A framework for the classification and the reclassification of electronic catalogs

Domenico Beneventano
Dipartimento di Ingegneria dell'Informazione
Università di Modena e Reggio Emilia - Via Vignolese 905
Modena 41100 Italy
+39 059 205 6141
beneventano.domenico@unimore.it

Stefania Magnani
Dipartimento di Scienze e Metodi dell'Ingegneria
Università di Modena e Reggio Emilia -Viale Allegri 15
Reggio Emilia 42100 Italy
+39 0522 52 2232
magnani.stefania@unimore.it

ABSTRACT

Electronic marketplaces are virtual communities where buyers may meet proposals of several suppliers and make the best choice. The exponential increment of the e-commerce amplifies the proliferation of different standards and joint initiatives for the classification of products and services. Therefore, B2B and B2C marketplaces have to classify products and goods according to different product classification standards. In this paper, we propose a framework to classify and reclassify electronic catalogs based on a semi-automatic methodology to define semantic mappings among different product classification standards and catalogs.

Categories and Subject Descriptors

H.2.5 [**Database Management**]: Heterogeneous Databases - *Data translation.*

Keywords

Semantic mappings, Product classification standards, Electronics Catalogs, Annotations, Catalog Reclassification

1. INTRODUCTION

In the latest few years, e-commerce has rapidly grown up, enabling company to be competitive on a large scale. One of the most promising activities are e-marketplaces, that enable buyers to analyze a wide range of products and, eventually, to obtain quickly products and services, reducing costs and times required by traditional trading activities. On the other hand, vendors may present a large amount of products, reduce selling costs and compete in large scale.

The exponential increment of the e-commerce amplifies the proliferation of different standards and joint initiatives for the classification of products and services. Some of these standards differ significantly on their coding systems, level of detail, granularity and so on.

Marketplaces have to classify all products according to a standard classification schema that help buyers and suppliers in communicating their product information. Some widely used classification schemas are UNSPSC and ECLASS. It is a difficult and mainly manual task to classify products according to a classification schema like UNSPSC or reclassify products according to another schema, i.e. to classify w.r.t. ECLASS a catalog that has already been classified w.r.t. UNSPSC [7]. This paper shows a methodology to solve these specific problems. The first step is to reach the interoperability of coding systems: exploiting the semantic mappings between their elements [8]. Manually finding such mappings is tedious, error-prone and clearly not possible at the scale of large product classification standards. We propose a semi-automatic methodology to define semantic mappings among different product classification schemas. Then we can exploit these mappings to obtain the classification and reclassification of an electronic catalogue.

This methodology is developed in the context of the MOMIS system [2, 3], a mediator system developed within the Intelligent Integration of Information research area, and was exploited in a preliminary way for the product classification standards integration in [4]. MOMIS is now evolving within the European project SEWASIE (SEmantic Webs and AgentS in Integrated Economies) (IST-2001-34825).

The paper is organized as follows: section 2 introduces the most used e-commerce standards and describes an example of electronic catalog. A framework for the semantic mapping between classification schemas is described in section 3. Section 4 analyzes the annotation phase of the standards w.r.t a lexical ontology, section 5 analyzes the mappings generation and section 6 shows the proposed methodology to obtain a catalog classification and reclassification.

2. PRODUCT CLASSIFICATION STANDARDS AND CATALOGS

Coding products and services according to standardized classification systems is useful for speeding up commerce among companies. A useful product classification schema should be hierarchical, so that individual commodities represent unique instances of larger classes and families [9].

In this section we present two proposals for the classification of products and services that have arisen in the context of e-commerce (UNSPSC and ECLASS schemas) and an electronic catalog from a popular e-commerce platform (eBay).

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 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.

ECLASS: An important European initiative that build a new classification schema for scratch is ECLASS, proposed by Cologne Institute for Business Research in cooperation with

SAC'04, March 14-17, 2004, Nicosia, Cyprus.

leading German industries. ECLASS 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. 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 ECLASS is the integration of attribute lists for the description of material and service specifications.

The eBay catalog. This catalog is structured in three kinds of elements, called categories, items and attributes. Catalog items are actual products sold by the e-marketplace. Attributes are defined on them with the 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 defined on them. The selected catalog is composed by five hierarchic levels with 2/3 levels of depth in the hierarchy of category. Catalogs are designed instead as classifications of products and services from the market point of view. Marketplaces like eBay have to classify electronic catalogs according a standard classification, but given the proliferation of standadization initiatives it is often requested a reclassification according to the other standards.

