

Ontology-based Integration for Sharing Knowledge over the Web ^{*}

D. Bianchini, V. De Antonellis

Università di Brescia
Dip. Elettronica per l'Automazione
Via Branze, 38
25123 Brescia - Italy
bianchin|deantone@ing.unibs.it

Abstract. In this paper, we propose a methodology developed in the framework of the VISPO project for engineering a three-layer ontology, based on the conceptualization, integration, synthesis and categorization of XML data descriptions provided by a number of sources in a virtual district, where different enterprises cooperate for business purposes. Ontologies are proposed as an unifying framework for different viewpoints by providing a shared understanding in a subject domain. Our methodology generates an ontology organized into concepts and concept relationships at different levels of detail, to provide multiple, unified views of the datasources containing heterogeneous information about the domain of interest.

1 Introduction

In the VISPO (Virtual district Internet-based Service Platform) project we addressed the problem of developing a service platform for a consortium of independent member enterprises, which operate in a cooperative way to exploit business opportunities, i.e., a virtual district. Enterprises cooperate by offering web services and sharing knowledge related to several distributed datasources and documents, possibly in different formats and generally XML-based or XML-compliant.

Consequently, for sharing XML data among different organizations or groups, a common frame of reference is required, where all concepts can be placed and understood. Ontologies are an extremely useful tool for expressing the meaning of distributed data and documents, allowing the creation of documents machine interpretable and not only machine readable [12]. This is the vision of the Semantic Web, that envisages the Web enriched with several domain ontologies, which specify formal semantics of data for different intelligent agents and services for information sharing, search, retrieval and transformation [3]. Main research issues in this field are concerned with the development of methods and tools for the construction of concept ontologies and the definition of thematic views to improve semantic interoperability and knowledge sharing [9]. Several efforts are also devoted to the development of techniques and approaches for the integration of heterogeneous datasources to provide global views on data provided

^{*} This work has been partially supported by the Italian VISPO (Virtual district Internet-based Service PlatfOrm [13]) projects.

from distinct organizations in a distributed environment, using ontologies to semantically organize the integrated knowledge about a particular domain [2, 4–6, 11].

In this paper, we consider the problem of sharing knowledge from a number of heterogeneous XML datasources over the Web and we rely on a three-layer ontology [7]. The ontology plays the role of interface between end-users and XML sources and has a twofold purpose: i) to provide a homogeneous, semantic view of the underlying XML data descriptions to support the formulation of queries at a semantic level, without to be aware of the structure and syntax of each specific description; ii) to define and maintain the mappings between the ontology schemas and the actual data in the underlying sources. In particular, we propose a methodology, developed in the framework of the VISPO project, for engineering a three-layer ontology, based on the abstraction, integration, synthesis and categorization of XML data descriptions provided by a number of sources in a virtual district.

This paper is organized as follows: Section 2 describes the ontology architecture and the proposed methodology. Sections 3 to 5 illustrate in more details the methodological phases. Concluding remarks are presented in Section 6.

2 A methodology for ontology construction

In the virtual district we consider a Web-based scenario, where XML is the standard adopted for information exchange among different datasources. We assume that, for each datasource, information to be exchanged is described by means of one of defined XML schema languages (e.g. DTD, DSD [10], XML Schema and RDF Schema). We propose an ontology architecture where information about XML schemas and their contents (e.g., meaning of elements, sub-elements, attributes) is organized in three layers:

- *semantic mapping layer*, where XML schema descriptions associated with different datasources are compared to find similarities between them; each description is abstracted into a set of XClasses according to a common conceptual formalism [5]; semantically related XClasses are clustered on the basis of their level of similarity, called *affinity*, computed considering terminological relationships (synonyms, hypernyms, etc) among their names and the structure of the XClasses featuring the involved datasources;
- *mediation layer*, where XClasses belonging to the same cluster are unified in global XClasses by means of integration techniques to obtain a unified view of exchanged knowledge; global XClasses are re-organized into ontological concepts and semantic relationships;
- *categorization layer*, where ontological concepts are related to subject categories, according to available standard taxonomies in the considered domain.

