A WEB SERVICE BASED FRAMEWORK FOR THE SEMANTIC MAPPING AMONGST PRODUCT CLASSIFICATION SCHEMAS

Domenico Beneventano
Dipartimento di Ingegneria dell'Informazione
Università di Modena e Reggio Emilia
beneventano.domenico@unimore.it

Francesco Guerra
Dipartimento di Ingegneria dell'Informazione
Università di Modena e Reggio Emilia
guerra.francesco@unimore.it

Stefania Magnani
Dipartimento di Scienze e Metodi dell'Ingegneria
Università di Modena e Reggio Emilia
magnani.stefania@unimore.it

Maurizio Vincini
Dipartimento di Ingegneria dell'Informazione
Università di Modena e Reggio Emilia
vincini.maurizio@unimore.it

ABSTRACT

A marketplace is the place where the demands and offers of buyers and sellers participating in a business transaction may meet. Therefore, electronic marketplaces are virtual communities in which buyers may receive proposals from several suppliers and make the best choice. In the electronic commerce world, the comparison between different products is not possible due to the lack of common standards, used by the community, describing and classifying them. Therefore, B2B and B2C marketplaces have to reclassify products and goods according to different standardization models. In this paper, we propose a semi-automatic methodology, supported by a web service based framework, to define semantic mappings amongst different product classification schemas (e-commerce standards and catalogues) and we provide the ability to be able to search and navigate these mappings. The proposed methodology is shown over fragments of UNSPSC and ecl@ss standards and over a fragment of the eBay online catalogue.

Keywords: Semantic mappings, Product classification schemas, Annotations, Lexical ontologies.

1. Introduction

In the last few years, e-commerce has rapidly grown, enabling companies to be competitive on a large scale. E-marketplaces are one of the most promising activities, produced by e-commerce, and represent a virtual place where applications improving business transactions are proposed. Marketplaces enable buyers to analyze a wide range of products and, eventually, to obtain products and services quickly, reducing costs and times required by traditional trading activities. On the other hand, sellers may present a large amount of products, reduce selling costs and compete on large scale. In order to normally provide this service it is necessary to solve the problem of the heterogeneity of the services and the goods managed by the marketplace. B2B players use different document standards to represent their business documents, and different content standards to specify their products. Thousands of the users meet together at B2B marketplaces, and, consequently, the marketplace has to be able to manage numerous different documents and content standards. The large number of the standards and their significant complexity make the integration problems difficult and require development of special integration architectures [Omelayenko 2001].

The exponential increase of e-commerce amplifies the proliferation of different standards and joint initiatives for the classification of products and services. Some of these standards differ significantly in their coding systems, level of details, granularity and so on. Building one large and consensuated product and service standard for the e-commerce market as a whole is extremely hard and intricate. Therefore the solution is to reach the interoperability of coding systems. The harmonization or interoperation of various standards is turning out to be extremely important especially considering the Web and the Semantic Web [Gangemi 2002].

Interoperation of product classification standards is based on the semantic mappings between their elements. Manually finding such mappings is tedious, error-prone and clearly not possible at the scale of large product classification standards. Hence the development of tools to assist the identification, discovery, validation, and utilization process of semantic relationships is crucial in the context of the harmonization of product classification standards [Maedche 2002].

From an architectural point of view, the opportunity of connecting existing systems based on different product classification standards relies on the availability of software designed to automate the integration process as much as possible. This kind of software, usually referred as a mediator based system, acts as an interpreter between different IT systems. A major challenge is to provide software frameworks for the integration of business processes based largely on the semantics expressed by the processes themselves rather than on the structure of the information to be integrated [Hammer 2001]. Furthermore, the well known standardization issue, as well as the fast evolution of e-commerce technologies suggest focusing on component based architectures leveraging technologies such as XML Web Services and SOAP (Simple Object Access Protocol) [SOAP 2001].

Existing applications can be integrated more rapidly, easily and less expensively, since XML Web Services reduce the interoperability requirements to the minimum. In fact, Web Services have been conceived of as self-contained, modular applications that can be described, published, located and invoked over a network, generally, the World Wide Web [Kreger 2001].

In this paper, we will face the problem of merging different classification schemas, by proposing the use of a semi-automatic methodology, supported by a group of tools, to define semantic mappings amongst different product classification schemas (e-commerce standards and catalogues) and we well provide the ability to be able to search and navigate these mappings as a Web Service. The proposed methodology is developed in the context of the MOMIS system [Beneventano 2000, Bergamaschi 2001], a mediator system developed within the Intelligent Integration of Information research area. MOMIS (Mediator envirOnment for Multiple Information Sources) provides an integrated virtual view, called Global Virtual View, of heterogeneous structured and semi-structured information sources. A preliminary idea of the use of MOMIS within the product classification standard integration appears in [Bergamaschi 2002]. MOMIS is now evolving within the European project SEWASIE (SEmantic Webs and AgentS in Integrated Economies) (IST-2001-34825), and aims at providing access to heterogeneous web information sources.

The paper is organized as follows: section 2 introduces the most commonly used e-commerce standards and describes an example of an electronic catalogue. Issues related to the map generation amongst different product classification schemas are dealt with in section 3. Section 4 analyzes the annotation phase of the sources w.r.t.(with respect to) a lexical ontology, section 5 shows the proposed methodology by using a real example. Section 6 introduces a web service based framework to support the proposed methodology. Section 7 discusses related work and Section 8 concludes.

2. Product Classification Schemas

Coding products and services according to standardized classification systems is useful for speeding up commerce amongst companies. In addition to this, the development of e-commerce solutions, and in particular the B2B marketplace, has rapidly increased the requirement of machine-readable products names that assists marketing and sales functions to find customers and distribution channel services.

By inserting the codes in various electronic trade documents and media such as product catalogues, Web sites, purchase orders, invoices, inventory/sales advices, and other types of documents, computer applications throughout an extended supply chain (seller, buyer, distributor, independent sales representative, end user) can process transaction data automatically and can perform management, analysis and decision functions in time-critical and labour-efficient ways that would not be possible without the codes. A useful product classification schema should be hierarchical, so that individual commodities represent unique instances of larger classes and families. Hierarchical organization allows a given company to focus on a level of detail that best suits its purposes and situation. In addition to maintain a hierarchical taxonomy, a classification schema has to be constantly maintained (to add new products and modify existing structures to adapt to changing market offers), it has to be responsive to the industry (because delays damage business), and code assignments to products and services must be unbiased (To prevent unfairly promoting one's company's products at the expense of others)[Granada 2002].

In this section we present three proposals for the classification of products and services that have arisen in the context of e-commerce: UNSPSC, NAICS, ecl@ss schema. Each standard describes its contents by using a hierarchical four-level classification. Finally we present an electronic catalogue from an e-commerce platform. Products are classified within the catalogue, by following rules finalized in favor of layout issues than to providing a clear and uniform representation.

