Semantically Valid Alignment in the "Semantic Web": The Problem of Grounding

Andrew U. Frank
Department of Geoinformation and Cartography
Gusshausstrasse 27-29/E127-1
A-1040 Vienna, Austria

Abstract

Integration of semantics is a precondition for the integration of data; efforts to integrate or align the semantics expressed in various ontology languages are reported in the literature. The approaches are characterized by the assumptions they make about the known commonalities between the two (or more) sets of classes, described by the ontologies to align, but no-one can proceed without some known common elements in the ontologies. Explicitly introducing the distinctions that differentiate between the classes reduces the number of common concepts that must be established before alignment. A few classes describing water bodies in three languages are used as an example.

1 Introduction

Semantics, i.e., the meaning of symbols and words, is the driving force in the “web 2.0” a.k.a. “Semantic Web”. Before data from different sources can be used intelligently they must be brought to a common set of meanings; this is often called semantic integration or alignment. The task is one of translation between web datasets that were created by independent users producing an expression of a real world situation (reality, including socially constructed reality (Searle 1995)) using their semantics for the symbols used. The problem is comparable to the translation of natural language semantics, a known hard problem with no immediate solution in sight.

Many papers, articles and PhD theses have been published, discussing the practically very important problem of alignment of data descriptions (for an extensive but still partial review see, e.g., Lemmens 2006). These approaches describe various methods how to align data descriptions and differ mostly in their tacit assumptions what commonality exists between the descriptions.

Alignment of semantics requires always some identified common concepts from which other commonalities can be inferred. It is useful to classify approaches to semantic alignment by what is assumed as common: often a common terminology is the starting point, other approaches compare structural similarities, etc. From this analysis emerges the approach presented here: the alignment of semantics of classes (i.e., a taxonomy) can be automated if the distinctions between the classes are aligned first: classes in a taxonomy must be distinct from other classes; these distinctions can be expressed and the distinctions used in two different taxonomies compared and the common distinctions aligned. Given that there are less distinctions than there are classes, focusing on aligning distinctions (from which the alignment of the classes follows automatically) reduced the non-automated, manual effort considerably. The approach is described using an example, originally analyzed by Mark (1993), of geographic terminology for standing water bodies in three languages (English, French, and Spanish).

The semantics of a dataset can be formalized in various way, Protégé (2000) or OWL is a popular tool respective a popular formalization. In general the description of semantics for current purposes can be seen abstractly as a graph where nodes are classes linked by *is_a* and *part_of* relations, and schema alignment as a graph mapping. It may be surprising that ontology research is very confused in its terminology; even using a single tool like Protégé to write an ontology for an application in

OWL one is faced with two different and only approximatively translatable terminology. It is therefore not easy to select a consistent terminology for a discussion of ontology alignment.

In the abstract, the alignment problem is: Given two datasets, which should be used jointly to answer some query. The semantics of the databases is given as hierarchies or heterarchies (single or multiple inheritance) of classes. The integration requires that their data description, for example, their relational database schema, which describes the semantics, are aligned or integrated. I will not differentiate between these two terms, because the mappings are not fundamentally different. In the classes from A and B are mapped to a common set of classes C by two mappings $g : A \rightarrow C$, and $h: B \rightarrow C$; the mapping $f : A \rightarrow B$ directly is the composed of $f = h'. g$ (where h' is the inverse mapping of h).