The proposed methodology for semi-automated extraction of ontological knowledge and for setting up the three ontology layers consists of four phases:

1. *data analysis and conceptualization*, to identify XClasses and cluster similar XClasses in different datasources;
2. *integration*, to obtain unified descriptions (global XClasses) of similar XClasses;

3. *synthesis and categorization*, (i) to define ontological concepts from global XClasses and to individuate semantic relationships between them in the mediation layer; (ii) to organize ontological concepts into subject categories in the categorization layer;
4. *implementation*, to formally represent the ontology in a Description Logic-based language and to iteratively refine and test the ontology concepts.

2.1 Running example

We present a running example to show how the phases of the methodology are applied. We consider the problem of integrating and sharing knowledge in the context of the industrial accessory and furnishing production, that we considered for the VISPO project experimentation. The information of interest is stored in two XML datasources, represented by an XML Schema (S1) and by a DTD document (S2). The graphical representation of the considered sources is shown in Figure 1. The source S1 contains information about the **Product Catalogue** that contains both zero or more **Product Categories** and one or more **Furnishing Components**; each **Product Category** can include one or more **Furnishing Components**. The source S2 contains the descriptions of **Furnishing Catalogue**, **Textile Catalogue** and **Metal Accessory Catalogue** in the **Industrial Accessory Production**; the three catalogue present respectively one or more **Furnishing Subcategories**, one or more **Textile Subcategories** and one or more **Metal Subcategories**, each of them describing one or more **Products**. In Figure 2, we show the ontology portion built on the two sources in the running example. In the following sections we will explain how to build it.

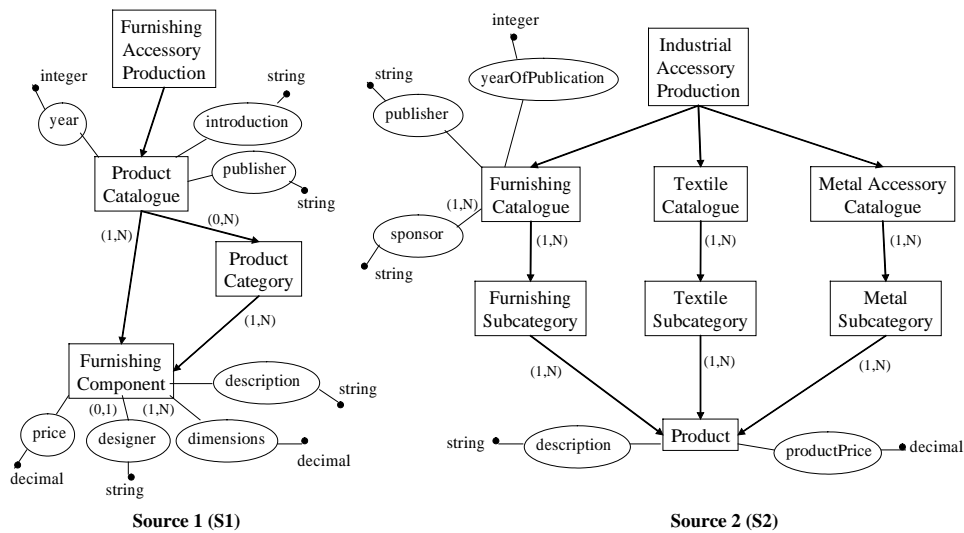


Fig. 1. Graphical representation of the two datasources considered in the running example.

3 Data analysis and conceptualization

Many languages have been proposed in order to describe schemas for XML documents. We extracted the most important features from the proposed schema languages and formalized them into a common conceptual formalism, called *X-Formalism*, presented in [5].

The X-Formalism tries to capture the main features in the various XML schema languages into a set of constructs, namely XClasses. Intuitively, an XClass represents a set of entities that have a common structure, described by a *name*, a *content model*, a set of *properties* (i.e., sub-elements with simple or built-in data types), a set of *attributes* and a set of references to other XClasses. Each schema language supports different content models, such as *empty*, *text*, *element* (if the content of an element includes other sub-elements), *mixed* (if the content includes either text or element content); in case of element content, the set of sub-elements can be ordered (*sequence*) or unordered (*all*); it is also possible to specify a choice among sub-elements, one and only one sub-element (*choice*) or any sub-element in any order (*any*). All schema languages control the occurrences of properties, attributes and references to other XClasses by specifying *cardinality constraints* (minimum and maximum occurrences).