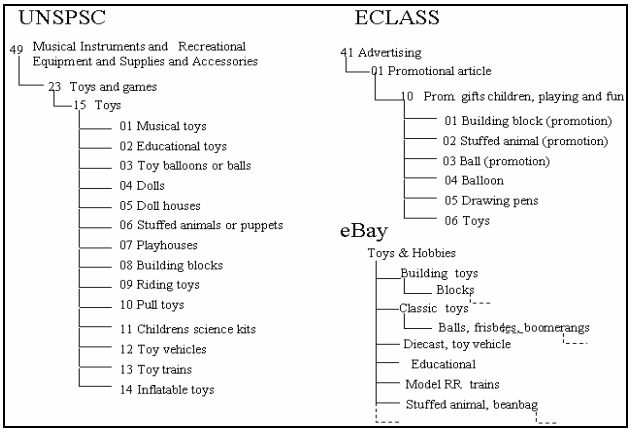

Figure 1. UNSPSC, ECLASS and eBay fragments

2.1 Running example

The proposed methodology is shown over fragments of UNSPSC and ECLASS standards and over a fragment of the eBay online catalog related to the "Toy" domain (Figure 1), but it is easily scalable to the whole standards and initiatives. In this example we assume that the electronic catalog has already been classified w.r.t. UNSPSC, i.e. every catalog category is associated to an UNSPSC code. This example will be used throughout the paper to explain and illustrate our main ideas; in particular:

- In section 3 we will show how the classification of eBay w.r.t. UNSPSC is represented in our framework;
- In section 5 we will show how to build a set of mappings between the UNSPSC and ECLASS standards;
- In section 6 we will show how to reclassify eBay w.r.t. ECLASS starting from the classification of eBay w.r.t. UNSPSC and by using the mappings between the UNSPSC and ECLASS standards.

3. A FRAMEWORK FOR SEMANTIC MAPPING

In this section we introduce a framework for semantic mapping between classification schema, developed in the context of the MOMIS system. To manage the information heterogeneity a mediator system typically encapsulates each source by a wrapper, which logically converts the underlying data structures to a common data model. The MOMIS system uses as common data model an object-oriented language called ODL_{I3} [3], an extension of the Object Definition Language, which is used to define interfaces to object types that conform to the Object Data Management Group (ODMG) object model. ODL_{I3} extends ODL with constructors, rules, and relationships that are useful for handling source heterogeneity.

3.1 Representation of classification schemas

The standards and the catalog introduced in the previous section, are described using different representation formats. The eBay catalog is available in HTML (taxonomy is presented visually); ECLASS and UNSPSC are available in Microsoft Excel format.

In the representation of a classification schema in ODL_{I3}, we take in account only the product classes and their hierarchical structure. This choice is motivated 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. For example, ECLASS contains a standard set of attributes only at the last level and UNSPSC is not descriptive on the attribute level. Consequently, the basic idea to obtain a representation of a classification schema in ODL_{I3} is straightforward: each level or product class of the classification schema corresponds to a ODL_{I3} class having as name the description of the level or product class and the hierarchical structure is represented by ISA relationships. Moreover each product class of a classification standard has a code associated to the related ODL_{I3} class. In this way, each product classification schema, considered as an information source, is represented as a set of ODL_{I3} classes, organized in ISA hierarchies; in the following a ODL_{I3} class will be also called product class.

3.2 Semantic Mappings

Once 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_{I3} relationships between classes are introduced in order to express intra- and inter-schema knowledge for the 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 the involved product classification schemas.
- **NT** (narrower classes) this relationship occurs when a class is a subclass of another class. 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 schemas. RT relationships are symmetric.

More formally, let $S_1,S_2,\ldots,Sn$ be product classification schemas. A product class C of a classification schema S, $C \in S$, will be denoted by S.C. Given two classes C_i and C_j of different schema,

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 → SYN | BT | NT | RT
The defined mappings enable to describe interactions between classification standards, between standards and catalogs and between catalogs:

• *Mappings between classification standards*. A mapping between two classes of (different) classification standards enables the interaction between systems using different standards. It also provides several means for classifying the same products; 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.

• *Mappings between catalogs*. A mapping between two classes of different catalogs enable the interaction between marketplaces that categorize products in different ways. Indeed, these kid of mappings establish a relationship between two category having different names but describing the same products. Building mappings between different catalogs is a first step toward the integration of electronic catalogs.