UNSPSC: Within the different standard classification systems proposed, the most used in the U.S. is the United Nation Standard Products and Services Code System (UNSPSC). UNSPSC is the result of a merger of the United Nations' Common Coding System (UNCCS) and Dun & Bradstreet's Standard Product and Services Codes (SPSC). The merger was completed in 1999 through the efforts of a team of analysts and researchers from both DB and the Inter-Agency Procurement Services Organization (IAPSO) of the UNDP. UNSPSC is considered an open standard, is available, free of charge, to anyone who wants to use it. Coding system is organized as five-level taxonomy of products. The levels allow user to search products more precisely (because searches will be confined to logical categories and eliminates irrelevant hits) and it allows managers to perform expenditure analysis on categories that are relevant to the company's situation. Each level contains a two-character numerical value and a textual description as follows:

XX Segment	The logical aggregation of families for analytical purposes
XX Family	A commonly recognized group of inter-related commodity categories
XX Class	A group of commodities sharing a common use or function
XX Commodity	A group of substitutable products or services
XX Business Function	The function performed by an organization in support of the commodity

Major obstacles in using the UNSPSC are that is rather shallow, not very intuitive, and not descriptive on an attribute level. A further disadvantage is that it is mainly developed in the US, leaving (for example) many European needs behind. In order to overcome some bottlenecks, there are initiatives to enhance the UNSPSC with local attribute. Example are the Eccma Global Attribute Schema (EGAS), managed by the ECCMA.

NAICS: Another standardization code is NAICS, it was created by the Census Office of USA in cooperation with the Economic National Classification Committee of USA, Statistics of Canada and the Instituto Nacional de Estadìstica, Geografia e Informatica de Mèjico. It describes products and services in general and is used in USA, Canada and Mexico. NAICS was created after revising the Standard Industrial Classification (SIC) standard. NAICS industries are identified by a 6-digit code, in contrast to the 4-digit SIC code. The international NAICS agreement fixes the first five digits of the code. The sixth digit, where used, identifies subdivisions. The hierarchical structure is the following:

XX	Industry Sector
XXX	Industry Subsector
XXXX	Industry Group
XXXXX	Industry
XXXXXX	U.S., Canadian, or Mexican National specific

Ecl@ss: An important European initiative that build a new classification scheme for scratch is ecl@ss, proposed by Cologne Institute for Business Research in cooperation with leading German industries. Ecl@ss is a standard for information exchange between suppliers and their customers and is characterized by a 4-level hierarchical classification system with a key-word register of 12,000 words. Ecl@ss maps market structure for industrial buyers and supports engineers at development, planning and maintenance. Through the access either via the hierarchy or over the key words both the experts as well as the occasional users can navigate in the classification. A unique feature of ecl@ss is the integration of attribute lists for the description of material and service specifications.

The eBay catalogue. We selected a catalogue of products from a popular e-commerce portal: eBay. This catalogue is organized into three kinds of elements, called categories, item and attributes. Catalogue items are actual products sold within the e-marketplace. Attributes are given to main characteristic of each product. Categories are groups of products (items) or groups of other categories. They are created with the aim of grouping products taking into account factors such as marketing, common use, etc. They have no attribute given to them. The chosen catalogue is composed of five hierarchic levels with 2/3 levels of depth in the hierarchy of category. Catalogues are designed instead as classifications of products and services from a market point of view. However, catalogues play an important role in the whole e-business process: they present a set of products offered by an e-commerce application and they are the interface used in the exchange of products in B2B and B2C environments.

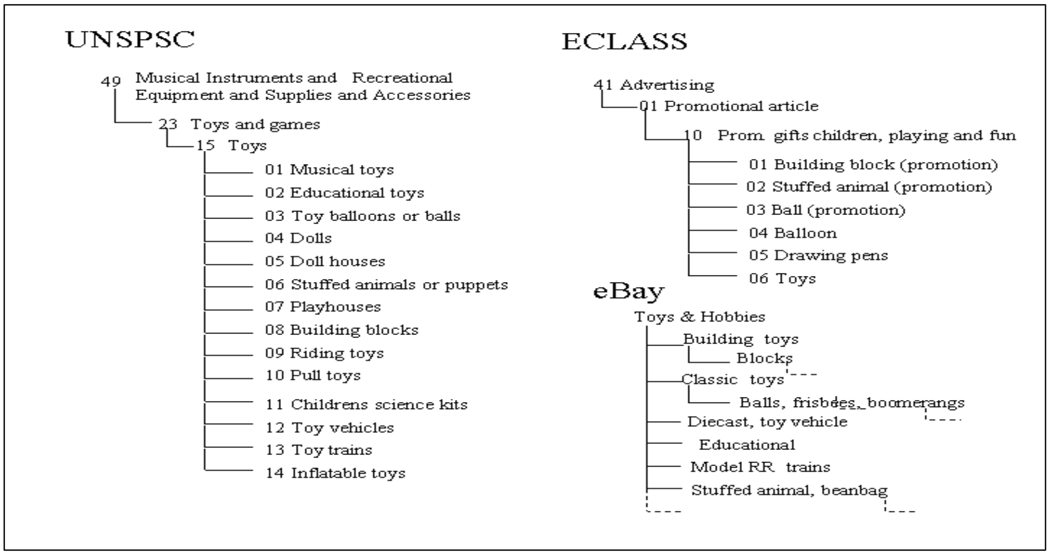


Figure 1:UNSPSC, ecl@ss and eBay fragment related to the "Toy" domain

2.1 Running example

The proposed methodology is shown over fragments of UNSPSC and ecl@ss standards and over a fragment of the eBay online catalogue related to the "Toy" domain (Figure 1), but is easy scalable to the whole standards and initiatives. The catalogue is structured in a very different way from the classification standards, it is similar to vertical classification systems, in fact there are about 500 different classes regarding the selected domain. So we will just analyze some significant classes from higher levels.

3. Semantic Mappings between Classification Schemas

In this section we face the problem of defining semantic mappings amongst different product classification schemas in the context of the MOMIS system [Beneventano 2000, Bergamaschi 2001]. In the first subsection below, we describe how a classification schema is represented. Then, in the next subsection, we define the semantic mapping that can be established between product classes of classification schemata.

3.1 Representation of classification schemas in MOMIS

The standards and initiatives introduced in the previous section, are described using different representation formats. NAICS and the eBay catalogue are available in HTML (taxonomy are presented visually); ecl@ss and UNSPSC are available in Microsoft Excel format. If we want to work with all this information together, we should use a common representation format, so that the treatment of this information can be performed homogeneously, no matter what its origin is [Corcho 2001]. To manage the information heterogeneity a mediator system typically encapsulates each sources by a wrapper, which logically converts the underlying data structures to the common data model. The MOMIS system uses as common data model an object-oriented language called ODL_I^3 [Bergamaschi 2001], an extension of the ODL language[1] which can be used to describe heterogeneous schemas of structured and semi-structured data sources. Due to the fact that the ODL_I^3 schema is composed of classes and simple binary relationships, the translation of ODL_I^3 descriptions into one of the Web standards such as RDF, DAML+OIL, OWL is a straightforward process; for example, the ISA ODL_I^3 relationship bring the same semantics of the concepts in the Semantic Web standards.

In our methodology semantic mappings are automatically generated according to the meanings of the product class names (*Lexicon-derived mappings)* and to the hierarchical organization of product classes (*Taxonomy-derived mappings)*. For this reason, representing a classification schema, we take in account only the product class names

[1] http:/www.service-architecture.com/database/articles/odmg_3_0.html

and the hierarchical structure of product classes. This choice is also supported from the fact that, in general, current standards do not include attributes for products; most of them just represent taxonomies of concepts, and other ones just include some attributes for them [Corcho 2001]; for example, ecl@ss contains a standard set of attributes only at the last level and UNSPSC is not descriptive on the attribute level. Concept hierarchies, i.e. ontologies without attributes and only with *is-a* relations between elements are becoming relevant in semantic mappings definition between ontologies as recently highlighted in [Giunchiglia 2003]. Consequently, the basic idea to obtain a representation of a classification schema in ODL_I^3 is straightforward: each level or product class of the classification schema corresponds to a ODL_I^3 class having as name the description of the level or product class and the hierarchical structure is represented by ISA relationships between classes. Moreover each product class of a classification standard has a code associated to the related ODL_I^3 class by means of a bijective function. We will use the notation *ClassificationSchema*.*ClassName* to denote a class with name *ClassName* in *ClassificationSchema*; for example, we will use UNSPSC.Toy_balloons_or_balls to denote the class "Toy balloons or balls" of the UNSPSC standard (code 49.23.15.03).

