

# A Data Integration Framework for e-Commerce Product Classification

S. Bergamaschi <sup>1</sup>, F. Guerra, and M. Vincini

Dipartimento di Ingegneria dell'Informazione  
Università di Modena e Reggio Emilia  
Via Vignolese 905 – Modena

{bergamaschi.sonia, guerra.francesco, vincini.maurizio}@unimo.it

**Abstract.** A marketplace is the place in which the demand and supply of buyers and vendors participating in a business process may meet. Therefore, electronic marketplaces are virtual communities in which buyers may meet proposals of several suppliers and make the best choice. In the electronic commerce world, the comparison between different products is blocked due to the lack of standards (on the contrary, the proliferation of standards) describing and classifying them. Therefore, the need for B2B and B2C marketplaces is to reclassify products and goods according to different standardization models. This paper aims to face this problem by suggesting the use of a semi-automatic methodology, supported by a tool (SI-Designer), to define the mapping among different e-commerce product classification standards. This methodology was developed for the MOMIS system within the Intelligent Integration of Information research area. We describe our extension to the methodology that makes it applicable in general to product classification standard, by selecting a fragment of ECCMA/UNSPSC and ecl@ss standard.

## 1 Introduction

The large amount of Internet sites, which have grown in the few last years, has increased the availability of information on the web, even if, due mainly to its structure, this information is less and less machine-readable and machine-understandable. Nevertheless, in the context of electronic commerce, many companies, organizations, and customers are exploiting the opportunities offered by Internet-based solutions and many more are expected to follow. Companies have been putting their databases and product catalogues on the web. Consequently, customers and suppliers have increased the amount of available information, but also the “noise” generated from these information sources has increased. This situation has allowed a third party, called the *marketplace*, to assume a key role in electronic commerce.

A marketplace is the place in which the demand and supply of buyers and vendors participating in a business process may meet. Therefore, marketplaces are virtual communities in which buyers may meet proposals of several suppliers and make the best choice. Then, marketplaces seem to be an interesting solution for e-commerce

---

<sup>1</sup> CSITE-CNR viale Risorgimento 2, 40136 Bologna, Italy

actors, because they show products, distributed by different vendors, but that may be compared since they have similar classification and they represent comparable products. In the e-commerce world, the comparison between different products is blocked due to the lack of standards describing and classifying them. Numerous proposals of classification standards have resulted in each supplier describing his own product in his own way (cf. [11], [18]).

Considering B2B e-commerce marketplaces, an economic transaction is further given difficult by the buyers' need to use only a standard to classify and describe products provided by different vendors, so as to assure an easy integration with its ERP system. Hence, the marketplace has to provide an environment using which to mediate among different standards used by the different participant to the transaction. In this way, each actor of the business process may exchange information using his own format. Therefore, the need, for B2B and B2C marketplaces, is to reclassify products and goods according to different standardization models.

This paper aims to face this problem by suggesting the use of a semi-automatic methodology to define the mapping among different e-commerce product classification standards. The methodology was developed for the MOMIS system ([4], [5], [6]) within the Intelligent Integration of Information research area. MOMIS (Mediator enviroNment for Multiple Information Systems) is a mediator-based system aiming to extract and integrate information from heterogeneous data sources, such as relational, object, semistructured sources (XML). Starting from source descriptions, the system generates an integrated global virtual schema of all data sources that is expressed in XML. MOMIS creates a global virtual schema by using different techniques, and by creating a common thesaurus of intra- and inter-schema relationships, which defines an ontology of the terms used to represent the information provided by the different sources. The common thesaurus contains intra-schema relationships extracted by using inference techniques, inter-schema relationships obtained using the lexical WordNet system ([www.cogsci.princeton.edu/wn](http://www.cogsci.princeton.edu/wn)) (which identifies the affinities between inter-schema concepts on the basis of their lexicon meaning) and inter-schema relationships explicitly given by the integration designer. In addition, MOMIS enriches the thesaurus using the Artemis system [10], which evaluates structural affinities among inter-schema concepts and ODB-Tools Engine [2], a tool based on Description Logics which performs checking consistency and subsumption computation. As an example of our integration methodology, we show how it is possible to define a mapping between a fragment of the ECCMA/UNSPSC and a fragment of the ecl@ss standard. With respect to previous works on MOMIS, we introduce a wrapper for semistructured data able to map XML/XML-Schema/RDF file into the common languages of MOMIS; a new method to create a mapping between XML/RDF sources and an XML representation of this mapping.

The paper is organized as follows. Section 2 introduces the two chosen e-commerce code product standards, namely ecl@ss and ECCMA/UNSPSC; section 3 describes our methodology and the results of the mapping process, section 4 presents related work and, finally, section 5 gives some concluding remarks.

## 2 Product Classification Systems and e-Commerce

Coding products and services according to standardized classification systems are useful for speeding up commerce among companies. In addition, the development of e-commerce solution has rapidly increased the requirement of machine-readable product names that assist marketing and sales functions to find customers and provide better customer and distribution channel services.

By inserting the codes in various electronic trade documents and media such as product catalogs, Web sites, purchase orders, invoices, inventory/sales advices, and others, 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 labor-efficient ways that would not be possible without the codes. A useful product classification scheme 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 specificity that best suits its purposes and situation. In addition to maintain a hierarchical taxonomy, a classification scheme must be constantly revised (to add new products and modify existing structures to adapt to changing market offers), it must be responsive to industry (because delays hurt business), and code assignments to products and services must be impartial (to prevent unfairly promoting one company's products at the expense of others) [15]. Within the different standard classification systems proposed, the most used into U.S. is the United Nation Standard Products and Services Code System (UNSPSC), a hierarchical classification with five levels. The levels allow users to search products more precisely (because searches will be confined to logical categories and eliminate 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:

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

In the e-commerce area, the ECCMA (*Electronic Commerce Code Management Association*) ([www.ucec.org](http://www.ucec.org)), has proposed an initiative to enhance the UNSPSC with local attributes to describe the bottom level. The current version consists of more than 16,000 terms.

