Nuno Silva and João Rocha
{Nuno.Silva, Joao.Rocha}@dei.isep.ipp.pt
DEI – ISEP - IPP
Porto – Portugal

Merging Ontologies using a Bottom-up Lexical and Structural Approach

Abstract: Information integration is the focus of different research domains for several years. With the emergence of the Internet as a very large database, this topic became of extreme relevance for a larger audience, seeking for mechanisms to align information from different data sources according to semantic needs and constraints. The classical integration approaches do not satisfy new operational requirements, thus new strategies should be proposed and developed. We suggest the adoption of a light alignment mechanism without merging data source. The alignment process occurs at ontological level, setting the components and transformation functions necessary to translate data from source to the target. Establishing the alignment process at ontological level allows the system to reason about the semantic embodied in the information and thus enhance the alignment results, according to both the source and target specifications. The resulting specifications are then applied in the transformation procedure, executed independently of data source entities.
Keywords: Semantic Web, Ontology, Interoperability, Alignment.

1. Introduction

Semantic Web vision propose the representation of data semantics and its exploitation in a machine processable way, contrasting with the current user-based approach, characterized by the discrepancy between the huge amount of existent data and the user capabilities to transform it in useful knowledge. Many standards (Semantic Web, 2002; W3C, 2002; UDDI, 2002) were proposed for the semantic web and knowledge interoperability, but entities have different information conceptualizations, consubstantiated in different information systems and respective ontologies (Guarino, 1994; Ushold and Grunninger, 1996) that can not be solve by these standards adoption. Accordingly, information integration research (Pinto, Gomez-Perez and Martins, 1999; Bergamaschi et al., 2001, Madhavan, Bernstein, and Rahm, 2001; Critchlow, Ganesh and Musick, 1998; Neumann et al., 2001; Visser and Tamma, 1999;Doan et al., 2002; Park, Gennari and Musen, 1997; Rahm and Bernstein, 2001), experienced new motivations and a very challenging application scenario. Additionally to the classical alignment problems, requirements in the Semantic Web are augmented due to decentralization (many entities describing eventually the same thing), dynamics (providers, consumers, content are always changing) and decontrol (no entity supervises the Web content). Typical alignment processes are very time consuming and the focus is mainly on accuracy and operational performance. We suggest that alignment process in semantic web should be fast and computational inexpensive, and must focus the relevant information components for each specific interoperability (Silva and Rocha, 2002). Additionally, we believe that in much situations the accuracy dimension of the alignment is not so fundamental and can be relaxed.

In next section we propose a conceptual alignment process describing the envisaged phases and propose different approaches for their operation. In section 3 we describe our understanding about the semantic bridge concept, both upon the process and the resulting specification. A description of a small case allows to present examples of our proposals. Section 4 summaries the current work and the main contributions, and suggest future steps.

2. Alignment Process

According to previous considerations we now introduce and describe the identified phases in the ontology alignment process.

Normalization phase respects the processes whereby both ontologies are synchronized to an unifying language and structural model. Additionally, minor lexical normalizations are executed. The prior concerns the constructs and may cause some changes in the represented knowledge, while the later deals with content (e.g. expansion of acronyms and abbreviations) and should avoid changes that might imply semantic modifications. In both cases it permits to maximize process abstraction (Omelayenko and Fensel, 2002).

Similarity measuring phase is the set of processes that specifies similarities between entities from both ontologies. Four different approaches exists in measuring ontologies similarities:

- Syntactic analysis concerns with measuring similarities between elements names. Simplest approaches considers names as set of characters (Levenshtein (Maedche and Staab, 2001) edit distance) but more complex ones attempt to infer syntactic relations between words. For example, considers the words "may" (month and simple present of verb "to may") and "might" ("power" and simple past of verb "to may"). The Levenshtein edit distance would result in a similarity of 4 (poorly similar), while a syntactical analysis would mention the relation centered in the verb sense. Although, it is difficult to translate humans readable information into computer quantifications;
- Lexical analysis measure similarity according to lexical relations and semantic meanings. Typical approaches use thesaurus, glossaries and other lexical tools like WordNet;
- Structural analysis consider the organization of (and relations between) elements in the ontology. Normally, an ontology has an underlying taxonomy, which is a good source of information for similarity measuring. Clustering techniques, based in syntactic and lexical similarities are typically used in this analysis [Bergamaschi et al., 2001];
- Extensional analysis evaluate similarity between instances of different ontology elements, which permits to evaluate a similarity of elements according to their instances.