In this way, each product classification schema, considered as an information source, is represented as a set of ODL_I^3 classes, organized in ISA hierarchies; in the following a ODL_I^3 class will be also called *product class*.

3.2 Semantic Mappings

Once we have described the considered product standards and their representation, we will make an analysis of the relationships that can be established between different product classes.

In ODL_I^3 relationships between classes and attribute names are introduced in order to express intra- and inter-schema knowledge for information sources. In our context, we use these relationships to define *mappings* between product classes. We consider the following *mappings*:

- **SYN** (*synonym of*) is a relationship defined between two product classes that are *synonyms*/*equivalent* in their product classification schema. This mapping corresponds to the *equivalence mapping* introduced in [Corcho 2001].
- **NT** (*narrower classes*) this relationship occurs when a class in a classification standard is a *subclass* of another class or classes in another classification schema. The opposite of NT is BT (broader classes).
- **RT** *(related classes)* is a relationship defined between two product classes that are generally used together in the same context in the considered classification schema. RT relationships are symmetric.

More formally, let $S_1, S_2, \ldots, S_n$ be classification standards. A product class C of a classification standard S, $C \in S$, will be denoted by $S.C$. Given two classes C_i and C_j of different standards, i.e. $C_i \in S'$, $C_j \in S''$, $S' \neq S''$, a mapping M between C_i and C_j is defined as $C_i \; M \; C_j$, where

$$M \rightarrow SYN \mid BT \mid NT \mid RT$$

A set of mappings between classes of $S_1, S_2, \ldots, S_n$ will be denoted by $M(S_1, S_2, \ldots, S_n)$.

A mapping can be established both between classes of product classification standards and between classes of the electronic catalogue and classes of a classification standards. These kinds of mappings bring different semantics and are exploited in different ways.

A mapping between two classes of (different) classification standards allows the interaction between systems using different standards. It also provides several means for classifying the same product; as an example, a SYN mapping between the class UNSPSC.Stuffed Animal or Puppets and the class ECLASS.Stuffed Animal (codes 49.23.15.06 and 41.01.10.02 respectively) means that these concepts are equivalent.

A mapping can be established between classes of different levels; as an example, we can define a SYN mapping between the class UNSPSC.Toy (49.23.15.00) that is at the third level of the classification and the class ECLASS.Toys (41.01.10.06) that is at the lower level. A class of a classification standards can also be mapped to more classes of another classification standards; as an example, we can have

ECLASS.Ball	NT	UNSPSC.Toy_balloons_or_balls
ECLASS. Balloon	NT	UNSPSC.Toy_balloons_or_balls

In this way, we state that the concept UNSPSC.Toy balloons or balls (code 49.23.15.03) is *specialized* in two concepts of ECLASS.

A mapping between a class of a classification standard and a class of a catalogue enables the access to items or attributes of any product category of the catalogue through the taxonomy of concepts of the classification, this will facilitate searches of products from many different points of view [Corcho 2001]. These kind of mappings are useful to classify products, starting from their original catalogue descriptions, in accord with an existing classification schema that helps buyers and suppliers in communicating their product information.

4. Annotations of product classification schemas

In order to semi-automatically map different product classification standards and catalogues, there is a clear need to annotate or make the meaning of product classes explicit with respect to a common lexical ontology. The annotation, in fact, creates a *semantic bridge* between the classification schemas involved and a reference lexical ontology.

Ontological structures may give additional value to the semantic annotations. They allow for additional possibilities in the resulting annotations, such as conceptual navigation. But also the reference to a commonly agreed set of concepts by itself constitutes an additional value. Furthermore, an ontology directs the attention of the annotator to a predefined choice of semantic structures and, thus, gives a guidance about what and how item residing in documents may be annotated [Staab 2001].

In the following, we first introduce the adopted reference lexical ontology, which is Wordnet [Miller 1995], then we show the annotation w.r.t. WordNet and, finally, we discuss the problem of extending WordNet when a product class description does not match with lexical ontology concepts.

4.1 WordNet

The WordNet database contains 146,350 lemma organized in 111,223 synonym sets. WordNet's starting point for lexical semantics comes from a conventional association between the forms of the words that is, the way in which words are pronounced or written and the concept or meaning they express. These associations give rise to several properties, including synonymy, polysemy, and so forth. The correspondence between the word forms and their meaning is synthesized in the so-called Lexical Matrix M, in which the word meanings are reported in rows (hence each row represents a synset) and columns represent the word forms (form/base lemma):

Table 1: The WordNet word form and meanings

	F1	**F2**	**F3**	**...**	**Fn**
M1	$E_{1,1}$	$E_{1,2}$			
M2		$E_{2,2}$			…
M3	…		$E_{3,3}$		
...				…	
Mm		…			$E_{m,n}$

Each element in the matrix implies that the form in that particular column can be used in an appropriate context to express the meaning in that particular row. Thus, entry $E_{1,1}$ implies that word form F_1 can be used to express word meaning M_1. If there are at least two entries in the same column then the corresponding word form is polysemous (i.e. it can be used to represent more than one meaning, exactly two in this case); if there are at least two entries in the same row then two word forms are synonyms relative to a context.

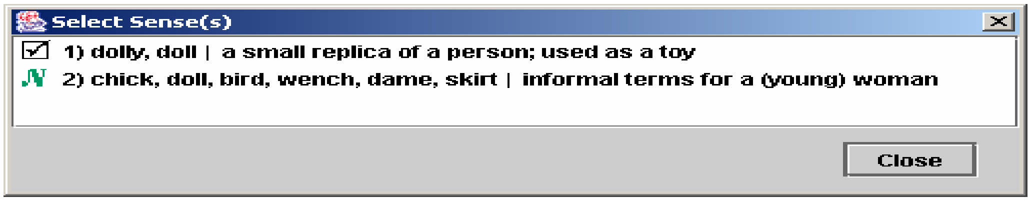
4.2 Annotation w.r.t. WordNet

The annotation w.r.t. WordNet consist of choose the correct (i.e. w.r.t. the context) WordNet meaning for each class. This is a two steps process that requires an interaction with the designer, i.e.the responsible person of the integration.

1. **Word form choice** In this step, the WordNet morphologic processor aids the designer by deriving the correct word form corresponding to the given term. More precisely, the morphologic processor stems (i.e. converts to a common root form) the term and checks if it exists as word form.
2. **Meaning choice** The designer can choose to map an element on zero, one or more senses.

As an example, in the annotation of the product class ECLASS.Dolls the WordNet morphologic processor derives the word form "Doll" and proposes two meanings (see Figure 2); the flag denotes the chosen meaning.

As an example of compound descriptive terms, in the annotation of the product class ECLASS.Doll_house the WordNet morphologic processor derives the word form "dollhouse" with two meanings and the designer chooses "a small model of a house used as a toy by children".