Another standardization code, used by the statistical agencies of the United States, is the North American Industry Classification System as the industry classification system (NAICS) ([www.ntis.gov/product/naics.htm](http://www.ntis.gov/product/naics.htm)). Finally, an important european initiative that built a new classification scheme from scratch is ecl@ss ([www.eclass.de](http://www.eclass.de)). Ecl@ss is a standard for information exchange 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 expert as well as the occasional user 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 previously mentioned product classification systems are only three of the many proposed and used in B2B marketplaces, where industrial standard are emerging to define the overall interchange process (RosettaNet, ebXML, OAGIS, BizTalk, xCBL, cXML, ...). Each of these proposals defines a protocol for the communication and the data description structure (often described in XML) in order to realize an e-commerce orchestration framework: by using different protocols a reconciliation of product information must be defined.

### 3 Reconciliation of Different Standards

In this paper, we propose an information reconciliation methodology, implemented within the MOMIS system, for the product mapping and reclassification among different code classification systems. The methodology is shown over a fragment of ECCMA/UNSPSC and ecl@ss standard but is easy scalable to the whole code system and can be used, without loss of generality, over any other hierarchical product classification system.

Our methodology uses MOMIS in order to obtain a mapping between elements of the different schemas that correspond semantically to each other. MOMIS is a system designed to provide a global virtual schema of a set of sources to be integrated. We show that the use of MOMIS at the metadata level, i.e. the schemas involved in the integration process describing the two chosen e-commerce standards, is effective to perform the mapping process between the two chosen standards in a semi-automatic way. In [7] it is provided a largely orthogonal classification of the algorithms used by match systems. On the basis of these criteria, our approach may be described as follows:

- ✂ *Schema derived*: our algorithm considers schema-level information. We are considering how to apply extensional knowledge in the process.
- ✂ *Matching granularity*: the match can be performed for combinations of objects, such as complex schema sub-graph. By setting specific parameters, we have the control of the dimension of sub-graph matched.
- ✂ *Language derived*: Our matcher uses a linguistic-based approach by interacting with a lexical database system (WordNet).
- ✂ *Auxiliary information based*: Our approach exploits further information given by the user input.

#### 3.1. Overview of the MOMIS System

MOMIS (see Fig. 1) follows a “semantic approach” to information integration based on the conceptual schema, or metadata, of the information sources, and on the mediator architecture [25]. In the MOMIS system, each data source provides a schema and a global virtual schema of all the sources is obtained in a semi-

automatical way. The global schema has a set of *mapping descriptions* that specify the semantic mapping between the global schema and the sources schema.

The system architecture is composed of functional elements that communicate using the CORBA standard. A data model,  $ODM_{13}$ , and a language,  $ODL_{13}$  are used to describe source schemas.  $ODL_{13}$  and  $ODM_{13}$  have been defined as subset of the corresponding ones in ODMG, augmented by primitives to perform integration.  $ODL_{13}$  is a source-independent language and it is used to describe heterogeneous schemas of data sources. In particular,  $ODL_{13}$  includes the following *terminological relationships*:

- SYN (synonym of) is a relationship defined between two terms  $t_i$  and  $t_j$  where  $t_i \neq t_j$  that are synonyms in every involved source.
- BT (broader terms) is a relationship defined between two terms  $t_i$  and  $t_j$  where  $t_i$  has a broader more general meaning than  $t_j$ . The opposite of BT is NT (narrower terms)
- RT (related term) is a relationship defined between two terms  $t_i$  and  $t_j$  that are generally used together in the same context in the considered sources.

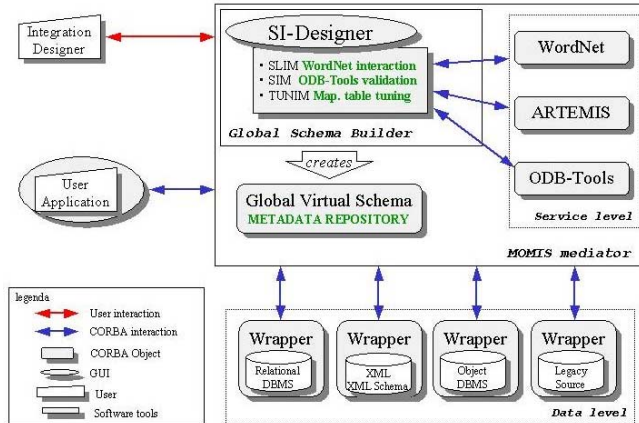


Fig. 1. The MOMIS Architecture

To interact with a specific local source, MOMIS uses a *Wrapper*, which has to be placed over each source. The wrapper translates metadata descriptions of a source into the common  $ODL_{13}$  representation. The core of the MOMIS system is the *Mediator*. The Global Virtual Schema (GSB) module processes and integrates descriptions received from wrappers to derive the global shared schema by interacting with different service modules, namely ODB-Tools, an integrate environment for reasoning on object oriented database based on Description Logics [2], WordNet lexical database that supports the mediator in building lexicon-derived relationships, and ARTEMIS tool that performs the clustering operation [10].

