters or be interpreted as linguistic objects. Structural techniques exploit ontology structures. This can be done at two levels, either by considering the internal structure of entities, e.g. attributes and their types, or by considering the relationships between entities. Extensional techniques work on data instances. Semantic techniques work on models. They require some semantic interpretation of the ontology and usually use some semantically compliant reasoner to deduce the correspondences. Furthermore, these current matching techniques can be complemented by using additional descriptions, called background knowledge. Some works assume that ontology matching can rely on a unique and predefined ontology that covers a priori all the concepts of the ontologies to be matched. Conversely, other works suppose that there does not exist a priori any suitable ontology. Hence, their idea is to dynamically select online available ontologies.

We illustrate now some of these matching techniques through TaxoMap [14] which has been used to match the two taxonomies built from the textual databases specifications mentioned in section 2. TaxoMap makes the assumption that most semantic resources are based essentially on classification structures which contain rich lexical information and hierarchical specification without describing specific properties or instances. Indeed, in practice, actual and available ontologies are mainly hierarchies of concepts, even if they could be much richer and more complex in theory. Hence, to find mappings in this context, we can only use the following available elements: labels of concepts and hierarchical structures. In TaxoMap, an ontology is considered as a pair (C, H_C) consisting of a set of concepts C arranged in a subclass hierarchy H_C. A concept c is defined by two elements: a set of labels and subclass relationships. The labels are terms that describe entities in natural language

and which can be an expression composed of several words. A subclass relationship establishes links with other concepts.

The matching process in TaxoMap is oriented from a source O_S to a target ontology O_T. It aims at finding one-to-many mappings between single concepts and establishing three types of relationships, equivalence, more specific and semantically related relationships.

TaxoMap mainly relies on terminological techniques. It performs a linguistic similarity measure between labels of concepts. The measure takes into consideration categories of words which compose a label. The words are classified as functional (verbs, adverbs or adjectives) and stop words (articles, pronouns) thanks to the use of TreeTagger [21], a tool for tagging text with part-of-speech and lemma information. Stop words categories enable to ignore these words in similarity computation. Functional words have less power than all the others (noun, etc.). The position of a word in the label is also of importance, a common word between two labels is less important after a preposition than a word that is a head.

Eight different matching techniques are implemented in TaxoMap, applied in a sequential way. The technique T_i will be performed on a concept C_S in O_S only if no correspondence with a concept C_T in O_T has been discovered with the techniques previously applied. Let C_S be a concept in O_S for which a correspondence has to be found and C_{Tmax}, C_{T2} and C_{T3} the three concepts in O_T having the best similarity measures with C_S, here is an overview of these techniques with some illustrations:

- Equivalence relationships identification technique (T_1): An equivalence relationship (C_S *isEq* C_{Tmax}) is generated when the similarity measure between one label of C_S and one label of C_{Tmax} is greater than a given threshold.

- Techniques based on label inclusion (T_2 ,T_3, T_7): These techniques consider inclusion of label words. According to T_2, (C_S *isA* C_{Tmax}) is proposed when one label of C_{Tmax} is included in one label of C_S without being behind a determiner. That way, *"*Route départementale*" isA "*Route*"*[3] is generated (cf. Fig. 3). Inversely, (C_S *isClose* C_{Tmax}) is proposed by T_3 when one label of C_S is included in one label of C_{Tmax}. T_7 is not performed immediately. Its aim is to identify hidden label inclusions.

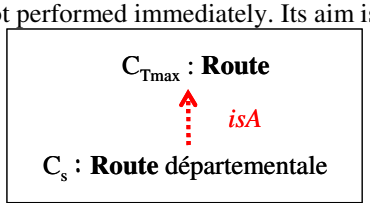


Fig. 3.: Illustration of the T_2 technique.

- Techniques based on relative similarity (T_4 ,T_5,T_6): These techniques are applied on C_S when no correspondence has been generated by the techniques T_2 and T_3 and when the similarity measure of C_{Tmax} is significantly higher than the measure of C_{T2}.