• *Mappings between a catalog and a classification standard*. These kind of mappings are used to represent the classification of a catalog w.r.t. a standard, i.e., a classified catalog will be represented in our framework by means of a set of mappings. Generally, a classification is composed by a set of corrispondences between the catalog and the classification standard, we represent these correspondences using SYN mappings. On the other hand, in [6] more complex correspondences have been proposed, we can represent them using different kind of mappings. As an example, every equivalence mapping will become a SYN mapping and every "Subclass-of" mapping will become a NT mapping. Figure 2 shows a set of SYN and NT mappings that represent the classification of the eBay fragment w.r.t. UNSPSC.

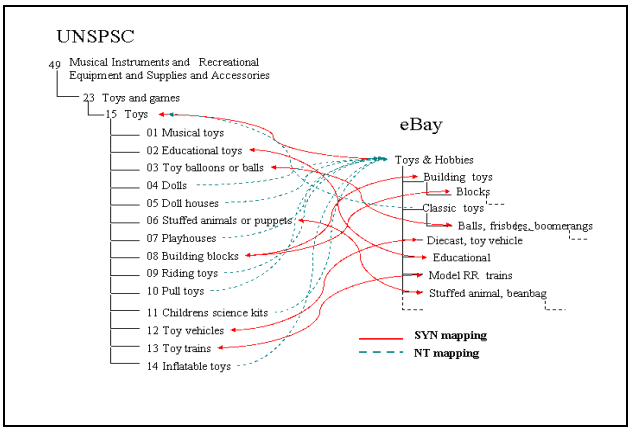

Figure 2. Classification mappings between UNSPSC and eBay

Manually finding semantic mappings is tedious and error-prone, hence, the development of techniques and tools to assist the designer in the identification, validation and utilization processes of semantic relations is crucial. Complete automation of these processes is unlikely to be possible: writing a correct mapping requires an understanding of the underlying semantics of the schema. In the following, we propose a methodology to extract semantic mappings among different product classification schemas and to propose them to the designer.

4. ANNOTATIONS OF PRODUCT CLASSIFICATION SCHEMAS

A remarkable set of mappings among classification schemas is derived from the meanings of the product class names; that is, the knowledge associated with product class names has to be exploited to built mappings. To this aim, we propose to annotate classification schemas in order to define the meaning of classes with respect to a common lexical ontology.

4.1 WordNet

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 concepts or meanings they express. These associations give rise to several properties, including synonymy, polysemy, and so forth.

4.2 Annotation w.r.t. WordNet

The annotation w.r.t. WordNet consists of choosing the correct (i.e. w.r.t. the context) WordNet meaning. This is a two steps process that requires an interaction with the designer.

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.
2. **Meaning choice** The designer can 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.

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.

4.3 Extending WordNet

Lexical semantic ontologies, such as WordNet, 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 belonging to different domains. If a source description element (i.e. a class name) does not find a correspondent within the reference lexical ontology, the designer is requested to adapt the element to an already existing concept or to ignore it. However both this choices cause loss of information. We need to add new concepts and relations to the existing ontology. We use a tool, WNEditor, developed in the MOMIS context, to make the designer able to efficiently create and manage new meanings and to create relationships between new meanings and pre-existing ones. A new synset can be created both starting from an existing word form and from a new word form.

- *creating a new synset starting from an existing word form*: the word form *"building_block"* is in WordNet with 2 meanings 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); moreover the designer can eventually add other word forms pertaining to this new synset, as, for example, *"block"* and *"building_toys"*.
- *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 the introduction of the word form and of a new synset. As an example of this case, we can insert the

lemma *"educational_toy"* and the related meaning: *"a toy with an educational purpose"*.

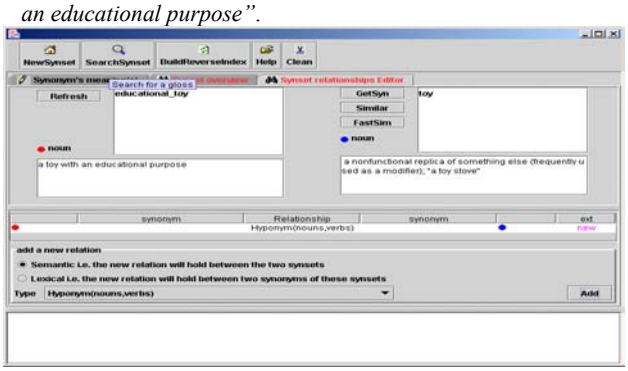

Figure 3. WNEditor: inserting new relationships

After inserting a new meaning, the designer can add some relationships holding between this new synset and those already existing in WordNet, by using a *"Synset Relationships Editor"* (Figure 3); in order to find candidate meanings for these relationships, WNEditor provides some search utilities based on information retrieval techniques [1]. For example the designer can search for meanings related to the keyword *"toy"*, to find the meaning and to define an Hyponym relationship. 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.