### 3.2. Wrapping of Source Schemas

To manage the information heterogeneity, a mediator system typically encapsulates each source by a wrapper, which logically converts the underlying data structure to the common data model. In this way, the wrapper architecture and interfaces are crucial for managing the diversity of data sources [24]. In particular, during the MOMIS integration process, the wrapper translates the schema of a source into the common data model of the mediator. For a conventional structured information source (e.g. relational databases, object oriented databases), schema description is always available and can be directly translated into the common data model. For semistructured information sources (e.g., Web data sources), a schema description is in general not directly available at the sources, in fact, a basic characteristic of semistructured data is that they are “self-describing”, hence the information associated with the schema is specified within data [8]. According to the different proposed models [8] [22], MOMIS represents semistructured information sources as rooted, labeled graphs with the data (e.g., an image or text) as nodes and labels on edges. A semistructured object can be viewed as a triple of the form  $\langle id, label, value \rangle$ , where  $id$  is the object identifier,  $label$  is a string describing what the object represents, and  $value$  is the value, that can be atomic or complex. The atomic value can be integer, real, ... while the complex value is a set of semistructured objects, that is, a set of pairs  $(id, label)$ . A complex object can be thought as the parent of all the objects that form its value (children objects). An object can have one or more parents. In semistructured data models, labels are descriptive as much as possible. Generally, the same label is assigned to all objects describing the same concept in a source. To represent the schema of a semistructured source  $S$ , we introduce the notion of *object pattern*. All objects  $so$  of  $S$  are partitioned into disjoint sets, such that all objects belonging to the same set have the same label. An object pattern is then extracted from each set to represent all the objects in the set. Then, an object pattern is representative of all different objects that describe the same concept in a semistructured source. According to our data model ( $ODM_{\mathcal{D}}$ ), we developed a wrapper to manage XML and RDF(S) files. The wrapper aims to map the data model of an XML file into the corresponding object pattern model.

The Extensible Markup Language (XML) is a W3C recommendation and it arises as a language to describe information sources by using a universal format. One of the main goals of this standard is to exchange files across the Internet. An XML file may be thought as self-describing like a semistructured data source. The main analogies may be summarized as follows:

- *object pattern* attribute ! XML tag
- *object pattern* ! DTD element
- atomic value of an *object pattern* attribute ! PCData value

By using this mapping, the XML data model allows to describe semistructured information sources according to our data model. The XML wrapper parses the DTD associated to each well-formed XML file and generates a translation from an XML statement into an  $ODL_{\mathcal{D}}$  statement. This mapping implies some critical aspects due to the lack of semantics of XML w.r.t.  $ODL_{\mathcal{D}}$ . In particular, the most relevant are: the order in which attributes are described in the DTD, the translation of the concept of attribute from XML language into  $ODL_{\mathcal{D}}$  language, the poor type system provided by XML and the weak semantics of intra-schema references. In this last case, to avoid

loss of information during the translation process, the designer may be asked to supply further information by a graphical interface. XML Schema, a recent W3C standard, allows to express more semantics on structures and datatypes, by using a XML-based language. Our wrapper is able to integrate XML-Schema sources generating a more significant translation in ODL<sub>13</sub>.

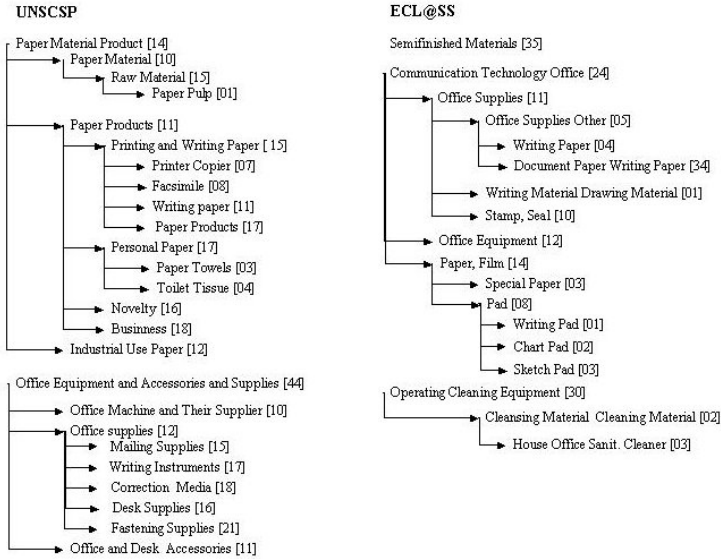


Fig. 2. The ECCMA/UNSPSC and ecl@ss writing paper fragment

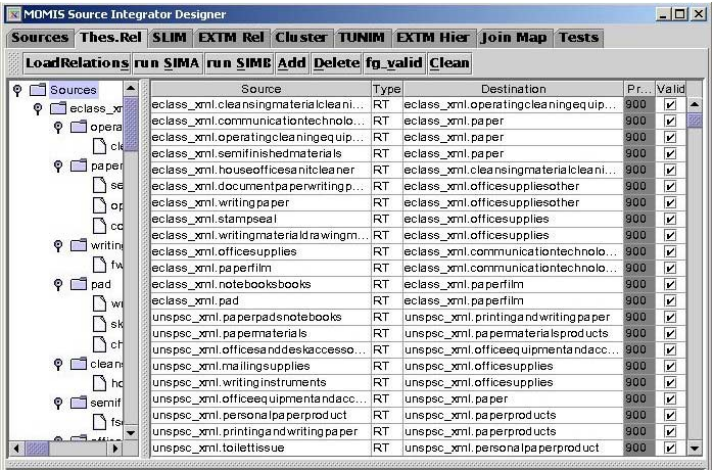
The Resource Description Framework (RDF) is a foundation for processing metadata; it aims at providing a method to describe metadata in a manner, which guarantees the interoperability among different sources on the web. RDF model and syntax may be expressed in XML. In this way, a standard model to represent knowledge and a standard language to describe this knowledge are provided for applications on the web. An RDF-Schema is given in order to interpret and manage the statements contained in a RDF data model. RDF Schema defines a *schema specification language* that may be used to model the specific domain of knowledge. The XML wrapper has been extended in order to manage the knowledge generated from the RDF description of the sources that use XML language to express their own information. In particular, the wrapper translates *RDFS classes* into *object pattern*, *RDF properties* into *object pattern attribute*, the *RDFS subclassOf* property into a *parent-child relationship* and the *RDFS seeAlso* property into a *part-of relationship*.