[3] In Enlish: "B-Road" isA "Road",

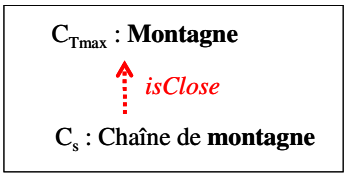


Fig. 4.: Illustration of the T_4 technique

For example, "Chaîne de montagne" *isClose* "Montagne"[4] is proposed according to T_4 (cf. Fig. 4). "Montagne" is included into "Chaîne de Montagne" but it is situated behind "de" denoting that this word is not of first importance in the expression "Chaîne de montagne".

- Techniques based on structure (T_8): This technique is applied after all those presented above. It is performed on C_S for which the similarity measure of C_{Tmax}, C_{T2} and C_{T3} is not very high (although greater than a given threshold) and when at least two of the concepts C_{Tmax}, C_{T2} and C_{T3} have a common father. In that case the relationship (C_S *isA CommonFather*) is generated, for example "Equipement de sport d'hiver" *" isA* "Entité topographique artificielle"[5] in Fig. 5.

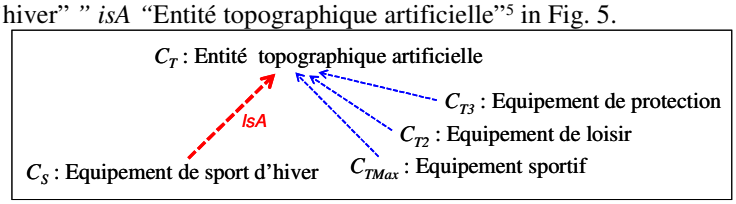


Fig. 5.: Illustration of the T_8 technique.

3.2 Insights from the analysis of alignments

The two taxonomies mentioned in section 2 have been aligned by means of techniques explained above. Concretely, we used the TaxoMap tool [14]. Some typical examples of what can be learned from this alignment illustrate its interest.

First, some effective alignments did map similar concepts expressed by different labels in the two taxonomies. These differences are examples of purely labelling differences and not conceptualization differences. This is the case for example for 'wooded area' mapped to 'clump'.

Some groups of concepts, existing in one taxonomy but not mapped in the other one, also illustrates the peculiarities of one taxonomy against the other one:

– This may emphasize the thematic choices behind the respective databases. For example, we identify such groups of concepts related to tourism ('tourist information office', 'seaside resort', 'seafront boardwalk' or 'historical downtown'), hydrographical details ('rivulet' or 'inlet'), or land use ('shrub area', 'rice swamp', or 'banana plantation').

[4] In English: "Mountain range" *isClose* "Mountain"

[5] In English : "Winter sport equipment" *isA* "Artificial topographic entity"

– This may also emphasizes that databases adopt different global approaches. In our experiments one database adopts a more functional point of view describing relations between features, compared to the other one that adopts a topographic point of view describing what is seen. Indeed, elements related to the description of network nodes appear more often in one taxonomy than in the other one (like 'diffluence' or 'cul-de-sac').
– The same type of analysis shows the difference of spatial level of detail between the databases: some high-level concepts may appear only in one taxonomy (like 'mountain range', 'town' or 'state').

A detailed analysis of the mapped concepts brings also some information about the different conceptualizations. Let us take a focused but significant example, the geomorphologic concept of 'cluse' (transverse valley, see fig. 6) exists in both taxonomies. However, in one taxonomy it is a subconcept of 'gorge', while it is more closely related to 'mountain pass' in the other one. This certainly originates from difficulties to classify such specific concepts. However, this may also be explained by the topographic point of view (a cluse is usually stip-sided valley, like a gorge) against the functional point of view (a cluse is a pass between valleys).