Select Sense(s)

☑ 1) dolly, doll | a small replica of a person; used as a toy

2) chick, doll, bird, wench, dame, skirt | informal terms for a (young) woman

Figure 2: "Dolls" annotation

If a class name is not available as word form, if there is an ambiguity, or the selected word form is not satisfactory, the designer can choose another word form of WordNet or manually search for a meaning of the class name. For a class name that does not find a meaning within WordNet the designer can choose:

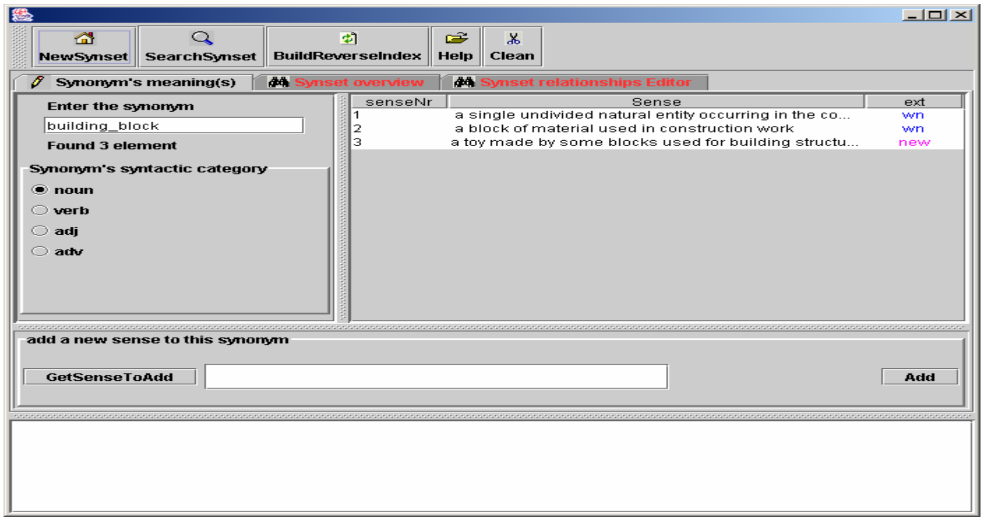
1. to consider the class name as *unknown*; in this case no lexicon-derived mappings will be derived for the class name (see section 5);
2. to extend the lexical ontology; we discuss this case in the next section.

4.3 Extending WordNet

Lexical semantic ontologies, such as WordNet, have proven very useful with many applications in Natural Language Applications. However, they usually only include general terms, as it would be impossible to extend them with every concept used in every domain of knowledge. In this context, we find very specific terms pertaining to different domains. If a source description element (i.e.a class name) does not find a correspondent within the reference lexical ontology (WordNet in our case), then the designer is requested to adapt the element to an already existing concept or to completely ignore it. However both these choices cause loss of information. We need to add new concepts and relations to the existing ontology.

We use WNEditor, a tool, developed in the MOMIS environment to make the designer able to efficiently create and manage new meanings and to create relationships between a new meaning and pre-existing ones. A new synset can be created both starting from an existing word form and from a new word form.

1. **creating a new synset starting from an existing word form:** the word form "building_block" is in WordNet with 2 meanings (meanings 1 and 2 in Figure 3) but there is not a *right* meaning related to the toy domain. In this case the designer can insert a new meaning for this word form (meaning 3, "A toy made of some blocks used for building structures", denoted with *new*, in Figure 3); moreover the designer can eventually add other word forms pertaining to this new synset, for example, "block" and "building_toys";



Enter the synonym

building_block

Found 3 element

Synonym's syntactic category

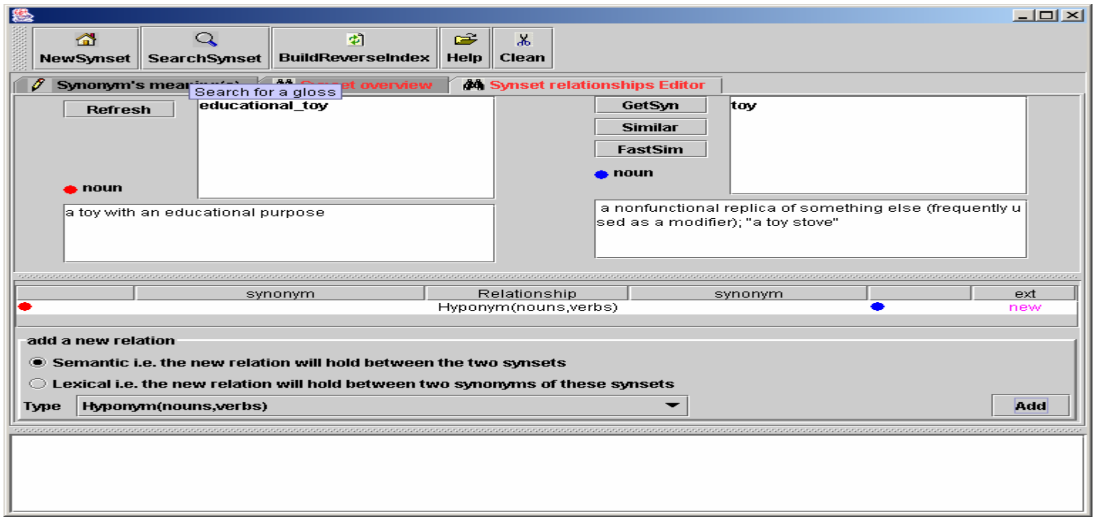
● noun
○ verb
○ adj
○ adv

senseNr	Sense	ext
1	a single undivided natural entity occurring in the co...	wn
2	a block of material used in construction work	wn
3	a toy made by some blocks used for building structu...	new

Figure 3: WNEditor: inserting new meanings

2. **creating a new synset starting from a new word form:** when the word form and the proper meaning are not in the lexical database the solution is to introduce the word form and of a new synset. As an example, we can insert the lemma "educational_toy" and the related meaning: "a toy with an educational purpose".

After inserting a new meaning, the designer can add some relationships existing between this new synset and those already existing in WordNet, by using a "Synset Relationships Editor"; in order to find candidate meanings for these relationships, WNEditor provides some search utilities based on information retrieval techniques [Baeza-Yates 1999]; for example the designer can search for meanings related to the keyword “toy”, to find the meaning and to define an Hyponym relationship (see Figure 4).



educational_toy

● noun

a toy with an educational purpose

toy

● noun

a nonfunctional replica of something else (frequently used as a modifier); "a toy stove"

	synonym	Relationship	synonym		ext
●		Hyponym(nouns,verbs)		●	new

add a new relation

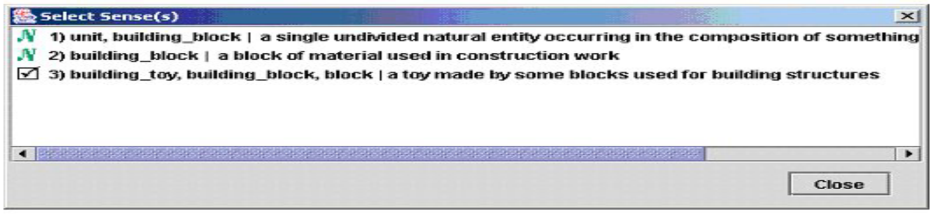
● Semantic i.e. the new relation will hold between the two synsets

○ Lexical i.e. the new relation will hold between two synonyms of these synsets

Type Hyponym(nouns,verbs) Add

Figure 4: WNEditor: inserting new relationships

All new inserted elements (synsets, word forms, relationships) are fully integrated in the WordNet and then can be used in the annotation process of all the sources. As an example, performing the annotation of the product class eBay.building_toys WNEditor proposes all the meanings shown in Figure 5, enabling the designer to choose the *new meaning* (3). As a consequence, a lexicon-derived SYN mapping between the product classes “Building_block” and “Building_toy” is generated (see Figure 6).