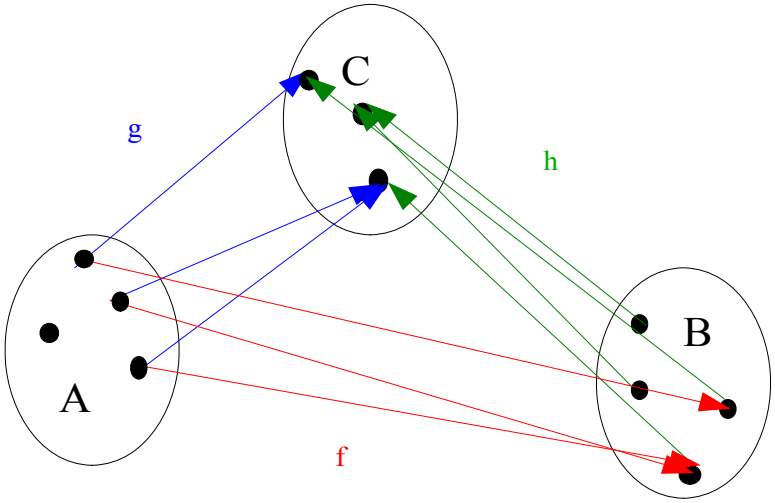

Figure 1: Mapping between two ontologies

2 Running Example: Geographic Terms for Standing Water Bodies

Mark has discussed the semantics of geographic terminology describing standing water bodies in three languages, namely English, French, and Spanish (Mark 1993); the set of terms includes in English “Lake”, “Pond”, and “Lagoon” and the corresponding French and Spanish terminology (see list 1 and 2). To build a pan-European geographic database to answer questions posed in one of the three languages, e.g., what is the largest lagoon in France, a semantically valid alignment between the terms is necessary.

List 1: BH080 (DGIWG 1992, p. A-66, quoted after (Mark 1993))

US	Lake/Pond
FR	Lac/Etang
SP	Lago/Laguna

List 2: BH190 (DGIWG, 1991, p.A-71, quoted after (Mark 1993))

US Lagoon/Reef Pool
FR Lagon/Lagune
SP Albufera

I will use this example throughout the paper, it seems to be more valid than a toy example of my own invention but still small enough and easy to understand.

3 Assumption about Communality

The alignment of semantics of data from different sources requires necessary some points of commonality from which other mappings can be inferred. Different approaches mostly differ, in what they assume as commonalities as inputs into the alignment process. Often commonalities are assumed tacitly. It appears useful to classify research in data description alignment by what they assume as common in the two descriptions. Such a classification could be exhaustive, as any alignment method must assume some commonalities.

3.1 Same Vocabulary with Same Semantics

Early research assumed that the classes where described by descriptive labels in both sources with the same words with the same semantics. This assumption simplifies the alignment between classes and this early research did concentrate on syntactic alignment between differences in the data models.

The assumption of common terminology (with the same semantics) is not realistic. Even if both data sources are described in the same natural language, the same words used in the two different contexts of their respective application field do not automatically mean the same thing. For example, in northern Germany, the German word 'Meer' means a lake, whereas in southern Germany 'Meer' means 'ocean'; the xx meer is a lake of xx sq km 30 km north of Bremen. Mark (1993) points out in his analysis that, for example, the French distinction between 'lac' and 'étang' is comparable to the English distinction between 'lake' and 'pond' in Quebec French, but not in France, where some 'étang' are very large; in Cajun French 'étang' is not a geographic term and 'ponds' are called 'marais' (which in France would be a swamp).

3.2 Data Descriptions Are (Mostly) Expressed in Terms of a Standardized Vocabulary

Kuhn (1996) advocates a semantic reference frame, comparable to the spatial reference frames used to indicate spatial position (e.g., with a GPS), is close to this approach: it suggests that some semantic framework is fixed and used by both (preferably all) databases. Most interpretations of the concept of a semantic reference frame implies that a fixed terminology with a standardized semantics is used.

Professional education and the corresponding organizations establish national and international standards (DIN, ISO), which include standardization of vocabulary. Some such efforts are directed primarily to standardize only the vocabulary, e.g., the names for geographic projection systems listed by the Petroleum engineers. The U.S. Digital Geographic Information Standard established in 92 by the USGS is a similar example of an attempt to authoritatively define classification and terminology (DGIWG 1992). Such taxonomies are not likely consistent, reflecting inconsistencies in natural languages. Newer approaches (e.g., in medical terminology) achieve more consistency by employing formal ontologies, in particular fixing an upper level ontology. Integration of definitions of semantics starting with the same upper level ontology is improved, but integration of definitions using different upper level ontologies require first an integration of the upper level ontologies. Surprisingly, mappings between the major upper level ontologies, e.g., SUMO (Niles 2003), Ontos

(ONTOS 2001), Cyc (CYC 2000), etc. are not know.

Some research in semantic alignment assumes that the same standardized vocabulary is used for both data sources. This approach is sometimes viable for new data collections for which the standardized terminology and semantics can be imposed during data collection. Later translations of existing data, collected with other description to a standardized terminology can be approximative at best—and is a very demanding task that requires intimate knowledge of the intentions of the descriptions used for data collection and the ones of the standardized terminology; this task is currently not suitable for programmed execution and requires human effort.