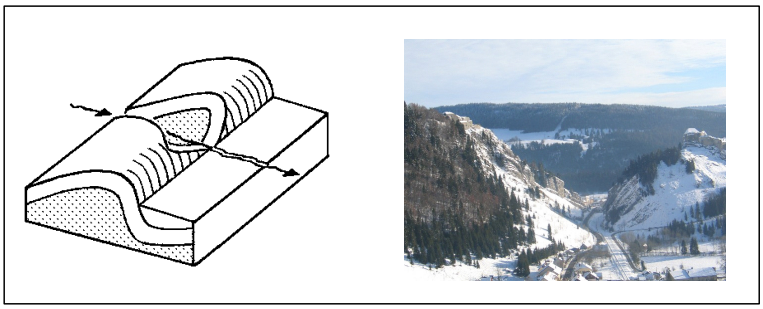


Fig. 6. Typical geomorphological shape of 'cluse'

We can hardly generalize too much from those focused examples. In order to make such a generalisation and assess the differences between conceptualizations underlying the two studied databases, some methods for a more systematic and global comparison of ontologies are needed or, in other words, distances between ontologies should be defined. Some directions for such distances are expressed in the next section.

3.3 Toward a global comparison of ontologies

Two main groups of distances between ontologies could be defined: the one relying on ontologies previously mapped, and the other ones [7]. Some distances, for example based on extensions of distances between concepts developed in the field of ontology alignment, are proposed and analysed according to their theoretical properties in [7]. Many measures can be defined. One major difficulty may be to interpret their meaning according to the intended use of the distance. For example: which measures

could be used to assess if ontologies follow the same conceptualization but with different levels of detail? Which measures could be used to assess that ontologies are compatible or not? We hereafter develop some general ideas for meaningful and interpretable measures of differences and commonalities between ontologies, based on various fields of research on ontologies.

Partition of ontologies. Works on ontology partitioning can help to compare ontologies, especially when the partitions are built by taking the alignment objective into account. In [13], two methods which transform the two ontologies to be aligned into two sets of blocks of a limited size are proposed (see fig. 7). Partitioning a set E consists in finding disjoined subsets E_1, E_2, …, E_n, of elements semantically close i.e. connected by an important number of relations. The realization of this objective consists in maximizing the relations within a subset and in minimizing the relations between the different subsets. The proposed partitioning methods are partially inspired by co-clustering techniques which consist in exploiting, besides the information expressed by the relations between the concepts within one ontology, the information which corresponds to the inter-ontology relations between concepts such as equivalence relationships. The partitioning process brings together the concepts that have relations between them in blocks. Both ontologies are partitioned one after the other. Blocks of the second ontology are built around sets of concepts whose label is equivalent to the label of concepts belonging to the same block in the first partition. That way, we obtain blocks in the two ontologies that correspond, e.g. containing concepts with equivalent labels and semantically close. An analysis of the two partitions and of the pairs of blocks can allow answering the following questions. Given two ontologies O and O', what is the corresponding part in O' of the block B_i in O? Are there parts with no correspondent? Are the parts similar according to the number of concepts with an equivalent label? Indeed, a high number of equivalent concepts may correspond to parts which are very closely related. Are corresponding parts different according to the number of concepts or to the depth of the concepts hierarchy? Is a description more refined or more precise than the other one?

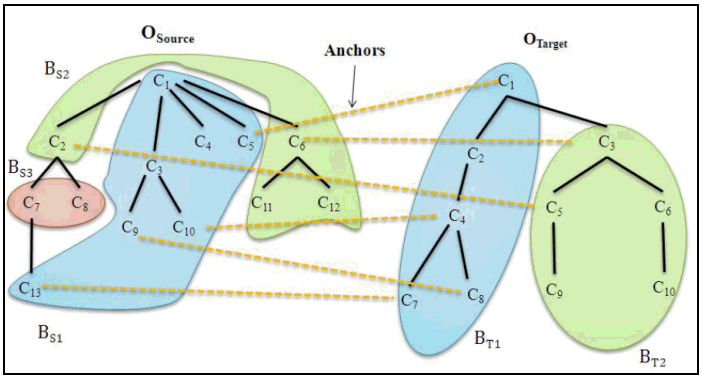


Fig. 7. Partition of ontologies in a matching context

A partitioning experiment has been done with the two ontologies mentioned in section 2. A preliminary analysis of the results attests that such an approach can