### 3.3. The MOMIS Integration Process

In order to create a global virtual schema of involved sources, MOMIS generates a common thesaurus of terminological intensional and extensional relationships describing intra and inter-schema knowledge about classes and attributes of the source

schemas. On the basis of the common thesaurus contents, MOMIS evaluates affinity between intra and inter-sources classes and groups similar classes together in clusters using hierarchical clustering techniques. A *global class*, that becomes representative of all the classes belonging to the cluster, is defined for each cluster. The global view for the involved source data consists of all the global classes. The MOMIS methodology is supported by a graphical tool, the Source Integration Designer, SI-Designer in short.

Let us apply the MOMIS methodology to a fragment of the two standard targets. Both the sources represent an extract of classification standard related to the domain of writing paper (Fig. 2). Since ecl@ss contains a standard set of attributes only at the last level and ECCMA/UNSPSC is not descriptive on the attribute level (even if it will be published the EGAS schema containing the attribute level), in the following we take into account only the category names. We assume that these standard schemas have been provided using XML-based files.



The screenshot shows the MOMIS Source Integrator Designer window. The 'Sources' tab is active, displaying a tree view on the left and a table of relationships on the right. The table has columns: Source, Type, Destination, Pr..., and Valid. The relationships are listed as follows:

Source	Type	Destination	Pr...	Valid
eclass_xml.cleansingmaterialcleani...	RT	eclass_xml.operatingcleaningequip...	900	✓
eclass_xml.communicationtechnolo...	RT	eclass_xml.paper	900	✓
eclass_xml.operatingcleaningequip...	RT	eclass_xml.paper	900	✓
eclass_xml.semifinishedmaterials	RT	eclass_xml.paper	900	✓
eclass_xml.houseofficesanitcleaner	RT	eclass_xml.cleansingmaterialcleani...	900	✓
eclass_xml.documentpaperwritingp...	RT	eclass_xml.officesuppliesother	900	✓
eclass_xml.writingpaper	RT	eclass_xml.officesuppliesother	900	✓
eclass_xml.stampseal	RT	eclass_xml.officesupplies	900	✓
eclass_xml.writingmaterialdrawingm...	RT	eclass_xml.officesupplies	900	✓
eclass_xml.officesupplies	RT	eclass_xml.communicationtechnolo...	900	✓
eclass_xml.paperfilm	RT	eclass_xml.paperfilm	900	✓
eclass_xml.notebooksbooks	RT	eclass_xml.paperfilm	900	✓
eclass_xml.pad	RT	eclass_xml.paperfilm	900	✓
unspsc_xml.paperpadsnotebooks	RT	unspsc_xml.printingandwritingpaper	900	✓
unspsc_xml.papermaterials	RT	unspsc_xml.papermaterialsproducts	900	✓
unspsc_xml.officesanddeskaccesso...	RT	unspsc_xml.officeequipmentandacc...	900	✓
unspsc_xml.mailingsupplies	RT	unspsc_xml.officesupplies	900	✓
unspsc_xml.writinginstruments	RT	unspsc_xml.officesupplies	900	✓
unspsc_xml.officeequipmentandacc...	RT	unspsc_xml.paper	900	✓
unspsc_xml.personalpaperproduct	RT	unspsc_xml.paperproducts	900	✓
unspsc_xml.printingandwritingpaper	RT	unspsc_xml.paperproducts	900	✓
unspsc_xml.toiletissue	RT	unspsc_xml.personalpaperproduct	900	✓

Fig. 3. A part of the intra-schema derived relationships

Building the Common Thesaurus

An important feature of semantic integration is the availability of a shared ontology providing a reference vocabulary on which to base the identification of heterogeneity and the subsequent resolution for conflicts. To achieve this goal we build a common thesaurus that expresses inter-schema knowledge in the form of terminological knowledge (such as SYN, BT, NT and RT) between classes and attribute names by exploiting WordNet-supplied ontology and Description Logics supplied by ODB-Tool. The common thesaurus is built through an incremental process during which relationships are added in the following order: schema-derived relationships (not modifiable by the designer), lexicon-derived relationships, designer-supplied relationships and inferred relationships. Using the SI-Designer tool, the designer is assisted during all the integration process and can refine lexicon-derived explicitly supplied relationships at each step of the integration process.