5. BUILDING MAPPINGS

In this section, we introduce our methodology for building a set of mappings between product classes of the classification schemas. This process is based on the following techniques of automatic derivation:

Lexicon-derived mappings. 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

Some lexicon derived mappings between two classification standards are shown in Figure 4.

Inferred mappings. We introduce the following straightforward inference rules between mappings:

R_1 : C_i M C_j , C_j M C_k → C_i M C_k
R_2 : C_i SYN C_j → C_j SYN C_i
R_3 : C_i RT C_j → C_j RT C_i
R_4 : C_i SYN C_j → C_i NT C_j ,C_j NT C_i and C_i RT C_j
R_5 : C_i NT C_j → C_i RT C_j

Given a set of mapping $M(S_1,S_2, \ldots,S_n)$, we define its closure $\mathbf{M}^+(S_1,S_2, \ldots, S_n)$ as the set of mappings obtained by applying inference rules R_1 to R_5.

Taxonomy-derived mappings. These mappings are derived from the hierarchical organization of product classes. 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.

As an example (for the complete definitions see [5]), the product classes *UNSPSC.Toys* (C_1) has 14 subclasses, 8 of those are NT of the subclasses of *ECLASS.prom_gifts_children_playing_fun* (C_2) then we have $SA(C_1,C_2,NT) = 8/14=0.57$, similarly, $SA(C_2,C_1,NT) = 0.33$; then, considering a *NT-Affinity Threshold* equal to 0.5, the system proposes the following mapping:

UNSPSC.Toys NT ECLASS.prom_gifts_children_playing_fun

New mappings can be supplied directly by the designer, in every step of the process, to capture specific domain knowledge; moreover, the designer can modify/delete a mapping of the current set.

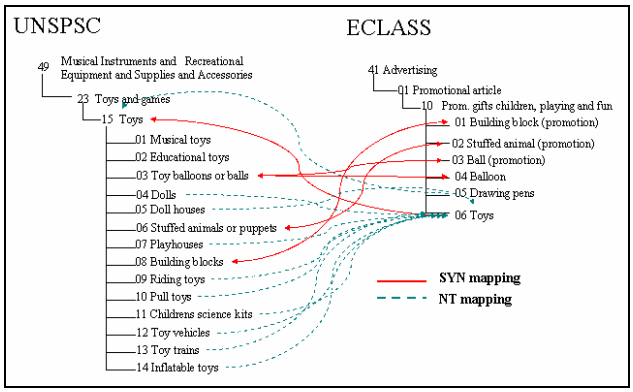

Figure 4. Lexicon-derived mappings between standards

6. CATALOG CLASSIFICATION AND RECLASSIFICATION

In this section, we show how to classify or reclassify electronic catalogs by exploiting semantic mappings among classification schemas. To classify a catalog S_1 w.r.t. a classification standard S_2, S_1 and S_2 have to be annotated w.r.t. a lexical ontology then the building mapping process has to be applied in order to detect a set of mappings $M(S_1,S_2)$ between the catalog and the standard.

The proposed methodology to reclassify a catalog exploits mappings between classification standards and inference rules introduced in section 5. More precisely, given a catalog S_1 and a classification standard S_2, let $M(S_1,S_2)$ be a set of mappings which represents the classification of S_1 w.r.t. S_2 (see section 3.2). Let S_3 be another classification standard and let $M(S_2,S_3)$ be a set of mapping between S_2 and S_3. In order to reclassify the catalog S_1 w.r.t. S_3, we consider: $M(S_1,S_2,S_3) = M(S_1,S_2) \cup M(S_2,S_3)$ and we compute its closure $\mathbf{M}^+(S_1,S_2, S_3)$. As an example, let us consider the classification of the catalog S_1=eBay w.r.t. S_2=UNSPSC (see Figure 2); this classification is represented by the set of mappings $M(S_1,S_2)$. Let $M(S_2,S_3)$ be the set of mappings between S2=UNSPSC and S3= ECLASS (see Figure 4). The obtained reclassification of eBay w.r.t. ECLASS is shown in Figure 5. Of course, if some product classes are not classified w.r.t. UNSPSC, no mappings will be built between these classes and ECLASS, so we can not obtain a complete reclassification. To classify these product classes we have to annotate them w.r.t. the lexical ontology, then we have to repeat a mapping generation process involving the classification standards and the catalogs (see section 5). The result of this process will be similar to that shown in Figure 6, where we can see the eBay catalog classified w.r.t. both classifications.