contribute to improve the understanding of the ontologies and to show a number of differences or common points. More precisely, the partitioning approach summarizes the themes described by one of the ontology (the first one which is partitioned) and allows answering the following question: are the same themes described in the second ontology? In our experiment, the first ontology has been decomposed into five blocks, each block having its correspondent in the second one. This forms five pairs of blocks grouping concepts related to: (1) natural topological entities (maritime space, element of the relief, ground hydrography), (2) administration buildings, entities with industrial vocation, sport equipments, … (3) infrastructures of transport, (4) industrial buildings, entities with agricultural vocation, equipments with military vocation, (5) cemetery, religious buildings, elements of the patrimony, equipments of leisure. These five blocks are indicators of five main topics described by both ontologies.

The number of concepts contained in the first three pairs of blocks is very similar, which means that the themes are equally described. Some blocks as that concerning the natural topological entities contain a high number of concepts linked by an equivalence relationship. The descriptions in both ontologies may be very close. They may correspond to knowledge expressed by the same point of view.

On the opposite, the number of concepts in the last two pairs of blocks differs. The corresponding themes are less prominent in the second ontology. Some blocks in the second ontology have no correspondent in the first, as that concerning the sea links. Some concepts are isolated as urban centre or natural obstacle. The observation of such isolated blocks or concepts leads to an analysis at various levels of granularity, either at a theme level or at a concept level. Furthermore, note that the maximal size of the blocks is a parameter in the partitioning process. Successive experiments can be made with different values. We can also apply the partitioning process on the whole ontologies and then reapply it on the pairs of blocks. This allows an analysis at various levels of detail and can lead to a more precise understanding of the geographical coverage of both ontologies. Small parts of models are easier to understand.

Evaluation of ontologies. Works on comparison of ontologies may benefit from those concerning the evaluation of ontologies which need to assess all their important features. Despite works in that domain lack of automatic, well grounded, methodologies, it seems to us important to briefly present them in order to identify a number of focus areas for future research. Ontologies may be assessed from different angles. Ontology evaluation can include aspects of ontology validation and verification, i.e. structural, functional, and usability issues [9]. An ontology can be evaluated against criteria based on its content coverage, for example by using a corpus describing the domain of interest. Some ontology search engines adopt a Page-Rank-like method to evaluate and rank ontologies by analysing links and referrals between the ontologies in the hope of identifying the most popular ones. Other systems for ranking ontologies are based on a number of measures that assess the ontology in terms of how well it represents the concepts of interest expressed by users. Such analysis metrics could be useful for our comparison purpose. For example, four ranking measures are applied in AKTiverank [3] in order to evaluate different representational aspects of the ontology and calculate its ranking. One of the

measures applied is the density measure (DM). It includes how well the concept is further specified (the number of subclasses), the number of properties associated with that concept, number of siblings, etc. DM is intended to approximate the representational-density or information-content of classes and consequently the level of knowledge detail. Another measure is the Centrality Measure (CEM) which aims to assess how representative a class of an ontology is. It assumes that the more central a class is in the hierarchy, the more likely it is for it to be well analysed and fully represented. These measures could be used to compare ontologies.