*Schema-derived relationships.* MOMIS extracts intensional relations from schemas structure knowledge, by analyzing each  $ODL_{i3}$  schema separately. In particular, MOMIS extracts each intra-schema RT relationships from the specification of foreign keys in relational source schema and from part-of relationship in hierarchical sources (i.e. XML files representation). When a foreign key is also a primary key both in the original and in the referenced relation, MOMIS extracts a BT-NT relationship. BT-NT relationships are also generated from the inheritance relationships in object-oriented schema and from ID-IDREF couples in XML file. Remember that in this latter case the designer, interacting with SI-Designer, has to identify each couple ID-IDREF. Fig. 3 shows some of the relationships extracted from the ecl@ss and ECCMA/UNSPSC writing paper fragments. For example, we observe:

```
<eclass_xml.semifinishedmaterials rt eclass_xml.paper>
<eclass_xml.notebooksbooks rt eclass_xml.paperfilm>
<unspsc_xml.personalpaperproduct rt unspsc_xml.paperproducts>
<unspsc_xml.mailingsupplies rt unspsc_xml.officesupplies>
```

MOMIS Source Integrator Designer

Sources
Thes.Rel
SLIM
EXTM Rel
Cluster
TUNIM
EXTM Hier
Join Map
Tests

Load
Save
Build
LoadST
AutoNotation

Name Path	Relation	Name Path
unspsc_xml.papertowels	RT	unspsc_xml.personal.papertproduct
unspsc_xml.papereproducts.personal.papereproduct	RT	unspsc_xml.personal.papereproduct.papertowels
unspsc_xml.personal.papereproduct	RT	unspsc_xml.personal.papereproduct.papertowels
unspsc_xml.officesupplies.mallingsupplies	NT	eclass_xml.communicationtechnologyoffice
unspsc_xml.mallingsupplies	NT	eclass_xml.communicationtechnologyoffice
eclass_xml.officesuppliesesother	SYN	eclass_xml.officesupplies
eclass_xml.officesuppliesesother	SYN	eclass_xml.officesupplies.officesuppliesesother
eclass_xml.officesuppliesesother	SYN	eclass_xml.communicationtechnologyoffice.officesupplies
eclass_xml.officesuppliesesother	SYN	unspsc_xml.papematerialproducts.papereproducts
eclass_xml.officesuppliesesother	SYN	unspsc_xml.officeequipmentandaccessoriesandsupplies
eclass_xml.officesuppliesesother	SYN	unspsc_xml.officeequipmentandaccessoriesandsupplies
eclass_xml.officesuppliesesother	SYN	unspsc_xml.officesupplies
eclass_xml.officesuppliesesother	SYN	unspsc_xml.officesupplies
eclass_xml.officesupplies	SYN	eclass_xml.officesupplies.officesuppliesesother
eclass_xml.officesupplies	SYN	eclass_xml.communicationtechnologyoffice.officesupplies
eclass_xml.officesupplies	SYN	unspsc_xml.papematerialproducts.papereproducts
eclass_xml.officesupplies	SYN	unspsc_xml.officeequipmentandaccessoriesandsupplies
eclass_xml.officesupplies	SYN	unspsc_xml.officesupplies
eclass_xml.officesupplies	SYN	unspsc_xml.officesupplies
eclass_xml.officesupplies.officesuppliesesother	SYN	eclass_xml.communicationtechnologyoffice.officesupplies
eclass_xml.officesupplies.officesuppliesesother	SYN	unspsc_xml.papematerialproducts.papereproducts
eclass_xml.officesupplies.officesuppliesesother	SYN	unspsc_xml.officeequipmentandaccessoriesandsupplies
eclass_xml.officesupplies.officesuppliesesother	SYN	unspsc_xml.officesupplies
eclass_xml.officesupplies.officesuppliesesother	SYN	unspsc_xml.officesupplies
eclass_xml.communicationtechnologyoffice.officesupplies	SYN	unspsc_xml.papematerialproducts.papereproducts

NO CONNECTION

SAVE TO FILE

SAVE AS XML

DTD translator

CONNECT

QUIT

**Fig. 4.** A part of the lexicon-derived relationships

*Lexicon-derived relationships.* MOMIS extracts the lexical relationships by analyzing different source schemas, according to the WordNet ([www.cogsci.princeton.edu/wn](http://www.cogsci.princeton.edu/wn)) ontology. 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, which are of the many-to-many kind, give rise to several properties, including synonymy, polysemy and so forth. Synonymy is the property of a concept or meaning that can be expressed with two or more words (a synonym group is called a *synset*). Only one synset exists for each concept or meaning. In contrast to synonymy, polysemy denotes the property of a single word that has two or more meanings. For

each element composing the schemas of the involved sources, the user has to choose the associated word. This choice consists of choosing both a base form and a meaning. Our system tries to automatically suggest a base form. Starting from the base form and the meanings associated to each sources' element, the system inserts into the common thesaurus the lexicon-derived relationships obtained by exploiting the names properties stored in WordNet. Fig. 4 shows some of the generated relationships. In particular, we have:

```
<eclass_xml.officeSupplies SYN unspsc_xml.officeSupplies >
<eclass_xml.pad SYN unspsc_xml.notebooksBooks>
<unspsc_xml.paperProducts unspsc NT eclass_xml.writingPaper>
<eclass_xml.paper RT unspsc_xml.officesupplies>
```

*Designer-supplied relationships.* The designer may supply further relationships to capture specific domain knowledge about the source schemas. For example in the ECCMA/UNSPSC the element *mailing supplies* may be considered as a more general concept of the element *stamp* in the ecl@ss standard. This relationship may be defined as follows:

```
<unspsc_xml.mailingSupplies BT eclass_xml.stamp>
```

This is a critical operation because new relationships are added to the common thesaurus and will be used to generate the global view. This means that if the designer supplies incorrect relationships the integration process can produce a wrong result.