In the data analysis and conceptualization phase, a first step consists in the identification of XClasses in the different XML datasource schemas. Once each available schema has been translated into a set of XClasses, these latter are compared to find semantic mappings between them (according to a schema matching process [11]). Our approach exploits the knowledge provided by XClasses and a thesaurus of weighted terminological relationships (e.g., synonymy, hyperonymy) to semi-automatically identify semantic mappings among XClasses of different sources. In ARTEMIS [1, 2] we have developed automated procedures for terminological relationship-based schema matching. A hierarchical clustering algorithm is used to find clusters based on the strength of the semantic mappings established between XClasses [6].

Applying the clustering algorithm to the set of XClasses that describe sources in Figure 1, we obtain the following clusters, as shown in the semantic mapping layer of the Figure 2: $cl_1 = \{\text{Furnishing Accessory Production, Industrial Accessory Production}\}$, $cl_2 = \{\text{Product Catalogue, Furnishing Catalogue, Textile Catalogue, Metal Accessory Catalogue}\}$ and $cl_3 = \{\text{Product Category, Furnishing Subcategory, Textile Subcategory, Metal Subcategory}\}$, derived from the similarity evaluation both of the names and of the structures (i.e., property and attribute names and data types) of the considered XClasses. `Furnishing Component` from S1 and `Product` from S2 are not clustered with other XClasses and constitute single-element (*singleton*) clusters.

4 Integration

The integration process is applied to obtain global unified descriptions (global XClasses), starting from clusters of XClasses. Basic reconciliation rules are introduced, which establish how to derive global features by unifying names, types and cardinality constraints of the XClasses in each cluster cl . Here, by *feature* we mean the property, referenced XClass or attribute of a given XClass.

Name reconciliation. The unified name of two feature f_1 and f_2 can coincide with the name of one of them or can be one of their hypernyms or synonyms.

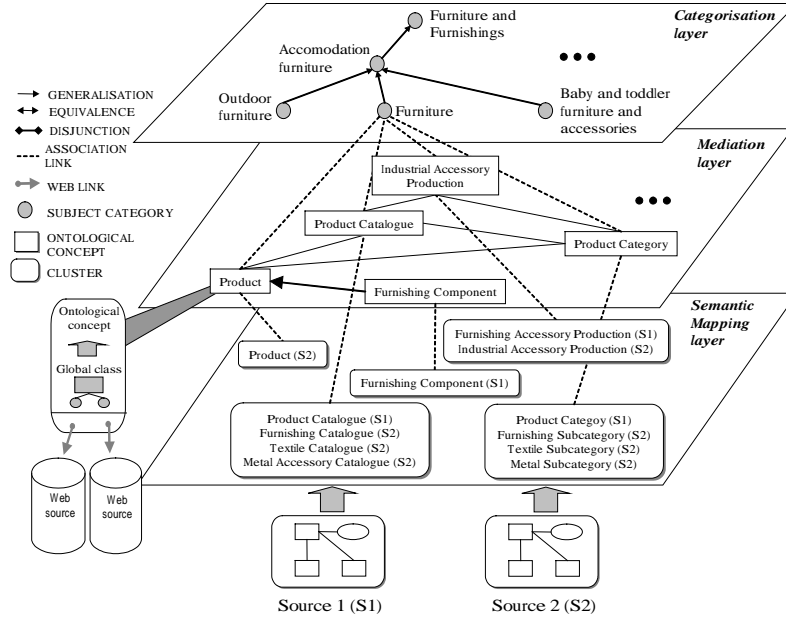


Fig. 2. A portion of the three-layer domain knowledge ontology.

Type reconciliation. The unified type of two features f_1 and f_2 coincides with the type of f_1 (or f_2) if they have the same type, otherwise, the selected type is the less restrictive among all the possible one (if they are comparable, since in general **integer** is not compared with **string**) or is chosen by the designer.

Cardinality reconciliation. The unified cardinality of two features f_1 and f_2 is defined as the less restrictive cardinality, i.e., the minimum (respectively, maximum) cardinality coincides with the minimum (respectively, maximum) value associated with f_1 and f_2 .