Ontologies and fitness for use. Other works aiming at detecting and retrieving relevant ontologies need means for measuring the similarity between ontologies. So, in [16] a set of measures that capture the similarity between ontologies at two different levels, the lexical and the conceptual levels, is proposed. Those similarity measures describe the coverage of one ontology specification by another. At the lexical level, labels of concepts are compared; their similarity based on the Levenshtein minimum string distance is computed in order to determine the best measure for each label. The average of the similarity measure of all the labels, AVG(SM(L_1,L_2)) determines the coverage of the vocabulary L_1 of an ontology O_1 by the vocabulary L_2 of the other ontology O_2. This is an asymmetric measure. When L_2 contains all the strings of L_1, but also plenty of others, then AVG(SM (L_1,L_2)) = 1 but AVG(SM (L_2,L_1)) may approach zero. The conceptual similarity is based on the intentional semantics of a concept C in an ontology O, IS(C,O), defined as the set of all its super- and subconcepts in O. When a concept C belongs to two ontologies, one can define the taxonomic overlap (TO) between O_1 and O_2 for this concept, denoted TO(C, O_1, O_2) and defined as the ratio between the number of common elements in the intentional semantics of C in O_1 and in O_2 and the total number of elements belonging to the union of these two sets. If a concept C is in O_1 but not in O_2, an optimistic approximation of TO(C,O_1,O_2) is defined as the maximum overlap obtained by comparing IS(C,O_1) to the intentional semantics of all the concepts in O_2. The average of the taxonomic overlaps allows comparing semantic structures of the two ontologies O_1 and O_2.

Visualisation of ontologies. Works on visualisation and understanding of very large and complicated ontologies can be helpful for comparison of ontologies because their objective is to provide global views. The main feature of the visualization approach described in [25] is that it presents a large-scale ontology by a holistic “imaging” which is semantically organized for quick understanding of the subject and the content. Furthermore, the approach has to assess the importance of classes because when displaying the layout, only the most important classes that fall into the screen are labelled. The importance is computed based on the class hierarchy by a formula which first part gives more importance to the classes higher in the hierarchy, while the second part gives more importance to the classes with more descendants. In [27], CARRank, an automatic ranking algorithm, which can be integrated in existing ontology visualization tools, is described. It is a tool to help understand ontologies based on the identification of potentially important concepts and relations user-independently. The importance of concepts and the weights of relations reinforce one another in CARRank in an iterative manner.

Analogy between conceptual and physical worlds. In order to compute and interpret distances between ontologies, an analogy could be made between spatial networks, like road networks, and ontologies that also rely on a graph structure. Indeed, an analogy between the geographic space and the conceptual space could be easily understood, at least for geographers and spatial analysts. In the field of spatial analysis, there exists a lot of works to analyse (like works studying accessibility or vulnerability of networks [12]), simplify (like works on generalisation of networks for mapping [24]) or compare spatial networks (like works on network matching (cf. fig. 8, [19]). These works use measures similar to the one previously mentioned about ontologies, or more specific measures taking advantage of the geometric aspect of spatial networks. We think that both fields of research, spatial analysis and ontology analysis, would gain from a mutual comparison and enrichment, for developing new methods.