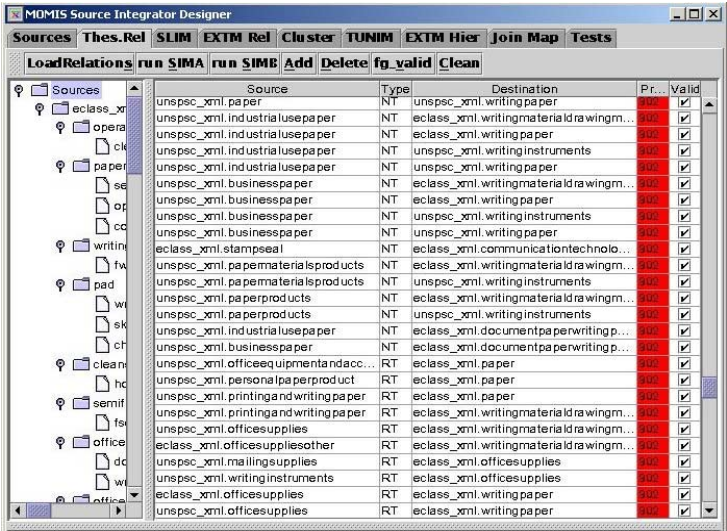


Fig. 5. A part of the new inferred relationships

*Checking consistency and inferring new relationships.* In this step, MOMIS performs reasoning about the common thesaurus relationships by exploiting the subsumption and inheritance computation, reasoning techniques of Description Logics performed by ODB-Tool [2]. Fig. 5 shows some of these relationships:

```
<unspsc_xml.industrialusepaper NT eclass_xml.writingpaper>
<unspsc_xml.businesspaper NT eclass_xml.writingmaterialdrawingmaterial>
<unspsc_xml.toilettissue RT eclass_xml.houseofficesanitcleaner>
```

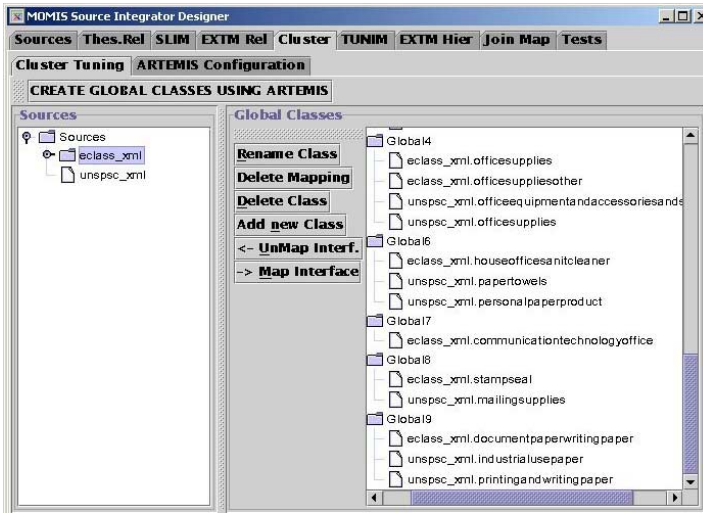


Fig. 6. The generated cluster

### Clustering $ODL_{I3}$ Classes

To integrate the  $ODL_{I3}$  classes of the different sources into a global  $ODL_{I3}$  classes, we employ hierarchical clustering techniques based on the concept of affinity. In this way, we identify  $ODL_{I3}$  classes that describe the same or semantically related information in different source schemas and give a measure of the level of matching of their structure. This activity is performed by ARTEMIS [10], evaluating a set of affinity coefficients (numerical values in the range [0,1]) for all possible pairs of  $ODL_{I3}$  classes on the basis of the relationships of the common thesaurus. Affinity coefficients determine the degree of semantic relationship between two classes based on their names (name affinity coefficient) and on their attribute refined by the data type validation (structural affinity coefficient). A comprehensive affinity value, called global affinity coefficient, is the linear combination of the name and structural affinity coefficients. The output of the clustering procedure is an affinity tree, where the classes themselves are the leaves and intermediate nodes.

### Global Virtual Schema Generation

The latter phase of integration methodology consists in the generation of a global virtual schema composed of  $ODL_{I3}$  global classes derived from the clusters. This is a synthesis activity performed interactively with the designer. Synthesis of clusters of  $ODL_{I3}$  classes requires taking into account semantic heterogeneity, which has to be treated properly to come up with an integrated and uniform representation at the global level. Let  $Cl_i$  be a selected cluster in the affinity tree and  $gc_i$  the global  $ODL_{I3}$  class to be defined for  $Cl_i$ . First, we associate with  $gc_i$  all classes belonging to  $Cl_i$  and a set of global attributes corresponding to the union of the attributes of these classes. The attributes having a valid terminological relationship are unified into a unique global attribute in  $gc_i$ . The attribute unification process is performed automatically for what concerns names according to the following rules: for attributes that have a SYN relationship, only one term is selected as the name for the corresponding global attribute in  $gc_i$ ; for attributes that have a BT/NT relationship, a name which is a

broader term for all of them is selected and assigned to the corresponding global attribute in gc. Furthermore, ODLI3 provides the designer with the syntax and semantics to define mapping rules among global and local attributes and to refine the unification process proposed by the system. As in our example, we considered only product categories (no attributes), we concentrate our attention on the global/local classes mapping. For example, global cluster “Global9”, re-defined by the designer as “Writing paper”, contains the categories “Printing and writing paper” and “Industrial use paper” of the ECCMA/UNSPSC standard and the class “Document paper, writing paper” of the ecl@ss one. In the following table, we show the global class “Writing Paper”, the involved categories and the corresponding code of the native standards.