Bridging phase intends to define which entities from the source ontology are semantically transposable with which entities in the target ontology, and how the instances associated to the first are transposed to instances in the second. Thus we distinguish two substantially different sub-phases, although closely related:

- **Association** sub-phase takes as input the similarity measures calculated in the similarity measuring phase, and defines associations between elements from both ontologies;
- **Rules specification** sub-phase concerns in specify, for each association, the process whereby the instances from the source ontology can be transformed into instances of the target ontology. In simplest case the necessary function would be just a copy of values, but in other cases complex rules should be applied in manipulating several entities from both ontologies.

Representation phase intends to represents semantic bridges using an external representation language. This representation is necessary when the execution phase is performed by other entity than that who specified the bridges. Hence, this phase is not mandatory if the bridging and execution phases are carried by the same entity. Nevertheless, in order to reuse those bridges it is recommend to represent them in a external, eventually widely accepted standard formalism. However, such language, besides a minimal proposal (Omelayenko, 2002), does not currently exists. A brief description of its requirements is presented in section 4.

Transformation phase intends to effectively translate instances from one data source to the other, according to the semantic bridges specified in the bridging phase. There are some relevant initiatives on the transformation filed in the context of semantic web, namely XSLT (W3C, 2002) and TRIPLE (Sintek and Decker, 2001). The prior is syntactic oriented, while

the later is an independent representation interpreter and rule processor. Its reasoning capabilities can be further expanded by connecting it to external inference and reasoning systems.

Transformation phase should be responsible to guarantee some quality in the resulting instances. This quality is hard to measure but some of its dimensions are: (i) unique instance identification (prevent identity duplication and same instance with different identity) and (ii) correct categorization of instances. Figure 1 presents a simple example of categorization refinement: at ontology level would not be possible to define semantic bridges between instance values, thus the value of Course attribute would not be filled if extensional analysis would not be performed. Later we describe a specific formalism that permits to expand the ontology.

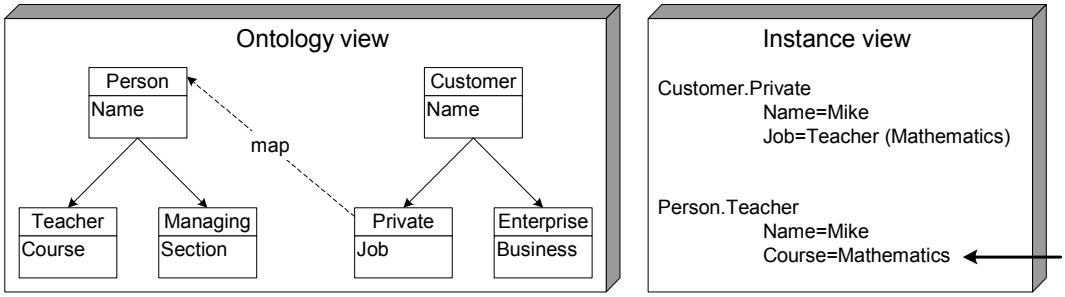

Figure 1 – Ontology view Vs. Instance view: data may complement semantic bridges

Negotiation phase concerns with the process whereby the "owners" of entities try to achieve a consensus about the semantic bridges. The negotiation process depends extensively in two components: the negotiation protocol and the arguing capabilities of the entities. Minimal research exists in this domain and none relating argumentation. The negotiation process may assume two distinct models according to the moment the negotiation occurs: (i) Progressively (for each derived bridge a negotiation procedure occurs in that moment) and (ii) At the end (each bridge is negotiated at the end of the bridging phase). A variation of these two approaches is possible, negotiating a group of bridges instead of individual bridges.

These two approaches are mutually contradictory in respect to advantages and disadvantages, but it is possible to combine both approaches, coping with each other disadvantages while maintaining the advantages.

