Building an integrated Ontology within the SEWASIE project: the Ontology Builder tool

D. Beneventano, S. Bergamaschi, A. Fergnani, D. Miselli, M. Vincini
Department of Informatics Engineering
University of Modena and Reggio Emilia
Via Vignolese 905 - 41100 Modena - Italy
Email: lastname@dbgroup.unimo.it

1 Introduction

SEWASIE (SEmantic Webs and AgentS in Integrated Economies) (IST-2001-34825) is a research project founded by EU on action line Semantic Web (May 2002/April 2005) (http://www.sewasie.org/). The goal of the SEWASIE project is to design and implement an advanced search engine enabling intelligent access to heterogeneous data sources on the web via semantic enrichment to provide the basis of structured secure web-based communication. A SEWASIE user has at his disposal a search client with an easy-to-use query interface able to extract the required information from the Internet and to show it in an easily enjoyable format. In this paper we focus on the Ontology Builder component of the SEWASIE system, that is a framework for information extraction and integration of heterogeneous structured and semi-structured information sources, built upon the MOMIS (Mediator envirOnment for Multiple Information Sources) [Bergamaschi *et al.*, 2001] system.

The Ontology Builder implements a semi-automatic methodology for data integration that follows the Global as View (GAV) approach [Lenzerini, 2002]. The result of the integration process is a global schema which provides a reconciled, integrated and virtual view of the underlying sources, GVV (Global Virtual View). The GVV is composed of a set of (global) classes that represent the information contained in the sources being used and the mappings establishing the connection among the elements of the global schema and those of the source schemata. A GVV, thus, may be thought of as a domain ontology [Guarino, 1998] for the integrated sources. Furthermore, our approach "builds" a domain ontology as the synthesis of the integration process, while the usual approach in the Semantic Web is based on "a priori" existence of an ontology (or a list of different versions of an ontology). The obtained conceptualization is a domain ontology composed of the following elements (see figure 1):

- local schemata of the sources: formal explicit descriptions with a common language, ODL_{I3} [Bergamaschi *et al.*, 2001], of concepts (classes), properties of each concept (attributes), and restrictions on instances of classes (integrity constraints).
- annotations of the local sources schemata: each element (class or attribute) is annotated with its meanings according to lexical ontology (we use WordNet [Miller, 1995]).
- a Common Thesaurus: is a set of intensional and extensional relationships, describing intra and inter-schema knowledge about elements of sources schemata. The kind of relationships are SYN (synonym of), BT (broader term / hypernymy), NT (narrower term / hyponymy) and RT (related term/relationship).
- a Global Virtual View (GVV): it consists of a set of global classes and the mappings between the GVV and the local schemata. In our approach, each Global Class represents a concept of the domain and each Global Attribute of a Global Class a specification of the concept. It is possible to define ISA relationships between Global Classes and to use a Global Class as domain of a Global Attribute.
- annotations of the GVV: the GVV elements (classes and attributes) meanings are semi-automatically generated from the annotated local sources.

With reference to the Semantic Web area, where generally the annotation process consists of providing a web page with semantic markups according to an ontology, in our approach we firstly markup the local metadata descriptions and then we produce the annotation of the GVV elements.

2 The Ontology Integration phases

1. **Ontology source extraction**
   The first step is the construction of a representation of the information sources, i.e. the conceptual schema of the sources, by means of the common data language ODLI3. To accomplish this task, the tool encapsulates each source with a wrapper that logically converts the underlying data structure into the ODLI3 information model. For conventional structured information sources (e.g. relational databases, object-oriented databases), schema description is always available and can be directly translated. In order to manage a semi-structured source we developed a wrapper for XML/DTDs files. By using that wrapper, DTD elements are translated into semi-structured objects in the same way as OEM objects [Papakonstantinou *et al.*, 1995].
2. **Annotation of the local sources**
   The designer has to manually choose the appropriate WordNet meaning for each element of local schemata.