1) unit, building_block | a single undivided natural entity occurring in the composition of something

2) building_block | a block of material used in construction work

☑ 3) building_toy, building_block, block | a toy made by some blocks used for building structures

Figure 5: Meanings provided by Extended Wordnet

5. Building mappings

After annotating product classification schemata, we introduce our semi-automatic process to build a set of mappings between the classes.

Lexicon-derived mappings

The first phase is the extraction of mappings based upon the lexical relations existing between product class names. These mappings are derived from the meanings of the product class names chosen by the designer in the previous phase of annotation, by considering the semantic relations between meanings coming from WordNet, according to the following correspondences:

Synonymy:	corresponds to a	SYN	mapping
Hypernymy:	corresponds to a	BT	mapping
Hyponymy:	corresponds to a	NT	mapping
Holonomy:	corresponds to a	RT	mapping
Meronymy:	corresponds to a	RT	mapping
Correlation:	corresponds to a	RT	mapping

In Figures 6 and 7, some of the lexicon derived mappings existing between the sources are shown.

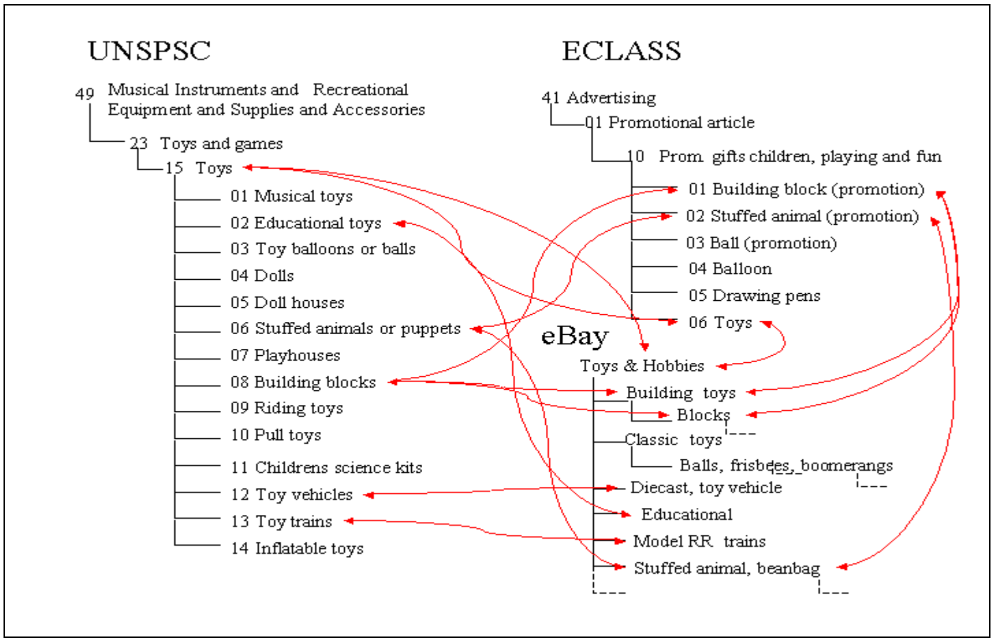


Figure 6: SYN lexicon-derived mappings

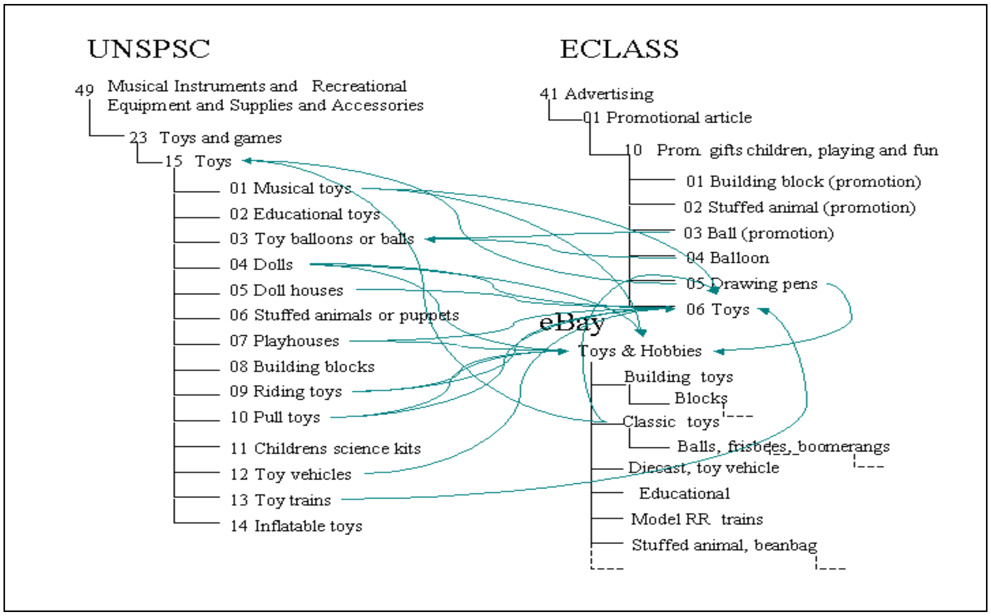


Figure 7: NT lexicon-derived mappings

Designer-supplied mappings

New mappings can be supplied directly by the designer, to capture specific domain knowledge. Moreover, the designer can modify/delete a mapping of the current set. As an example, the designer can insert the mapping:

UNSPSC.Toy_ball NT eBay.Balls_frisbees_boomerangs

Inferred mappings

By considering the meaning of mappings as explained above, we can define the following straightforward inference rules between mappings:

R $_1$: for each mappings the transitive property holds:

$$C_i \ M \ C_j \, , C_j \ M \ C_k \rightarrow C_i \ M \ C_k$$

$\mathbf{R}_2$: a SYN mapping is symmetric:

$$C_i \text{ SYN } C_j \rightarrow C_j \text{ SYN } C_i$$

$\mathbf{R}_3$: a RT mapping is symmetric:

$$C_i \text{ RT } C_j \rightarrow C_j \text{ RT } C_i$$

$\mathbf{R}_4$: a SYN mapping implies other mappings:

$$C_i \text{ SYN } C_j \rightarrow C_i \text{ NT } C_j \text{ , } C_j \text{ NT } C_i \text{ and } C_i \text{ RT } C_j$$

$\mathbf{R}_5$: a NT mapping implies a RT mapping:

$$C_i \text{ NT } C_j \rightarrow C_i \text{ RT } C_j$$

Given a set of mapping *M(S$_1$,S$_2$, ...,S$_n$)*, we define its *closure* ***M***$^+$*(S$_1$,S$_2$, ..., S$_n$)* as the set of mappings obtained by applying inference rules R$_1$ to R$_5$. As an example, from the mappings:

1) UNSPSC.Toy_balloons_ball NT eBay.Balls_frisbees_boomerangs
2) ECLASS.Ball NT UNSPSC.Toy_balloons_ball

the following mapping can be inferred:

ECLASS.Ball NT eBay.Balls_frisbees_boomerangs

Starting from mapping (1) which connects the classification standard UNSPSC and the catalogue eBay and the mapping (2) which connects two classification standards, we have built a mapping between ECLASS and eBay. The inferred mapping can be considered as a re-classification of the eBay catalogue with respect to the ECLASS coding system. Therefore if an electronic catalogue is mapped to a classification standard (i.e. UNSPSC), following the introducted inference rules it can be automatically mapped to all classification standards mapped with the first one (i.e. ecl@ss).