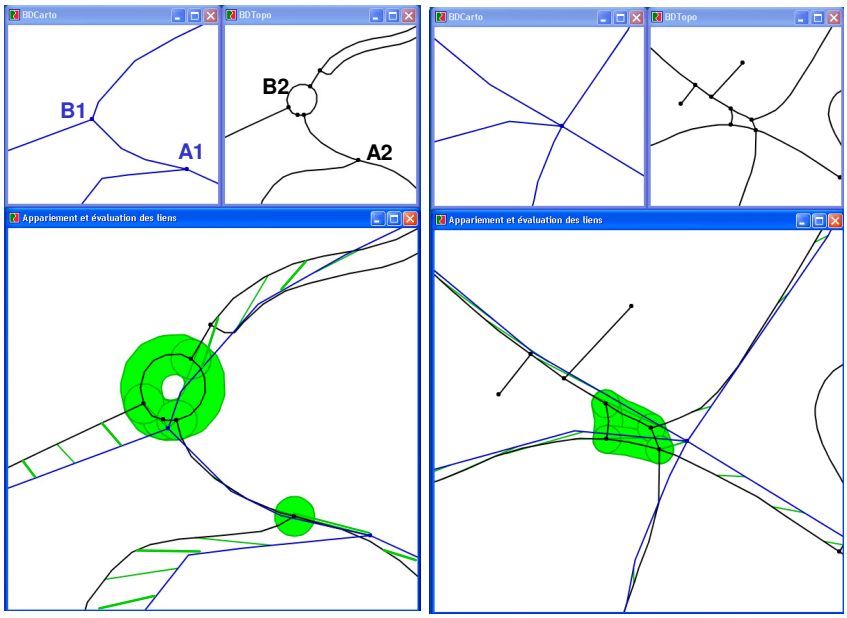


Fig. 8. Matching of networks with different levels of detail, from [19]. Examples where a roundabout (on the left) or a square (on the right) can be mapped to a single node in a less detailed network, like some groups of concepts in an ontology could be matched to a single concept in another one.

3 Conclusion

In this paper we argue that assessing how much two geographic databases reflect the same point of view is a key issue for data integration. This necessitates distinguishing

in data schemas between differences originating from actual conceptual choices from those originating from technical reasons. We have shown in some experiments that ontologies, or at least taxonomies, reflecting the point of view of a database can be derived from textual documents like data specifications. The derivation can possibly be assisted by means of natural language processing techniques. These ontologies can then be linked and compared by means of ontology matching techniques. Some of our experiments show that interesting information, like differences between the thematic points of view or the conceptual levels of detail could be derived from the analysis of ontology matching. We then claim that methods to globally compare ontologies are needed, and different field of researches could be fruitful sources of inspiration for this, from ontology visualization to spatial network matching.

We are convinced that such methods for comparing ontologies should have two main properties, even more important than being precise and efficient: the first one is the possibility to make a meaningful interpretation of the results of the methods; the second one is to be adapted to 'light' but actually existing ontologies, which are more or less hierarchical taxonomies.

Acknowledgement

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

References

1. Abadie N. and Mustière S. : Constitution d'une taxonomie géographique à partir des spécifications de bases de données. In : SAGEO'2008, Montpellier, France (2008).
2. Aerts K., Maesen K. and van Rompaey A.: A practical example of semantic interoperability of large-scale topographic databases using semantic web technologies. In: 9th AGILE Conference on Geographic Information Science. College of Geoinformatics, University of West Hungary, pages 35-42 (2006).
3. Alani H., Brewster C., Shadbolt N.: Ranking Ontologies with AKTiveRank. In Cruz I., Decker S., Allemang D.,Preist C., Schwabe D., Mika P., Uschold M., Aroyo L.M. (Eds). ISWC 2006, LNCS. vol. 4273, Springer, Heidelberg (2006).
4. Doan A.-H., Madhavan J., Domingos P., Halevy A.: Ontology Matching: a machine-learning approach. In Steffen Staab and Rudi Studer, editors, Hondbook on Ontologies, chapter 18, p. 385-404, Springer Verlag, Berlin (DE), (2004).
5. Egenhofer M.J., Clementini E. and Di Felice P.: Evaluating inconsistencies among multiple representations. In: 6th International Symposium on Spatial Data Handling (SDH'94), pp. 901-920 (1994).
6. Euzenat J., Shvaiko P.: Ontology Matching. Springer-Verlag Berlin and Heidelberg GmbH & Co. (2007).
7. Euzenat J.: Quelques pistes pour une distance entre ontologies. In atelier « Mesures de Similarité Sémantique » de la conférence EGC, Extraction et Gestion de Connaissances. (2008).