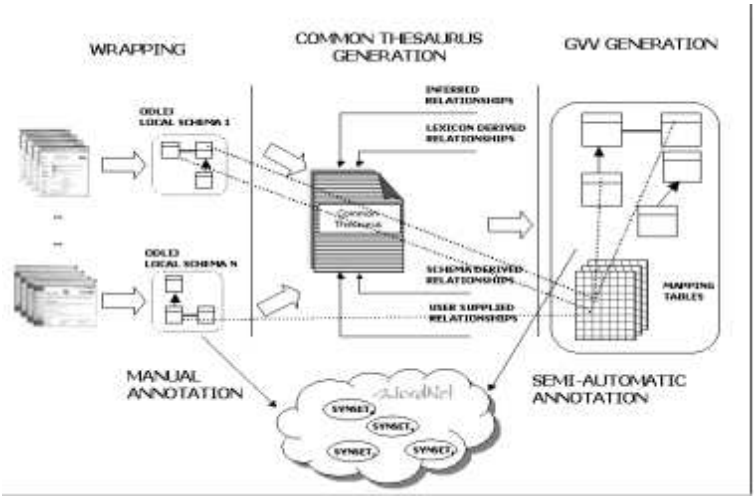

Figure 1: The Ontology Integration phases

First, the WordNet morphologic processor aids the designer by suggesting a word form corresponding to the given term, then the designer can choose to map an element on zero, one or more senses. If a source description element has no correspondent in WordNet, the designer may add a new meaning and proper relationships to the existing meanings.

3. **Common Thesaurus generation**
   The relationships of the Common Thesaurus are automatically extracted by analyzing local schemata description (for example in XML data files, ID and IDREF generate a BT/NT relationship and nested elements RT relationships), from the lexicon, on the basis of source annotation and of semantic relationships between meanings provided by WordNet, and inferred by using description logic inference techniques provided by ODB-Tools [Beneventano *et al.*, 1997].

4. **Affinity analysis of classes**
   Relationships in the Common Thesaurus are used to evaluate the level of affinity between classes intra and inter sources. The concept of affinity is introduced to formalize the kind of relationships that can occur between classes from the integration point of view. The affinity of two classes is established by means of affinity coefficients based on class names, class structures and relationships in Common Thesaurus.

5. **Clustering classes**
   Classes with affinity are grouped together in clusters using hierarchical clustering techniques. The goal is to identify the classes that have to be integrated since describing the same or semantically related information.

6. **Generation of the mediated schema (GVV)**
   For each cluster C, composed of a set S of local classes, a Global Class GC and mappings between global and local attributes are automatically defined. In particular, attributes of local classes in S related by SYN and BT/NT relationships in the Common Thesaurus are grouped and mapped into a single global attribute of GC.

7. **Annotation of the GVV**
   GVV elements (classes and attributes) meanings are semi-automatically generated from the annotated local sources. For a Global Class, the annotation is performed by considering the set of all its "broadest" local classes w.r.t. the relationships included in the Common Thesaurus. In particular the union of the meanings of the local class names in are proposed to the designer as meanings of the GVV and the designer may change this set, by removing some meanings or by adding other ones. For a Global Attribute, we use the same method starting from the set of local attributes which are mapped into it.

References

[Bergamaschi *et al.*, 2001] Sonia Bergamaschi, Silvana Castano, Domenico Beneventano and Maurizio Vincini. Semantic Integration of Heterogeneous Information Sources. DKE, Vol. 34, Num. 1, pages 215–249, Elsevier Science B.V., 2001.

[Lenzerini, 2002] Maurizio Lenzerini. Data Integration: A Theoretical Perspective. PODS, pages 233–246, 2002.

[Guarino, 1998] Nicola Guarino. Formal Ontology in Information Systems. In *N. Guarino (ed.)*, FOIS'98, 1998.

[Miller, 1995] A. G. Miller. A lexical database for English. Communications of the ACM, 38(11):39:41, 1995.

[Papakonstantinou *et al.*, 1995] Y. Papakonstantinou, H. Garcia-Molina, J. Widom. Object exchange across heterogeneous information sources. In *Proceedings of ICDE '95*, 1995.

[Beneventano *et al.*, 1997] Domenico Beneventano, Sonia Bergamaschi, Claudio Sartori and Maurizio Vincini. ODB–QOPTIMIZER: A tool for semantic query optimization in oodb. In *Proceedings of ICDE '97*, 1997.