Taxonomy-derived mappings

These mappings are derived from the hierarchical organization of product classes: classes are analyzed and compared by means of an *Affinity Coefficient* which allows us to determine the level of similarity between them on the basis of the mappings existing between their subclasses. In other words, we define an Affinity Coefficient of two classes *C* and *C'*, denoted $SA \ (C, C', M)$, as the measure of the level of matching of *C* and *C'* based on mappings between their subclasses. If the *Affinity Coefficient* is greater than an *Affinity Threshold*, fixed by the designer, a mapping can be built between classes.

More formally, given a class *C*, with *S(C)* we denote the set of subclasses of *C*. We want to consider the different kinds of mappings existing between subclasses, then, we consider a mapping M and we define the set $MSUB(C_1, C_2, M)$ of subclasses of C_1 for which it exists a mapping M with a subclass of C_2 as follows:

$$MSUB(C_1, C_2, M) = \{C \in S(C_1) \mid \exists C' \in S(C_2) \wedge \ C \ M \ C' \in M^+\}$$

In order to evaluate if it is possible to establish a SYN mapping between the classes C_1 and C_2, we define a *Synonymy Affinity Coefficient* $SA(C_1, C_2, SYN)$ as follows:

$$SA \ (C_1, C_2, SYN) = \frac{|MSUB(C_1, C_2, SYN)| + |MSUB(C_2, C_1, SYN)|}{|S(C_1)| + |S(C_2)|}$$

The system proposes a mapping $C_1 \ SYN \ C_2$ if $SA(C_1, C_2, SYN)$ is greater than or equal to an *SYN-Affinity Threshold* fixed by the designer. Notice that, if $SA(C_1, C_2, SYN)$ = 1, all subclasses of C_1 are mapped by a *SYN* mapping with a subclass of C_2 and viceversa.

In order to evaluate if it is possible to establish a *NT* mapping between the classes C_1 and C_2, we define $SA \ (C_1, C_2, NT)$ as follows:

$$SA \ (C_1, C_2, NT) = \frac{|MSUB(C_1, C_2, NT)|}{|S(C_1)|}$$

The system proposes a mapping $C_1 \ NT \ C_2$ if $SA \ (C_1, C_2, NT)$ is greater than or equal to an *NT-Affinity Threshold* fixed by the designer.

In order to evaluate if it is possible to establish a RT mapping between the classes C_1 and C_2, we define $SA \ (C_1, C_2, RT)$ as follows:

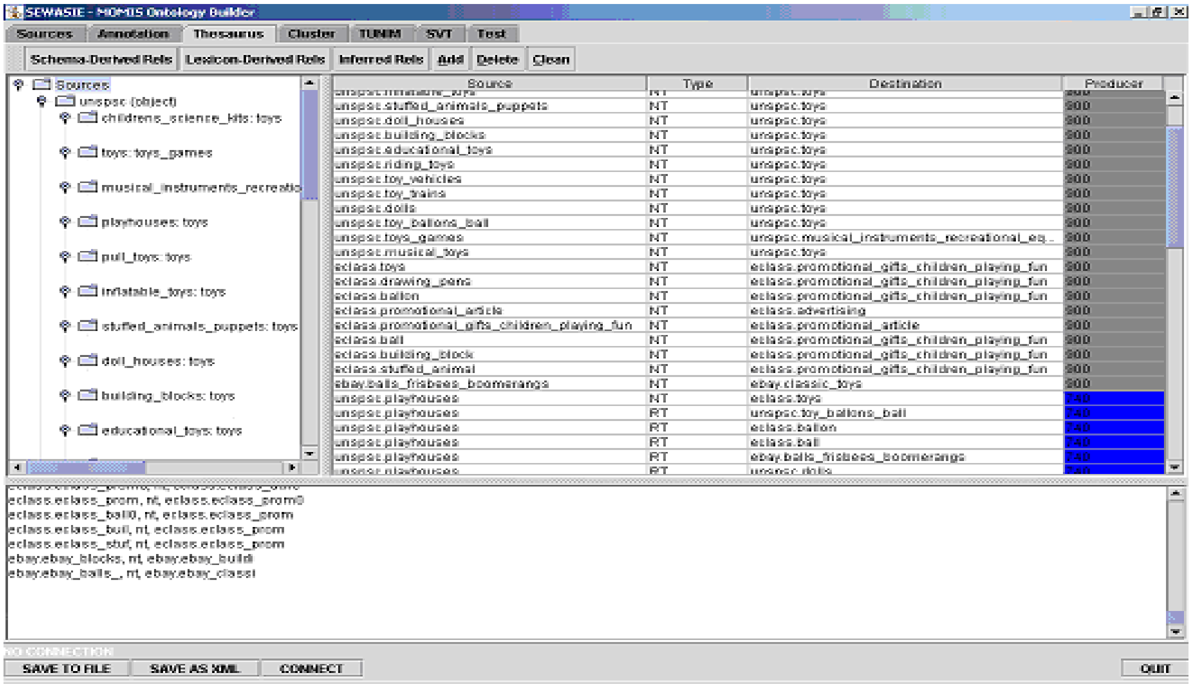
$$SA \ (C_1, C_2, RT) = \frac{|MSUB(C_1, C_2, RT)| + |MSUB(C_2, C_1, RT)|}{|S(C_1)| + |S(C_2)|}$$

The system proposes a mapping $C_1 \ RT \ C_2$ if $SA \ (C_1, C_2, RT)$ is greater than or equal to an *RT-Affinity Threshold* fixed by the designer.

As an example, for the product classes UNSPSC.Toys (C_1) and ecl@ss.prom_gifts_children_playing_fun (C_2), we have $SA\ (C_1, C_2, NT)$ = 0.57 and $SA\ (C_2, C_1, NT)$ = 0.33; then, considering a *NT-Affinity Threshold* equal to 0.5, the system proposes the following mapping:

UNSPSC.Toys NT ecl@ss.prom_gifts_children_playing_fun

In figure 8, the tool to generate and visualize mappings is shown.