Sources	Class (Category) name	Code
Mediator Shared Level	Writing Paper	Writing Paper
ECCMA/UNSPSC	Printing and writing paper	14111500
ECCMA/UNSPSC	Industrial use paper	14120000
<a href="#">ecl@ss</a>	Document paper, writing paper	24113400

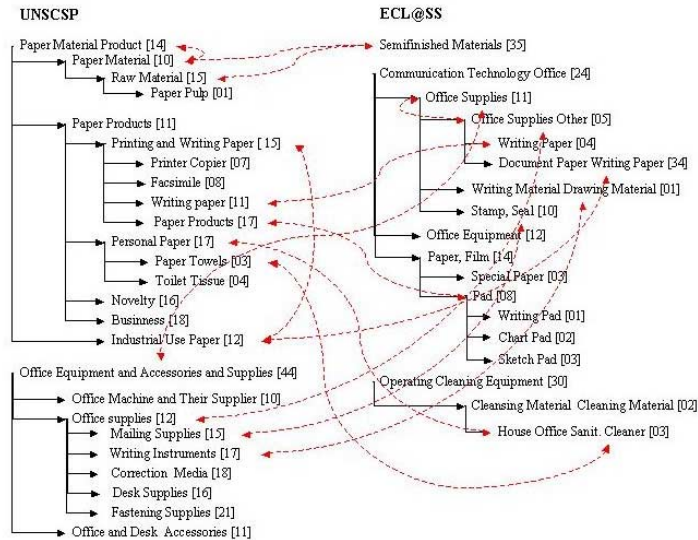


Fig. 7. The obtained mapping

The mapping between concepts of the two standards is graphically shown in Fig. 7, with dashed lines. The system produces the whole global shared schema in an XML-like format, where, for each global class the mapping into the local sources’ classes is described. In particular, the MOMIS output for the obtained mapping is an XML file according to this trivial DTD:

```
<!ELEMENT cluster (interface)>
<!ATTLIST cluster name CDATA #IMPLIED>
<!ELEMENT interface EMPTY>
<!ATTLIST interface name CDATA #REQUIRED>
<!ATTLIST interface code CDATA #REQUIRED>
```

For example, we show how “Writing Paper” cluster is exported using our formalism:

```
<cluster name="Writing Paper">
<interface name="eclass_xml.documentpaperwritingpaper" code="24113400"/>
<interface name="unspsc_xml.industrialusepaper" code="14120000" />
<interface name="unspsc_xml.printingandwritingpaper" code="14111500"/>
</cluster>
```

By considering the mapping among the clusters of the global virtual schema and the local product code, it is possible to compare two different standard categories or to consider as a unique entity all the elements contained in the cluster. In this paper, we have exemplified our mapping approach by referring to a fragment of standard classification related to the paper domain. Nevertheless, due to the scalability of MOMIS methodology, the process may easily extended to map every term between the two considered standards. The approach may be further extended in order to obtain mapping among other product classification standards.

## 4 Related Work

The integration of information has become the main prerequisite for a scalable business process in B2B e-commerce area. The problem to be overcome is related both to the structural and applicative heterogeneity, as well as to the lack of a common ontology, that causes semantic differences among information sources. "Virtual Catalogs" synthesize this approach, as are defined as instruments for the dynamic retrieval of information from multiple catalogs which have been created with the goal to present product data in a unified virtual manner, without directly storing information of the single catalogs [17]. The most popular, the Tsimmis project[21], follows a structural approach and uses a self-describing model to represent heterogeneous data sources. The Garlic project [16] builds a wrapper-based architecture to describe the local source data using an OO language, while the Sims project [1] proposes to create a global schema definition by exploiting the use of description logics. Another approach based on Description Logics is taken in the OBSERVER system to support semantic interoperation and formulation of rich queries over distributed information repositories where different vocabularies are used [20]. Here the idea is that each repository has its own ontology. Inter-ontology relationships are specified in a declarative way (using Description Logics) in an inter-ontology manager module to handle vocabulary heterogeneities between ontologies of different information repositories for query processing. In this respect, our approach tries to extract as much information as possible from source descriptions and from WordNet and we show how this information can be used for integration purposes.

The analysis, discovery, and representation of inter-schema properties are another critical aspect of the integration process and research proposals have appeared on this topic. In DIKE system [23], semi-automatic techniques for discovering synonyms, homonyms and object inclusion relationships from database schemas are described and an algorithm for integrating and abstracting database schemas is proposed. More specific contributes concerning the product integration of information in B2B is provided by Fensel et al. [12]. Their first proposal apply XSL-T [13] technology to B2B document interchange, by defining specific XSL-T rules that directly translate each XML element or attribute of one catalog into XML element of another ones. These rules try to carry out the complete transformation process in one shot, so the

different aspects of integration, like syntax heterogeneity, structural and granularity level of representation mismatches are encapsulated into the rules. For this reason any re-use of such rules is practically impossible. To overcome this problem, the same authors developed a multi-layer framework [19], where three layers (syntax layer, object layer and ontology layer) are proposed for information modeling on the Web. The integration is performed by using only the syntax and the object layer through the translation of the XML catalog (syntax layer) into its normalized RDF data model (object layer), the transformation between a pair of RDF data models of different catalogs and the translation from the data model back into XML target catalogs [14]. In our approach the syntax layer represents the wrapper level, while the object layer and ontology layer correspond to the global virtual schema; in addition, we propose a semi-automatic methodology for the creation of the common ontology.