3. Semantic Bridges

Semantic associations are categorized according to (i) Entity type, (ii) Cardinality and (iii) Scope. This section describes several issues in dealing with these dimensions.

Entity type dimension respects to the type of elements being associated in the bridge, which are of a single type. Typically there are 4 types of entities being described:

- Concept corresponds to compound objects of attributes, relations and axioms. It correspond to the OO class entity (RDF, 1999);
- Relation corresponds to a connection between two concepts, the subject and the object forming a triple (RDF, 1999). Each concept has its own identification, scope, etc.;
- Attribute corresponds to a connection between one concept (the subject) and an atomic value (the object) (RDF, 1999). The attribute do not has identity, and its scope is the concept it belongs to;
- Extensional Patterns intends to describe data source instances. This arises from the fact that ontologies describes entities from different perspectives and detail. Hence, sometimes it is necessary to expand the ontology characterization to meet the other ontology specification. The case from Figure 1 is a good example of this requirement.

Some other type of entities like statements, rules, functions, etc. are allowed in some ontology languages, which may appear in some ontologies. One may suggest these entity types should

also be bridged. However ontology alignment does not intend to provide alignment between ontologies but between data sources described according to ontologies. I.e., such elements are part of the ontology and not of the data source, thus the alignment process should not align these elements. On the other hand, these elements should be considered in the alignment process, once they describe information about the entities in the ontology.

Cardinality dimension relates to the number of objects intervening in the bridge from both ontologies, ranging from 1:1 to m:n. Besides it is theoretically possible, we did not found any real case where a m:n bridge was necessary that could not be substituted by 1:n and n:1 bridges whether entity type the bridge respects to.

Scope dimension relates to the bridges inter-relation and application, similar to the OO modeling perspective (composition, inheritance, abstraction, etc.). We suggest two modeling approaches that corresponds to three constructs:

- Composition, allows to specify a bridge as being composed by several others. For example, the concept bridges presented in Figure 2 are both composed of other bridges. Functionally it means that, for complete the outer bridge, the inner bridges should also be completed;
- Inheritance, allows to specify an hierarchy of bridges, from super-classes to sub-classes. Functionally it implies that the super-class bridges should be called by the sub-class and executed prior to itself. Abstraction is the third construct and represents a variation of the type of the super-classes. When this attribute is set, the specified bridge should be used only as super-class of another, and should not be executed independently. This situation is described in Figure 2 in the relation between the two concept bridges. Notice that the first bridge is stated as abstract, once it is considered that each instance of Staff should always belong to either Employee or Manager.