4.1 Mediation layer design process

Reconciliation rules are used for the mediation layer design process. We consider clusters containing more than one XClass: the corresponding global XClass is obtained by considering pairs of attributes and properties having semantic mappings in the cluster and by applying the name, type and cardinality reconciliation rules to pairs of them. In this phase, the set of referenced XClasses of a global XClass is not yet determined. In a refinement phase, we consider global XClasses obtained from the previous phase and we identify the referenced global XClasses by replacing each referenced XClass name occurrence with the name of the corresponding global XClass defined in the previous phase. To complete the global XClass definition, information for mapping global attributes, properties and referenced XClasses to corresponding features of XClasses in the cluster are specified in form of *mapping rules*, expressed as persistent mapping tables

whose columns represent the set of the local XClasses belonging to the cluster associated to a given global XClass and whose rows represent the global XClass features.

Product Catalogue Global XClass

Name: Product Catalogue
Content Model: (introduction,year,publisher,sponsor,Product Category,Furnishing Component)
Properties: {(introduction,string,(1,1)),(year,interger,(1,1)),(publisher,string,(1,1)),(sponsor,string,(1,N))}
RefX: {(Product Category,(0,N)),(Furnishing Component,(1,N))}
Attributes: { }

Mapping Table

Product Catalogue (Global XClass)	Product Catalogue (S1)	Furnishing Catalogue (S2)
introduction	introduction	
year	year	yearOfPublication
publisher	publisher	publisher
sponsor		sponsor
Product Category	Product Category	Furnishing Subcategory
Furnishing Component	Furnishing Component	

Fig. 3. Example of global XClass and the associated mapping table, derived from the integration of XClasses Product Catalogue and Furnishing Catalogue in Figure 1.

For example, if we apply the first phase of the unification process to the XClasses Product Catalogue and Furnishing Catalogue shown in Figure 1, we obtain the global XClass Product Catalogue shown in Figure 3 and in the mediation layer of Figure 2: the name of the class is derived by applying the name derivation rule; the properties year and publisher are obtained by applying the three reconciliation rules previously presented, while the properties sponsor and introduction are simply added to the global XClass. The same process is applied to the XClasses Product Category, Furnishing Subcategory, Textile Subcategory and Metal Subcategory to obtain the global XClass Product Category. In the second phase of the unification process the names of referenced XClasses Product Category and Furnishing Subcategory in the global XClass Product Catalogue are replaced with the corresponding global XClass name Product Category. Furnishing Component XClass belongs to a singleton cluster and then became directly a global XClass with the same name. The mapping table for the global XClass Product Catalogue is also shown in Figure 3.

5 Synthesis and categorization

Global XClasses provide uniform representation of heterogeneous datasources. To express better the semantics of unified viewpoint on the sources, global XClasses are organized into *ontological concepts* with semantic relationships among them. Ontological concepts are described by a name and a set of attributes (each of them with a name, a type and cardinality constraints); they are generated from global XClasses according to their different content models.

In the case of *sequence* and *all* content models, mapping between the global XClass and the ontological concept is one-to-one: features of the global XClass

(e.g., properties, attributes and so on) become attributes of the corresponding ontological concept, with associated types and cardinality constraints; in the case of *choice* and *any* content models the global XClass is mapped to more ontological concepts: for every possible combination of different alternatives a concept is generated; a further concept is created as generalization of all concepts previously generated from the considered global XClass.

After generation of ontological concepts, semantic relationships between them are derived according to their components and their structure. We consider three kinds of relationships.

Generalization. A concept α generalizes another concept β if the set of instances of α includes the set of instances of β ; this means that for every attribute x of α there exists an attribute y of β such that names of x and y are equal or synonyms, the type of y is equal or more restrictive of the type of x and the cardinality constraints of y are equal or more restrictive of cardinality constraints of x .

Disjunction. Two concepts α and β are disjoint if the sets of their instances are disjoint, i.e., there exist an attribute x of α and an attribute y of β such that names of x and y are equal or synonyms and the types of x and y are mutually exclusive or cardinality constraints of x and y represent disjoint ranges.

Equivalence. Two concepts α and β are equivalent if the sets of their instances coincide, i.e., α generalizes β and viceversa.