Source	Type	Destination	Producer
unspsc.stuffed_animals_puppets	NT	unspsc.toys	900
unspsc.doll_houses	NT	unspsc.toys	900
unspsc.building_blocks	NT	unspsc.toys	900
unspsc.educational_toys	NT	unspsc.toys	900
unspsc.riding_toys	NT	unspsc.toys	900
unspsc.toy_vehicles	NT	unspsc.toys	900
unspsc.toy_trains	NT	unspsc.toys	900
unspsc.dolls	NT	unspsc.toys	900
unspsc.toy_ballons_ball	NT	unspsc.toys	900
unspsc.toys_games	NT	unspsc.musical_instruments_recreational_eq...	900
unspsc.musical_toys	NT	unspsc.toys	900
eclass.toys	NT	eclass.promotional_gifts_children_playing_fun	900
eclass.drawing_pens	NT	eclass.promotional_gifts_children_playing_fun	900
eclass.ballon	NT	eclass.promotional_gifts_children_playing_fun	900
eclass.promotional_article	NT	eclass.advertising	900
eclass.promotional_gifts_children_playing_fun	NT	eclass.promotional_article	900
eclass.ball	NT	eclass.promotional_gifts_children_playing_fun	900
eclass.building_block	NT	eclass.promotional_gifts_children_playing_fun	900
eclass.stuffed_animal	NT	eclass.promotional_gifts_children_playing_fun	900
ebay.balls_frisbees_boomerangs	NT	ebay.classic_toys	900
unspsc.playhouses	NT	eclass.toys	740
unspsc.playhouses	RT	unspsc.toy_ballons_ball	740
unspsc.playhouses	RT	eclass.ballon	740
unspsc.playhouses	RT	eclass.ball	740
unspsc.playhouses	RT	ebay.balls_frisbees_boomerangs	740
unspsc.playhouses	RT	unspsc.dolls	740

```
eclass.eclass_prom, nt, eclass.eclass_prom0
eclass.eclass_ball0, nt, eclass.eclass_prom
eclass.eclass_buil, nt, eclass.eclass_prom
eclass.eclass_stuf, nt, eclass.eclass_prom
ebay.ebay_blocks, nt, ebay.ebay_build
ebay.ebay_balls_, nt, ebay.ebay_classi
```

Figure 8: The interface to manage the mappings

6. Web Services infrastructure

A distributed computing model consists of a message exchange model, a communication protocol, and mechanisms for describing, defining, and discovering services. In particular it is composed of:

- Web Services that use Internet-based application level protocols (Hypertext Transfer Protocol -HTTP) for communication between applications;
- Simple Object Access Protocol (SOAP) [SOAP 2001] that is the message-exchange protocol;
- Web Services Description Language (WSDL) [WSDL 2002] that is the standard for describing and defining, services;
- Universal Description, Discovery and Integration (UDDI) that is the standard for discovering services.

SOAP is a lightweight protocol for the exchange of information in a decentralized, distributed environment. It is an XML based protocol which consists of three parts: an envelope that defines a framework for describing what is in a message and how to process it, a set of serializing rules for expressing instances of application-defined data types, and a convention for representing remote procedure calls and responses. All SOAP messages are encoded using XML. Simplicity and extensibility are major design goals for SOAP. This means that there are several features from traditional messaging systems and distributed object systems that are not part of the core SOAP specifications. SOAP defines a message-processing model but does not itself define any application semantics, such as programming model or implementation-specific semantics.

The SOAP specification defines also the relationships between HTTP messages and SOAP. This HTTP transport binding is important because HTTP is supported by almost all modern operating systems and makes it attractive for industrial uses. Since most organizations are familiar with HTTP and already have it incorporated into their network infrastructure, SOAP fits right in without the complex changes to the network or firewalls that many other protocols require. One of the most relevant uses of SOAP is to enable XML Web services. An XML Web service is a function that is exposed through a SOAP interface so that other SOAP-based application on the Web can call it to access the service.

WSDL (Web Services Description Language) [WSDL 2002] consists of two distinct parts – service definition and service implementation. Service definition is an XML-style description of what the service intends to provide, i.e., names of messages and their parameters, type of messages, etc. Service implementation specifies binding to a particular protocol or data type, i.e., syntax of the messages exchanged, protocols used to transfer messages, etc. By dissociating service definition from its implementation, WSDL allows re-use of the service description interface by clients that might be using other programming models to implement the service. Currently WSDL supports only SOAP and HTTP protocols for message communication.

In Figure 9 we show our framework, implementing three different Web Services by using SOAP and WSDL as description language: *Information Service, Metadata Service and Search Service.*

The *Information Service* provides names, kinds and the descriptions of the product standards or catalogues that are managed by the system. A generic application queries this service in order to obtain the integration process identifier, the list and a brief description of involved sources (product classification schemas).

The *Metadata Service* provides all the obtained semantic mappings existing amongst classification schemas; moreover, it provides the annotation of the involved classification schemas, i.e., the meaning of product classes with respect to the common lexical ontology Wordnet.

The *Search Service* provides the ability to be able to search and navigate semantic mappings amongst the different product classification schemas (mapped by the system). This service represents an entry point for a user application who wants to know how product classes of different classification standards are mapped or how a product of a catalogue is classified in the different standards. In particular by specifying the integration process identifier is possible to search for all kinds (filtering for a specific kind) of mappings related to a concept.

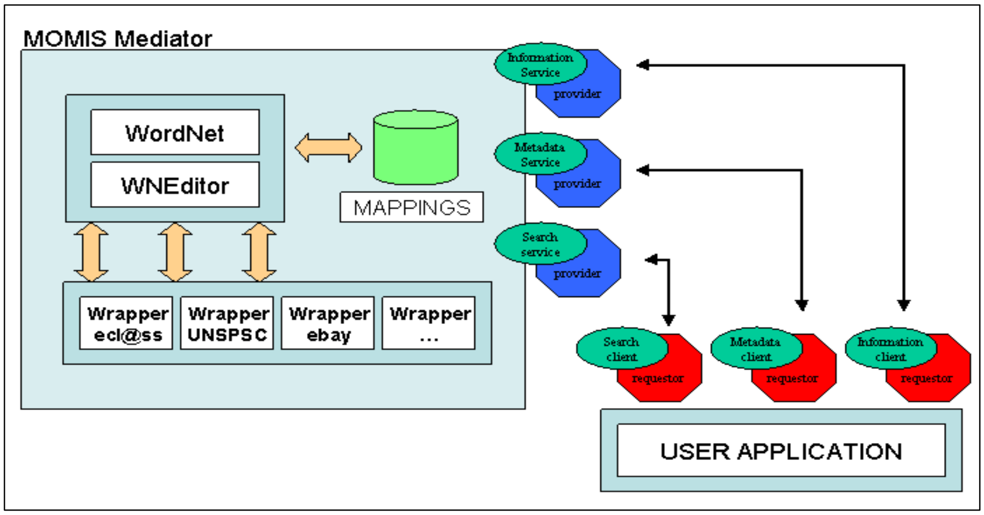


Figure 9: Access Web Services to product classification mappings

By means of these services all the knowledge provided by classification schema can be exported, therefore external portals, in particular marketplaces, can exploit results obtained by the framework in order to automatically provide contents to external users. By combining the provided Services a thirty parts portal may automatically obtain the mapping of its whole catalogue (or of a single element) respect to the other standards and catalogues integrated by the proposed methodology. For example, by using the 'toy domain' integration mapping described in the paper, a thirty parts portal may automatically query both the eBay web-site and the B2nB marketplace (www.b2nb.com.au/UNSPSCSearch.asp), that groups companies by using the UNSPSC Code.

7. Related work

Semantic mappings have been proposed in the area of heterogeneous databases [Sheth, 1990], in the area of ontology integration [Mena et al, 2000] and they are one of the main approaches for aligning ontologies on the World Wide Web [Stuckenschmidt, 2002]. Compared to these papers, the main differences of our work are that:

- we focus on the semantic mappings that are conceived of as important in the context of e-commerce standards and catalogues integration [Corcho 2001];
- we focus on the generation process of such mappings proposing a semi-automatic methodology where mappings are generated according to meanings of the product class names and to the hierarchical organization of product classes.