## 5 Conclusions

In this paper, we proposed a methodology to define the mapping among different e-commerce product classification standards. We have exemplified the methodology by showing how it is possible to create a mapping between a fragment of the ECCMA/UNSPSC and a fragment of the ecl@ss standard, but the process may be successfully applied to other product classification standards. The obtained XML mapping file may be inserted within an electronic marketplace in order to define automatic rules to manage products classified using the ECCMA/UNSPSC and ecl@ss standards. These rules may generate automatic data translation to give to the marketplace seller a unique code representing the same product that is classified by the vendors in different manners. In this way, by given a common manner to describe goods involved in the e-commerce process, the marketplace really becomes the place in which sellers and vendor may interact without any change in their data management system. In the future, we are going to investigate this idea by using the obtained XML mapping file within the MOMIS Query Manager [3] that permits optimized query processing in distributed information system.

## References

- [1] Ambite J. L., Knoblock C. A., *Flexible and scalable cost-based query planning in mediators: A transformational approach*, Artificial Intelligence 118(1-2), (2000).
- [2] Beneventano D., Bergamaschi S., Sartori C., Vincini M., *ODB-Tools: A Description Logics Based Tool For Schema Validation and Semantic Query Optimization in Object Oriented Databases*, Proc. of Int. Conf. on Data Engineering (ICDE-97), Birmingham UK 1997.
- [3] Beneventano D., Bergamaschi S., Guerra F., Vincini M., *Exploiting extensional knowledge for query reformulation and object fusion in a data integration system*, Proc. of the Convegno Nazionale Sistemi di Basi di Dati Evolute (SEBD2001), Venezia, June, 2001.
- [4] Beneventano D., Bergamaschi S., Castano S., Corni A., Guidetti R., Malvezzi G., Melchiori M., Vincini M., *Information Integration: the MOMIS Project Demonstration*, Proceedings of 26th Int.Conf.on Very Large Data Bases (VLDB2000), 000, Cairo, Egypt.

- [5] Bergamaschi S., Castano S., Beneventano D., Vincini M., *Semantic Integration of Heterogeneous Information Sources*, Special Issue on Intelligent Information Integration, Data & Knowledge Engineering, Vol. 36, Num. 1, 215-249, Elsevier Science B.V. 2001.
- [6] Benetti I., Beneventano D., Bergamaschi S., Guerra F., Vincini M., *An Information Integration Framework for e-commerce*, IEEE Intelligent Systems, (Jan/Feb 2002).
- [7] Bernstein P.A., Rahm E., *A survey of approaches to automatic schema matching*, VLDB Journal 10(4): 334-350 (2001).
- [8] Buneman P., *Semistructured Data*, Proceedings of Symposium on Principles of Database systems (PODS97), Tucson, Arizona, (1997), 117-121.
- [10] Castano S., De Antonellis V., De Capitani Di Vimercati S., *Global Viewing of Heterogeneous Data Sources*, IEEE Transactions TKDE 13(2): 277-297 (2001).
- [11] Fensel D., *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*, Springer-Verlang, Berlin(2001).
- [12] Fensel D., Ding Y., Schulten E., Omelayenko B., Botquin G., Brown M., Flett A., *Product Data Integration in B2B E-commerce*, IEEE Intelligent System, 16(4), 2001.
- [13] Omelayenko B., Fensel D., *Ontologies: An Analysis of Integration Problems of XML-Based Catalogues for B2B E-commerce*, In Proc. of the 9th IFIP 2.6 Working Conference on Database (DS-9), April 2001.
- [14] Omelayenko B., Fensel D., *Ontologies: A Two-Layered Integration Approach for Product Information in B2B E-commerce*, Second Int. Conference on Electronic Commerce and Web Technologies, EC-Web 2001 Munich, Germany, September 4-6, 2001.
- [15] Granada Research, *Why Coding and Classifying Products is Critical to Success in Electronic Commerce*, White paper, 2002. [www.un-spsc.net](http://www.un-spsc.net).
- [16] Haas L. M., Miller R. J., Niswonger B., Roth M. T., Schwarz P. M., Wimmers E. L., *Transforming Heterogeneous Data with Database Middleware: Beyond Integration*, IEEE Data Engineering Bulletin, 22(1):31-36 (1999).
- [17] Hull R., *Managing Semantic Heterogeneity in Databases: A Theoretical Perspective*, ACM Symp. on Principles of Database Systems, pp. 51-61, 1997.
- [18] Li H., *XML and Industrial Standards for Electronic Commerce*, Knowledge and Information Systems, 2 (2000), 487-497.
- [19] Melnik S., Decker S., *A Layered Approach to Information Modeling and Interoperability on the Web*, in ECDL 2000 Workshop on the Semantic Web.
- [20] Mena E., Illarramendi A., Kashyap V., Sheth A. P., *OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies*, Distributed and Parallel Databases 8(2): 223-271 (2000), Kluwer.
- [21] Garcia-Molina H., Quass D., Rajaraman A., Sagiv Y., Ullman J., Vassalos V., Widom J., *The TSIMMIS approach to mediation: Data models and Languages*, in Journal of Intelligent Information Systems, JIIS(8)(2): 117-132 (1997).
- [22] Papakonstantinou Y., Garcia-Molina H., Widom J., *Object Exchange Across Heterogeneous Information Sources*, Proc. of Int. Conf. on Data Sources, Taiwan (1995)
- [23] Rosaci D., Terracina G., Ursino D., *A Semi-automatic Technique for Constructing a Global Representation of Information Sources Having Different Formats and Structure*, Proc. of 12th Int. Conf. Database and Expert Systems Applications (DEXA 2001), 2001.
- [24] Roth M. T., Schwarz P. M., *Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources*, Proceedings of 26th Int. Conf. on Very Large Data Bases, 266-275, 1997.
- [25] Wiederhold G., Genesereth M. R., *The Conceptual Basis for Mediation Services*, IEEE Expert 12(5): 38-47 (1997).