In our example, consider the XClasses `Furnishing Component` and `Product` in Figure 1, that become directly global XClasses with the same names, since they belong to singleton clusters; their content models are *sequences*, so the mapping to ontological concepts is one-to-one and every global XClass becomes an ontological concept. If we consider the semantic relationships between concepts, we note that all the attributes of `Product` are also attributes of `Furnishing Component` (`price` and `productPrice` are synonyms), with the same types and cardinality constraints; moreover, `Furnishing Component` has two further attributes, `designer` and `dimensions`. So we can conclude that `Furnishing Component` is a specialization of `Product` and this is represented in the mediation layer of Figure 2.

After having organized the mediation layer in ontological concepts and semantic relationships, the categorization phase is devoted to the identification of the subject categories as provided in available standard taxonomies. A subject category provides a topic-based view of underlying ontological concepts. In particular, an association link is maintained between an ontological concept in the mediation layer and a subject category in the categorization layer (see Figure 2). The association links are identified on the basis of the domain expert knowledge. In our work we considered the UNSPSC taxonomy [8]. Finally, concepts and semantic relationships are implemented in a suitable Description Logic-based language to support automated reasoning tasks [2].

6 Concluding remarks and future work

In this paper we have presented a methodology for the design of a three-layer ontology architecture for sharing knowledge in a virtual district. The ontology provides a unified view on sources which contain heterogeneous information about

a particular domain of interest. Information is extracted from datasources, it is integrated by use of reconciliation rules and is organized at three different layers. The proposed ontology provides a structured search space and allows the user to query multiple Web datasources without knowing in advance their location, vocabulary and contents and according to different search modalities. Future work includes strategies for ontology maintenance and designing of inference engines which can extract new information from ontological concepts and semantic relationships between them, revealing possible inconsistencies and ambiguities.

References

1. The ARTEMIS Project Home Page. http://www.ing.unibs.it/~deantone/interdata_tema3/Artemis/artemis.html.
2. D. Beneventano, S. Bergamaschi, S. Castano, V. De Antonellis, A. Ferrara, F. Guerra, G. Ornetti and M. Vicini. Semantic Integration and Query Optimization of Heterogeneous Data Sources. In *Invited Paper at 1st Int. Workshop on Efficient Web-based Information Systems (EWIS 2002)*, Montpellier, France, September 2nd 2002.
3. T. Berners-Lee, J. Hendler and O. Lassila. The Semantic Web. *Scientific American*, May 2001.
4. A. Cali, D. Calvanese, G. De Giacomo, M. Lenzerini, P. Naggar and F. Vernacotola. IBIS: Semantic Data Integration at Work. In *Proc. of the 15th Int. Conference on Advanced Information Systems Engineering (CAiSE 2003)*, Klagenfurt, Austria, June 16th-18th 2003.
5. S. Castano, V. De Antonellis, S. De Capitani and M. Melchiori. Semi-automated Extraction of Ontological Knowledge from XML Datasources. In *Proc. IEEE DEXA 2002 of Int. Workshop on Electronic Business Hubs (WEBH2002)*, Aix-en-Provence, France, pages 852–860, 2002.
6. S. Castano, V. De Antonellis and S. De Capitani di Vimercati. Global Viewing of Heterogeneous Data Sources. *IEEE Transactions on Knowledge and Data Engineering*, 13(2), 2001.
7. S. Castano, V. De Antonellis, S. De Capitani di Vimercati and M. Melchiori. Designing a Three-Layer Ontology in a Web-based Interconnection Scenario. In *Proc. IEEE of Int. Workshop WEBH2001*, Munich, Germany, 2001.
8. ECCMA. Universal Standard Products and Services Classification (UNSPSC). <http://www.eccma.org/>.
9. D. Fensel. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, Berlin, German, 2000.
10. N. Klarlund, A. Moller and M. I. Schwatzbach. DSD: a schema language for XML. In *Proc. of the 3rd ACM Workshop on Formal Methods in Software Practice*, Portland, Oregon, USA, August 2000.
11. J. Madhavan, P. A. Bernstein and E. Rahm. Generic schema matching with Cupid. In *Proc. of the Int. Conference on Very Large Data Bases (VLDB2001)*, pages 49–58, Rome, Italy, September 2001.
12. A. Maeche and S. Staab. Ontology Learning for the Semantic Web. *IEEE Intelligent Systems*, 13, 2001.
13. The VISPO Project Home Page. <http://cube-si.elet.polimi.it/vispo/index.htm>.