A major criticism against the use of standardized terminology is based on cultural differences. Mark points out—quoting Whorf (Carroll 1956)—that meaning of words is a cultural agreement necessary for effective communication. Societies construct appropriate classifications (taxonomies) that accentuate the distinctions of importance in the environment; for example, an agricultural society makes distinctions between classes of animals important for farming (cows, heifers, oxen, bulls, calves), whereas—in the same language—inhabitants of towns hardly differentiate between cows and goats. Imposing a standardized vocabulary may occult relevant distinctions and accentuate distinctions that are not important for a society with different climatic or technological environment. Newer research by Mark et al. on ethnophysiography (2007) demonstrates carefully and in detail the differences between cultures in the way they conceptualize and classify geographic features (Mark 1993).

3.3 Constructing Same Vocabulary with Structural Similarity

After common terms are identified, the alignment problem is restricted to align the terms used in the data descriptions that are not part of this vocabulary. Alignment research proposed to exploit the structure, expressed in the *is_a* and *part_of* relations and match these; starting from nodes that were identified by same vocabulary. Such an approach may as often fail as it succeeds; structural similarity at this level does not indicate semantic similarity, it is easy to construct realistic descriptions of data where such alignments produce misleading matches and, as a consequence, wrong database query results see (Frank 2008).

3.4 Comparing Attributes of Classes

Matching terms by comparing the attributes is another option (assuming tacitly that the attribute names come from the same vocabulary). The terms representing similar concepts are typically characterized with the same attributes. Terms that are characterized with the same attributes are good candidates for matching, but again, the indication can easily be misleading. Having the same attributes is neither a necessary nor a sufficient condition for matching. For example, '*Building*' can be characterized in one dataset with surface area and street address, and in another dataset with an number of floors and area of inhabitable space; in this second dataset, *parcels* have surface area and street address; but matching *building* and *parcel* based on same attributes is clearly in error.

3.5 Comparison of Measurement Units

If the attribute names are not from a common vocabulary (for example when merging datasets from different countries), then identifying matching attribute names is a prerequisite for matching classes. Object properties are measured on particular scales and expressed with specific measurement unites. Matching classes must have—as far as present—the same properties expressed with the same measurement units. Therefore combination of measurement units used for the properties of an object can be used to identify object classes (Fallahi, et al. 2008). This approach is probably more suitable for confirming identifications between object classes and only exceptionally serve to detect

identical object classes.

3.6 Distinction Grounded

The discussion in Mark's COSIT contribution is based on the properties of features that are used for the classification. For example, the distinction between the English terms 'lake' and 'pond' is one of size: Lakes are big, ponds are small standing water bodies. Such distinctions can be systematically used to construct the taxonomy (Frank 2006): Classes are distinguished by some property: a canal is different from a road for example by the presence of water, a building is different from a tent by use of solid vs textile material for the walls, etc. Starting from a set of 'distinctions' a (large) set of classes are generated (from n binary distinctions n^2 classes are generated). Distinctions can be used to construct an alignment between two taxonomies. The approach is attractive to align the distinctions used in two taxonomies, and then to construct automatically the alignment between the taxa, because there are much less distinctions than classes, which reduces the effort that requires human input and cannot be automated currently. The next section describes the necessary process using the example from section 2.

4 Alignment Using Distinctions

The alignment using distinctions proceeds in the following steps:

1. Describe the classes of both datasets by the distinctions necessary to generate them.
2. Identify the distinctions that are common for both semantics.
3. The power of the union of all distinctions, yielding all potential classes.
4. Identify the classes from both sources among the potential classes. Some of these classes that are linked to both sources, some only to one. The set of all classes linked to one or the other source schema are the used classes. With the distinctions the *is_a* and *part_of* relations between the used classes are determined and queries against this database schema are possible.

This process is grounded by the identification of distinctions (step 1 and 2), which is done by human experts. The process is effective, because the number of distinctions is much smaller than the number of classes and the effort to identify the common classes is harder than for the distinctions. The rest of the alignment process (step 3 and 4) is fully automated. The following subsections apply the method to the terminology of standing water bodies.

4.1 Describe Classes by Distinctions

English: Following Mark's analysis, I suggest that *lake* is used polysemous with two meanings: *lake*$_1$ as a generic 'enclosed standing water body', superordinate to *lake*$_1$, which is a large fresh water body, distinct from *lagoon*, which is a coastal water and a *pond*, which is a small water body.