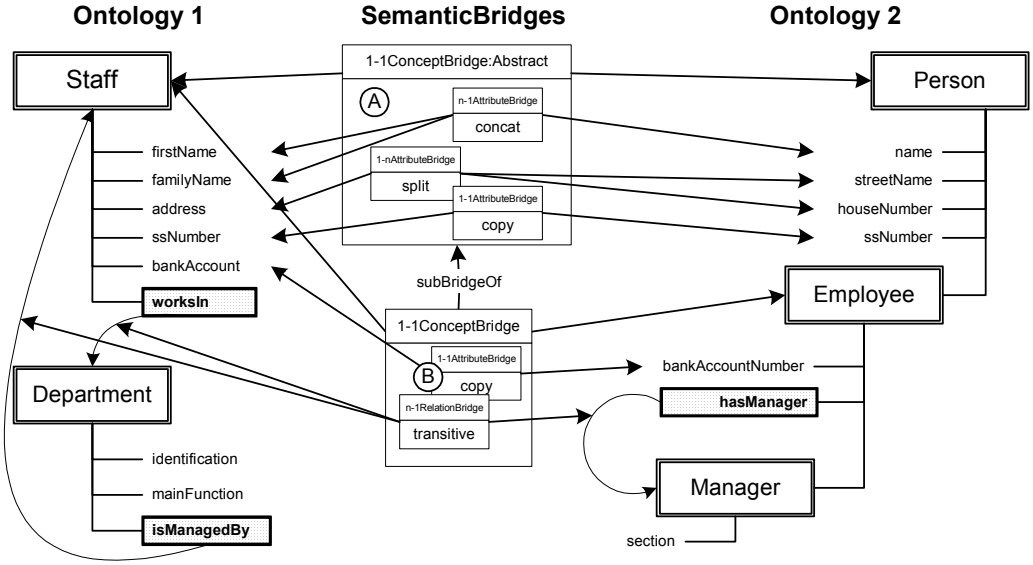

Figure 2 – Ontology bridging example

The n-1RelationBridge defined in the example from Figure 2 represents a bridge between two relations (shaded) in the source ontology and one relation in target ontology. It may be understood as:

```
var o1.Staff s={"Jim","Juice","Orange 23, 4234 orange grove",122,323,"#Food"};
var o2.Employee e;
e.name=concat(s.firstName,s.familyName);
...
var o1.Department d=s.worksIn;
var o1.Staff m=d.isManagedBy;
e.hasManager=m;
...
```

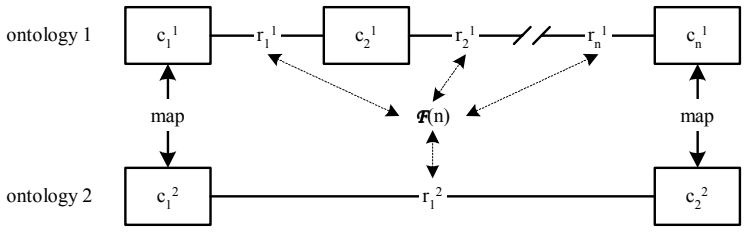

Figure 3 – Relation bridge (validity conditions)

Notice that a relation bridge makes sense only when three conditions are verified (Figure 3):

- The sequence of the relations are set and maintained;
- Exists a bridge between the concept holding the first relations from both ontologies;
- Exists a bridge between the concepts that play the object role in the last relation of the bridge.

Additionally to previous constructs, we believe that is necessary to specify a formalism to represent constraint rules associated with the bridge. This formalism allows to specify several bridges for the same entities, but prevent its execution according to such constraint rule. In the example from Figure 3, we could define a bridge C similar to bridge B, but associating Manager instead of Employee. A constraint rule would be associated, constraining it to be executed only if the Staff instance is referred as object in at least one instance of the Department.isManagedBy. In that case a Manager instance is created instead of Employee.

4. Summary and Future Work

We suggested a decomposition of a traditional ontology merging process, but we suggest the independence of specification and execution of semantic bridges. The proposed strategy permits to align heterogeneous data sources or heterogeneous ontologies in an inexpensive and effective way. Transformation relations are specified between source and target, dynamically and independently of contents, permitting an on-demand fast query mechanism. This strategy was firstly introduced and better justified in (Silva and Rocha, 2002). Accordingly, we suggests the existence of a representation formalism that would allow to exchange correctly and unambiguously the specified semantic bridges. We suggest to describe such formalism in a meta-ontology, which would allow to represent and exchange them as an instantiation. Hence, this meta-ontology would serve simultaneously as representation and validation formalism.

Currently we are focused in formalize the semantic bridges conceptualizations described in 3, and in specify this meta-ontology. The real semantic bridges case tests allow to test such language, and specially, test the usefulness and completion of the one composed by the described constructs for the scope dimension. This work is also progressing automating as much as possible the association phase.

Acknowledgment: This work is partially supported by the Portuguese MCT-FCT project POCIT/2001/GES/41830.

References

Bergamaschi et al., 2001; S. Bergamaschi, S. Castano, D. Beneventano e M. Vincini; "Semantic Integration of Heterogeneous Information Sources", Special Issue on Intelligent Information Integration, Data & Knowledge Engineering, Vol. 36, Num. 1, Pages 215-249, Elsevier Science B.V.; 2001.

Critchlow, Ganesh and Musick, 1998, T. Critchlow, M. Ganesh, R. Musick; “Automatic Generation of Warehouse Mediators Using an Ontology Engine”; in Proceedings of the 5th International Workshop on Knowledge Representation meets Databases (KRDB'98); May 1998.