[Corcho 2001] claims that with the current state of affairs it is more suitable to establish ontological mappings between existing standards and initiatives than to pretend to build *the* unified knowledge model from scratch. They focus on the semi-automatic integration of existing standards and catalogues in a multilayered knowledge model for e-commerce applications through ontological mappings. They define different kinds of mappings in order to represent the different relationships existing between items of existing standards and catalogues but they do not discuss how such mappings are derived. In our paper, we consider similar mappings and we describe a semi-automatic methodology to build such mappings by using the semantics of classification schemas.

Few works discussed techniques to help automate the task of catalogues and standards integration. [Agrawal 2001] shows how a Naïve Bayes classification can be enhanced to incorporate the similarity information present in source catalogues. [Ding 2002] introduces GoldenBullet a software environment targeted to support product classification according to certain content standards. It is currently designed to automatically classify the products, based on their original descriptions and existent classification standards (such as UNSPSC). It integrates different classification algorithms from the information retrieval and machine learning areas and some natural language processing techniques to pre-process data and index UNSPSC so as to improve the classification accuracy.

These approaches are substantially based on syntactic affinity between involved terms; hence, they can be considered complementary to our approach, which is mainly based on the meaning of product classes with respect to a common lexical ontology.

8. Conclusions

E-business applications are adopting standards and initiatives that allow interoperation and the interchange of information between information systems.

In this paper, we proposed a semi-automatic methodology to define semantic mappings amongst different e-commerce product classification standards and e-marketplace catalogues. We exemplified the methodology by showing how it is possible to build mappings between a fragment of the UNSPSC standard, a fragment of ecl@ss standard and a fragment of an online catalogue. These mappings may be exploited to give the marketplace seller a unique code representing the same product that is classified by vendors in different manners.

The proposed methodology comprises of the following steps:

- *Acquiring and representing sources in a common format:* we face the problem of the format heterogeneity using specific wrappers to translate classification schemas and catalogues from their original format into the format required by our system.
- *Disambiguating content:* in order to semi-automatically map different product classification standards we annotate product classes with respect to a common lexical ontology. The annotation constitutes, therefore, a preliminary semantic bridge between the single classification standard and a lexical ontology like WordNet.
- *Extending WordNet:* if a source description element does not find a correspondent within the reference lexical ontology then the designer has to add a new concept and relations from the concept to the existing ones belonging to the lexical ontology. We propose WNEditor, to make the designer able to efficiently browse and to extend WordNet with his own new lexicons, meanings and relations.
- *Building mappings:* different kinds of mappings have been defined, in order to represent different kinds of relationships existing between items of the classification schemas. A semi-automatic methodology to build semantic mappings amongst different product classification schemas is proposed.
- *Providing Web Services:* by means of these services the knowledge provided by classification schema can be exported, therefore external portals, in particular marketplaces, can exploit the results obtained by the framework in order to automatically provide content to external users.

Future work will try to extend existing methodology in order to obtain the reclassification of catalogues w.r.t. different classification standards [Corcho 2001], [Ding 2002]. We will focus on the possibility to reclassify product catalogues according to different product classification standards, exploiting the existing mappings between the involved classifications. Another important future work is to extend our approach in order to consider multilingual classification schemas. The starting point for this activity will be the use of EuroWordNet, a multilingual database resembling WordNet that stores semantic relations between words in different languages of the European Community (www.illc.uva.nl/EuroWordNet).

REFERENCES

R. Agrawal and R. Srikant. On integrating catalogs. World Wide Web 2001, pages 603-612.

Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. Modern Information retrieval. ACM Press / Addison-Wesley, 1999.

Beneventano D., S. Bergamaschi, C. Sartori, and M. Vincini. Odb-tools: A description logics based tool for schema validation and semantic query optimization in object oriented databases. In Proc. of Int. Conf. on Data Engineering, ICDE'97, Birmingham, UK, April 1997.

Beneventano D., S.Bergamaschi, S.Castano, A.Corni, R. Guidetti, G. Malvezzi, M.Melchiori, and M.Vincini. Information integration: The MOMIS project demonstration. In The VLDB Journal, pages 611-614, 2000.

Bergamaschi S., S. Castano, D. Beneventano, and M. Vincini. Semantic integration of heterogenous information sources. Journal of Data and Knowledge Engineering, 36(3):215-249, 2001.

Bergamaschi S., F. Guerra, M. Vincini, A Data Integration Framework for E-commerce product classification, 1st International Semantic Web Conference (ISWC2002), Sardegna, Italy, 9-12 June 2002.

Corcho O. and A. Gomez-Perez. Solving integration problems of e-commerce standards and initiatives through ontological mappings, 2001. IJCAI-02 Workshop on E-Business and Intelligent Web,Seattle, August 5.

Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, B., Zykov,V., Klein, M., Schulten, E., Fensel, D.. GoldenBullet: Automated Classification of Product Data in E-commerce. Withold Abramowicz (ed.),Business Information Systems, Proceedings of BIS 2002,Poznan, Poland.

Gangemi A., N. Guarino, and M.Doerr. Harmonization perspectives of some promising content standards. WP3 Content Standardization-Deliverable 3.4.

Giunchiglia F., P. Shyaiko, Semantic Matching. *IJCAI-03 Workshop on Ontologies and Distributed Systems,* Acapulco, August 9, 2003.

Granada Research. Why coding and classifying products is critical to success in electronic commerce. White Paper, 2002.

Hammer K., “Almost Perfect: Where Middleware and XML May Fail to Deliver”, eAI Journal, June 2001, 12-16.

Hovy E.H.. Combining nd standardization large-scale, practical ontologies for machine translation and other uses. In Proceedings of the 1th International Conference on Language Resources Evaluation (LREC), Granada, May 28-30 1998.

Kreger, Web Services Conceptual Architecture (WSCA 1.0), May 2001 – IBM Software group- http://www-3.ibm.com/software/solutions/webservices/pdf/WSCA.pdf.

Maedche A., B. Motik, N. Silva, and R.Volz. MAFRA - A MApping FRAmework for distributed ontologies. Lecture Notes in Computer Science, 2473:235-??, 2002.

Mena E., A. Illarramendi, V. Kashyap, A. Sheth, OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Preexisting Ontologies. *International Journal Distributed and Parallel Databases (DAPD)*, 8(2), pp. 223-271, April 2000.

Miller A.G. Wordnet: A lexical database for english. Communications of the ACM, 38(11):39-41, 1995.

Omelayenko B. and D. Fensel. An analysis of b2b catalogue integration problems. In ICEIS (2), pages 945-952, 2001.

Schulten E., H.Akkermans, N. Guarino, G. Botquin, N. Lopes, M. Dorr, and N. Sadeh. The E-Commerce products classification challenge. Final version v1.0, july 2001. Intended for IEEE Intelligent System Magazine.

Sheth A., J. Larson, Federated Databases: Architectures and Issues, ACM Computing Surveys, 22 (3), September 1990, pp. 183-236.

Staab S. , A. Maedche, and S. Handschuh. An annotation framework for the semantic web, 2001. In Proceedings of the First Workshop on Multimedia Annotation, Tokyo, Japan, January 30-31.

W3C, Web Services Description Language (WSDL) 1.1, W3C Note 15 March 2001

Stuckenschmidt, H.; Timm, I.J.: Adapting Communication Vocabularies Using Shared Ontologies. In Cranefield, S. et al. (Eds.) "Proceedings of the Second International Workshop on Ontologies in Agent Systems", Workshop at 1st International Conference on Autonomous Agents and Multi-Agent Systems, 15-19 July 2002, Bologna, Italy, pp. 6-12

W3C, Simple Object Access Protocol (SOAP) 1.2, W3C Working Draft - 26 June 2002.