French: Mark, in figure 2, suggests that the distinction between *lac* and *étang* is not the size of the water body, but the presence of marshy edges. *Étang* is used polysemous and *étang*$_2$ describes coastal lagoons (brackish water). *Lagune* is, following again figure 2, used for round enclosures of sea water.

Spanish: *Lago* covers the same meaning as English *lake*, but *laguna* is polysemous and means “1. Pond, lake, a large diffusion of stagnant water, marsh. 2. an uneven country, full of marshes” (Velazques de la Cadena, 1973, p 402—quoted after (Mark 1993). Man-made small water bodies are called *tanque* in Spain large man-made water bodies are called *embalse* in Spain. Excluding the use of Mexican terminology, a coastal water body is called an *albufera* (lagoon

would be used in Mexico and Texas).

4.2 Identification of Distinctions Used

The distinctions used are:

- English: size (large-small), inland-coastal.
- French: edge (marshy or not), water (fresh-sea), form (round)
- Spanish: marshy or not, natural—man-made, inland-coastal.

Assuming that the concepts described here in English terms match reasonably the differentiations a native speaker would make and relate to the same real world aspects of geographic objects, then these 2 or 3 distinctions used in each of the 3 languages, can be reduced and identifies to the following 5 distinctions.

- size (large-small),
- inland-coastal (including fresh or sea water),
- edge (marshy edges or sharp),
- natural (vs man-made)
- form (round).

4.3 Potential Classes

The four binary distinctions produce a lattice with maximally 32 taxa, not all of them meaningful or used. Only 9 taxa, are actually used in one of the three languages (or form superclasses to existing taxa); it does not show, for example, the theoretically possible class of coastal waters with marshy edge distinguished form coastal waters with sharp edges, or small, marshy edge water body distinguished from small sharp edge water bodies, because such a class does not appear in any of the three languages.

4.4 Identify the Classes and Characterize Them by Distinctions

The taxa in the three languages can be characterized by their distinctions, as shown in the following table:

	Inland + coastal -	Large + small -	Marshy + edge	Natural + man-made -	Round +
EN					
$Lake_1$	O	O	O	O	O
$Lake_2$	+	+	O	O	O
Pond	+	-	O	O	O
Lagoon	-	O	O	O	O
FR					
Lac	+	O	-	O	O
$Étang_1$	+	O	+	O	O
$Ètang_2$	-	O	O	O	-
Lagoon	-	O	O	O	+
SP					
Lago	+	O	O	+	O
Laguna	O	-	+	O	O
Albufera	-	O	O	O	O
Embalse	+	+	O	-	O
Tanque	+	-	O	-	O

Table 1: The term characterized by distinctions (+ = applies; - = does not apply; O *= indifferent*

4.5 Mapping between Taxa

From the Table 1 one can read off the relations between the taxa in the different languages. For example, *$Lake_1$* is superordinate to all of the taxa included (except laguna, which is not a body of standing water according to the dictionary entry quoted).

One can also observes that there are nearly no clean mappings between taxa: English *$Lake_2$* does not separate man-made from natural and large, nothing equivalent in other languages and only *embalse* as a new subordinater *Lago* is natural, but independent of size; neither is in a superordinate/subordinate relation. Alignments are thus more likely, if the focus of the application restricts the distinctions that are meaningful, respective the distinctions that can be left out.

5. Conclusion

Aligning two descriptions of datasets in order to process queries against the two datasets jointly

cannot be fully automated, at least not till a full, automatic understanding of human natural languages is achieved. The various methods proposed to establish mappings or alignments between data descriptions differ in what they assume as commonality between the dataset descriptions.

Approaches assuming a common vocabulary, either standardized or 'natural language' are problematic, as a standardized vocabulary fix a conceptualization, which is not necessarily appropriate for an application and the use of natural language terms glosses over differences in the conceptualization hidden under the common vocabulary (remember the saying “England and America are separated by a common language” attributed to George Bernard Shaw).

To show the limitations of approaches that use structural similarity, it is sufficient to construct example cases, where strong structural similarity is found among semantically unrelated classes. Structural approaches could only be useful when grounded with some already identified classes using other methods and run even then the risk to produce nonsensical mappings, because structural similarity is neither a necessary nor sufficient condition for semantic similarity.