8. Fisher P. F.: Models of uncertainty in spatial data. In: Geographical Information System, P.A. Longley, M.F. Goodchild, D.J. Maguire, D.W. Rhind (eds), vol. 1, Second Edition, pp. 191-203. Wiley (2003).
9. Gangemi A., Catenacci C., Ciaramita M., Lehmann J. : A theoritical framework for ontology evaluation and validation. In Proceedings of SWAP 2005, (2005).
10. Gesbert, N.: Etude de la formalisation des spécifications de bases de données géographiques en vue de leur intégration. Phd thesis, University of Marne-la-Vallée (2005)
11. Giunchiglia F., Shvaiko P.: Semantic Matching. The Knowledge Engineering Review. 18(3):265-280, (2003).
12. Gleyze J.-F., 2008, Using structural approach to understand transportation networks vulnerability. Proceedings of European Geosciences Union 2008, Vienna, Austria.
13. Hamdi F., Safar B., Zargayouna H., Reynaud C.: Partitionnement d'ontologies pour le passage à l'échelle des techniques d'alignement. 9èmes Journées Francophones « Extraction et Gestion des Connaissances », Strasbourg, France (2009). Paper having received the price of the best application paper.
14. Hamdi F., Zargayouna H., Safar B., Reynaud C.: TaxoMap in the OAEI 2008 alignment contest. Proc. of OAEI 2008 in cooperation with the ISWC Ontology matching Workshop, Karlsruhe, Germany (2008).
15. Madhavan J., Bernstein P., Rahm E.: Generaic Schema Matching with Cupid. In Proc. 27th International Conference on Very Large Data Bases (VLDB), p. 48-58, Roma (IT), (2001).
16. Maedche A., Staab S.: Measuring Similarity between Ontologies. In proceedings of the 13th International Conference, EKAW 2002, Sigüenza, Spain. A. Gomez-Perez, V.R. Benjamins (Eds.), Springer Berlin 2473 . pp. 251-263, (2002)
17. Malczewski J.: GIS-based multicriteria decision analysis: a survey of the literature. International Journal of Geographical Information Science, vol.20, n.7, pp.703-726 (2006).
18. Mark, D., Turk, A. and Stea, D.: Does the Semantic Similarity of Geospatial Entity Types Vary Across Languages and Cultures? In: Workshop on Semantic Similarity Measurement at the Conference on Spatial Information Theory (COSIT 2007), Melbourne, Australia (2007).
19. Mustière S., Devogele T.: Matching networks with different levels of detail. GeoInformatica, vol.12, n.4, 12/2008, pp.435-453 (2008).
20. Partridge, C.: The role of ontology in integrating semantically heterogeneous databases. Technical Report 05/02 LADSEB-CNR, Padoue (2002)
21. Schmid H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. International Conference on New Methods in Language Processing (1994).
22. Sheeren, D., Mustière, S. and Zucker, J.-D.: A data-mining approach for assessing consistency between multiple representations in spatial databases. International Journal of Geographical Information Science, DOI: 10.1080/13658810701791949, first published online 11/09/2008. Taylor and Francis (2008).
23. Stumme G., Mädche A.: FCA-Merge: Bottom-up merging of ontologies. In Proc. 17th Internatinal Joint Conference on Articicial Intellignece (IJCAI), p. 225-234, Seattle (WA US), (2001).
24. Thomson R. and Brooks R., 2007. Generalisation of Geographical Networks: In Generalisation of Geographic Information: Cartographic Modelling and Applications, Mackaness, Ruas, Sarjakoski (eds), Elsevier, 2007.
25. Tu K., Xiong M., Zhang L., Zhu H., Zhang J., Yu Y. : Tpwards Imaging Large-Scale Ontologies for Quick Understanding and Analysis. In Gil Y., Motta E., Benjamins V.R., Musen M.A. (Eds). ISWC 2005. LNCS 3279. Springer, Heidelberg (2005).

26. Visser P. R.S., Jones D. M., Bench-Capon T. J. M., Shave M. J. R.: An analysis of ontological mismatches: Heterogeneity versus interoperability. In AAAI 97 Spring Symposium on Ontology Engineering, Stanford, USA, 1997.
27. Wu G., Li J., Feng L., Wang K.: Identifying Potentially Important Concepts and relations in an Ontology. 7th International Semantic Web Conference (ISWC2008), Karlsruhe, Germany. Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (Eds.), LNCS 5318, Springer (2008).