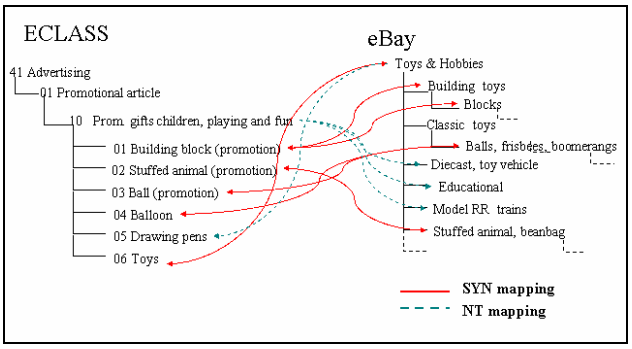

Figure 5. Reclassification mappings between ECLASS and eBay

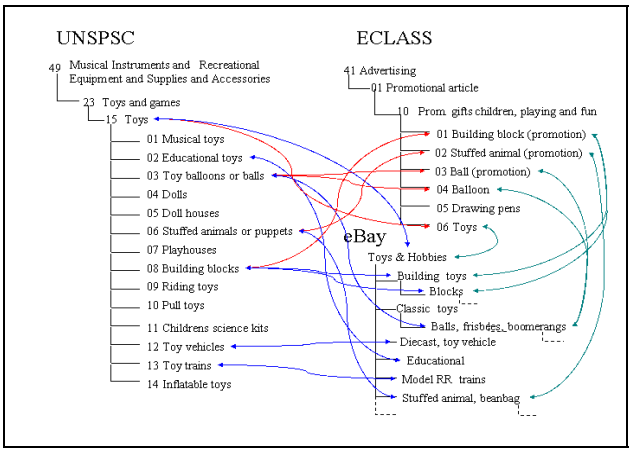

Figure 6. Catalog classification and reclassification w.r.t. ECLASS and UNSPSC (SYN mappings only)

7. CONCLUSIONS

E-business applications are adopting standards and initiatives for allowing interoperation and interchange of information between information systems. Electronic catalogs must be classified according to a standard classification schema that help buyers and suppliers in communicating their product information. Anyway there are too many standards and none of them is an actual standard, therefore it is important to classify and to reclassify catalogs according to different standards. In this paper, we proposed a semi-automatic methodology to classify\ reclassify catalogs exploiting semantic mappings among different e-commerce product classification standards. The proposed methodology is composed by 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 catalogs 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.
- *Extending WordNet*: if a source element has not a correspondent meaning within the reference lexical ontology then the designer has to add a new concept and some relations. We propose a tool, WNEditor, to make the designer able to efficiently browse and to extend WordNet with his own new lexicons, meanings and relations among them.
- *Building mappings*: different kinds of mappings have been defined, in order to represent different kinds of relationships holding between items of the classification standards and catalogs. A semi-automatic methodology to build semantic mappings among different product class is proposed.
- *Classification and reclassification an electronic catalog w.r.t. a classification standard*: we defined a methodology to classify or reclassify (starting from its preliminar classification) an electronic catalog, exploiting the semantic mappings between standards and classifications and applying simple inference rules.

8. REFERENCES

[1] Baeza-Yates, R. A., And Ribeiro-Neto, B.A. 1999. Modern Information retrieval. ACM Press / Addison-Wesley.

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

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

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

[5] Beneventano, D., Magnani, S. A framework for the classification and the reclassification of electronic catalogs. DB-Group Technical report: www.dbgroup.unimo.it\paper\BenMag2003.pdf

[6] Corcho, O.; Gòmez-Pérez, A. Solving Integration Problems of e-commerce Standards and Initiatives through Ontological Mappings IJCAI'01 Workshop on Ontologies and Information Sharing. Seattle (USA). August 2001.

[7] 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.

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

[9] Granada Research 2002. Why coding and classifying products is critical to success in electronic commerce.

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

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

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