Doan et al., 2002; AnHai Doan, Jyant Madhavan, Pedro Domingos, Alon Haley; “Learning to Map between Ontologies on the Semantic Web”, in Proceedings of the World-Wide Web Conference (WWW-2002); 2002.

Guarino, 1994; N. Guarino; “The Ontological Level”; Invited paper Presented at IV Wittgenstein Symposium; Kirchberg, Austria; in R. Casati, B. Smith and G. White (eds.), Philosophy and the Cognitive Sciences; Vienna, Austria; 1994.

Madhavan, Bernstein, and Rahm, 2001; Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm; “Generic schema matching with Cupid”; in Proceedings of the 27th International Conferences on Very Large Databases, pages 49-58; 2001.

Maedche and Staab, 2001; A. Maedche and S. Staab; “Measuring similarity between ontologies”; In Technical Report E0448; University of Karlsruhe; 2001.

Mitra et al., 2000; Prasenjit Mitra, Gio Wiederhold and Martin Kersten; "A Graph-Oriented Model for Articulation of Ontology Interdependencies"; in Proceedings of Conference on Extending Database Technology, (EDBT 2000); Konstanz, Germany; March 2000.

Neumann et al., 2001; H. Neumann, G. Schuster, H. Stuckenschmidt, U. Visser, T. Vögele and H. Wache; “Intelligent Brokering of Environmental Information with the BUSTER System”; in Proceedings of the International Symposium Informatics for Environmental Protection, pages 505-512; Zürich, Switzerland; October 2001.

Omelayenko, 2002; Omelayenko B.; “Integrating Vocabularies: Discovering, Representing and Compiling Vocabulary Maps“; in Proceedings of the First International Semantic Web Conference (ISWC-2002); Sardinia, Italy; June 9-12, 2002.

Omelayenko and Fensel, 2002; Omelayenko B. and Fensel D.; “Scalable Document Integration for B2B Electronic Commerce”; submitted; 2002.

Park, Gennari and Musen, 1997; J. Y. Park, J. H. Gennari, & M. A. Musen; “Mappings for Reuse in Knowledge-based Systems”; SMI Report Number: SMI-97-0697; 1997.

Pinto, Gomez-Perez and Martins, 1999; H.S. Pinto, A. Gomez-Perez, and J.P. Martins; “Some issues on ontology integration”; in Proceedings of the IJCAI'99 Workshop on Ontology and Problem-Solving Methods: Lesson learned and Future Trends; volume 18, pp. 7.1--7.11; Stockholm, Sweden; August 1999.

Rahm and Bernstein, 2001; E. Rahm and P. A. Bernstein; “A survey of approaches to automatic schema matching”; in The VLDB Journal 10: 334-350; 2001.

Resource Description Framework

RDF, 1999; World Wide Web Consortium; “Model and Syntax Specification: W3C Recommendation”; [http://www.w3.org/TR/REC-rdf-syntax]; accessed 22/03/2002; 1999.

Semantic Web, 2002; The Semantic Web Community Portal; [http://www.semanticweb.org]; accessed 22/03/2002.

Silva and Rocha, 2002; Nuno Silva and João Rocha; “Ontology mapping using multiple dimension approach”; in Proceedings of the International Conference on Fuzzy Systems and Soft Computational Intelligence in Management And Industrial Engineering (FSSCIMIE’2002); Istanbul, Turkey; May 2002.

Sintek and Decker, 2001; M. Sintek and S. Decker; “TRIPLE - An RDF Query, Inference, and Transformation Language”; Deductive Databases and Knowledge Management (DDLP'2001); Japan; October 2001.

UDDI, 2002; [http://www.uddi.org]; accessed 22/03/2002.

Ushold and Grunninger, 1996; Uschold, M. and Gruninger, M.; “Ontologies: Principles, Methods and Applications”; The Knowledge Engineering Review, 11(2): 93-136; 1996.

Visser and Tamma, 1999; Visser, P.R.S & Tamma, V.A.M.; “An Experience with Ontology-based Agent Clustering”; in Proceedings of the IJCAI 99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends, pp.12-1-12-13; Stockholm, Sweden; August 1999.

W3C, 2002; World Wide Web Consortium; [www.w3.org]; accessed 22/03/2002.