Most of data descriptions come in form of taxonomies. The observation that taxa must be distinguished from each other and few distinctions generate a rich ontology suggest that alignment starts with an alignment of the distinctions, not the taxa. Aligning distinctions by human experts is less arduous than aligning the taxa. From an aligned set of distinctions the aligned taxonomy is produced automatically. The approach focusing on distinctions is therefore affective; the necessary reasoning can, for example be done in the Protégé or OWL description language framework and extensions to include other semantic inputs than the taxonomic *is_a* relations should be researched.

The example given by Mark in a paper 15 years ago hinted at distinctions, but did not use a formal approach. Formal concept analysis (Mineau et al. 1999) and its application to ontologies (Frank 2006) gives a framework in which this example can be reworked. It shows, first, that alignment using distinctions is possible and the procedure effective, helping to focus the attention to the relevant distinctions and providing a formalism for notation. It shows, second, that alignment between taxa is not necessarily working well. The example shows very few clear cases of alignment —different languages make different distinctions, which only allow a mapping, if the application indicates which distinctions are important and which other can be neglected. It is appropriate to remind interoperability researchers of a book by Umberto Eco “Dire quasi la stessa cosa” (To Say Nearly the Same Thing) (Eco 2003) in which he argues that a perfect translation is always a trade-off between what can be kept and translated and what must be sacrificed.

Acknowledgments

I thank David Mark for the example and other contributions he made over the years to further our understanding of semantics of geographic terminology.

References:

Carroll, John B. Language, Thought and Reality - Selected Writing of Benjamin Lee Whorf. Cambridge, Mass.: The MIT Press, 1956.

CYC. The Cyc Corporation Web Page 2000 [cited 21 November 2000]. Available from http://www.cyc.com/tech.html.

DGIWG. "The Digital Geographic Information Standard (Digest): Part 4, Feature and Attribute Coding Catalog (Facc)." Washington, DC: U.S. Defense Mapping Agency., 1992.

Eco, Umberto. Dire Quasi La Stessa Cosa. Milano: Bombiani, 2003.

Fallahi, G., M. Mesgari, A. Rajabifard, and A.U. Frank. "A Methodology Based on Ontology

for Geo-Service Discovery." World Applied Sciences Journal 3 (2008): 2.

Frank, A. "Distinctions Produce a Taxonomic Lattice: Are These the Units of Mentalese?" Paper presented at the International Conference on Formal Ontology in Information Systems (FOIS), Baltimore, Maryland, 9.-11. Nov. 2006.

Frank, A.U. "Similarity Measures for Semantics: What Is Observed?" Paper presented at the Special Issue on Semantic Similarity Measurement and Geospatial Applications 2008.

Hornsby, Kathleen, and Max J. Egenhofer. "Qualitative Representation of Change." In Spatial Information Theory - a Theoretical Basis for Gis (International Conference Cosit'97), edited by S. C. Hirtle and A. U. Frank, 15-33. Berlin-Heidelberg: Springer-Verlag, 1997.

Kuhn, W. Semantics of Geographic Information. Edited by A.U. Frank and P. Haunold. Vol. 7, Geoinfo Series. Vienna: Dept. of Geoinformation, 1996.

Lemmens, Rob. Semantic Interoperability of Distributed Geo-Services. Enschede, Netherlands: ITC, 2006.

Mark, D.M. "Toward a Theoretical Framework for Geographic Entity Types." In Spatial Information Theory: A Theoretical Basis for Gis (European Conference Cosit'93, Elba, Italy), edited by A.U. Frank and I. Campari, 270-83. Berlin: Springer-Verlag, 1993.

Mark, David M., Andrew G. Turk, and David Stea. "Progress on Yindjibarndi Ethnophysiography." Paper presented at the Spatial Information Theory COSIT 2007, Melbourn, Australia, Sept. 19-23 2007.

Mineau, G., G. Stumme, and R. Wille. "Conceptual Structures Represented by Conceptual Graphs and Formal Concept Analysis." In Conceptual Structures: Standards and Practices, 423-41. Heidelberg: Springer, 1999.

Niles, Ian. "Mapping Wordnet to the Sumo Ontology." 7. Palo Alto: Technowledge Corporation, 2003.

ONTOS. The Ontos, Inc. Web Page 2001 [cited 17 January 2001. Available from http://www.ontos.com.

Protégé. The Protégé Project Web Page 2000 [cited 21 November 2000]. Available from http://www.smi.stanford.edu/projects/protege.

Searle, John R., ed. The Construction of Social Reality. New York: The Free Press, 1995.