A Modeling Approach for Product Classification Systems

Joerg Leukel, Volker Schmitz, Frank-Dieter Dorloff
Department of Information Systems, University of Essen, Germany
{joerg.leukel | volker.schmitz | dorloff}@uni-essen.de

Abstract

Standardized product classification systems play a major role for searching and comparing offered products on electronic markets. Especially in case of large multi-vendor product catalogs classified data becomes an important asset and success factor. The most known systems are UNSPSC and eCl@ss, however they are still developing, and new systems are emerging as well. Classification systems differ not only in content but also in structure from each other. The management and exchange of the systems between market partners must be able to get along with these differences. A common structure model, which can be used to specify XML business documents, is missing so far. This paper discusses the design of classification systems and develops a data model using XML Schema. The model can be used for the transmission of classification systems, thus it is an innovative extension of existing product catalog standards.

1. Introduction

The task of product classification is to assign each product to a product group corresponding to common attributes or application areas. Though classification systems are not a new phenomenon of B2B e-commerce; they are already in use as an instrument of structuring since decades. The field of application is very broad. It ranges from manufacturing, costing and sale (e.g. product catalogs) up to national and international economic statistics.

In B2B e-commerce classification systems gain a new meaning and function. They are an instrument for the access to large e-catalogs. Standardized and supplier-independent classification is an elementary requirement for efficient product search and qualified comparison of products in electronic markets and other catalog-based procurement systems [1]. To describe products in a uniform manner, some classification systems define so-called *sets of attributes*. A set of attributes is assigned to a classification group and contains the necessary product attributes. In e-catalogs that claim to support the classification system each product has be described by the group-depend set of attributes.

Just as little as there is a generally accepted XML standard for business documents today, it is not to be expected, that a single classification system will prevail worldwide and for all branches of industry. Rather we see the development of specific classification systems along the requirements of branches and markets. In consequence classification systems become an object of data management and data exchange [2]. A model for the efficient handling of classification systems is missing so far, that describes classification systems universally and that can be used for the transmission of classification systems. Catalog applications implement search and navigation mechanisms on basis of classification systems, however they cannot import the classification system definitions in a standardized format.

2. Paper Organization and Related Work

Goal of this paper is the development of the outlined model. Aftwerwards we will formulate the requirements on the model (Section 3). Then we will discuss the two essential design aspects in detail, on the one hand the product group hierarchy and on the other hand the sets of attributes (Section 4 and 5). The specification of the model takes place using XML Schema. In Section 6 we will utilize the developed model for an empirical analysis of selected classification systems and XML catalog standards as well. Finally, we will draw some conclusion.

Research literature show two main working areas regarding classification systems. The first area deals with the application of classification systems and the requirements on them (e.g. [3]). Structural and content wise design aspects are described in [4]; they should serve for the evaluation and development of new systems. The structural aspects are derived on the basis of three systems. In [5] the problem of different schemas for the categorization of products and their attributes is examined from the view of relational databases. The conclusion is that concepts of schema integration cannot be transferred directly.

Approaches for the integration of classification systems on a semantic level form a second area. Classification systems are seen as product ontologies that enable a common and accepted communication in the respective product domain [6]. The specification of ontologies takes place using concepts and languages of knowledge representation. An integration approach is presented in [7]; the system imports classification system data using a Wrapper. The supply of this data in a standardized format is not demanded. In [8] a similar, integration-oriented path is taken. [9] proposes an information retrieval approach. The import process of UNSPSC data is not described.

3. Modeling of Classification Systems

Taking in mind, that new classification systems will emerge as well as existing systems are in an on-going process of extension or could even merge with other systems, we believe that classification system are becoming an object of data management in e-business. Application, exchange and integration of classification systems as well as mapping one system to another system makes a data model necessary, that can be used for the named purposes. In the following we outline the requirements on such a model.

The first requirement emerges from the usage in market-based B2B relationships: The model must have a *general validity*, i.e. the model must be able to represent all existing classification systems and their relevant system properties. The general validity leads to a model consisting of components. A real classification system makes use of more or less of these components (e.g. set of attributes).

The model should not only serve for the representation of real-world classification systems, but may also contain components and properties, which show up from the requirements of B2B e-commerce meaningfully and are still little or not realized in commercial systems (e.g. attribute groups).

Since classification systems are not limited to a single language, *language-independency* is to be called for the model, i.e. no language-specific keys may occur.

As stated above attributes are defined for each group on the leaf level. To minimize the number of attributes in the system and to come to non-redundant attributes, classification systems use a defined attribute pool. This is realized by a *dictionary concept*.

The steady semantic extension and changes of classification systems lead to different versions of the same system over time. Consequently the management of

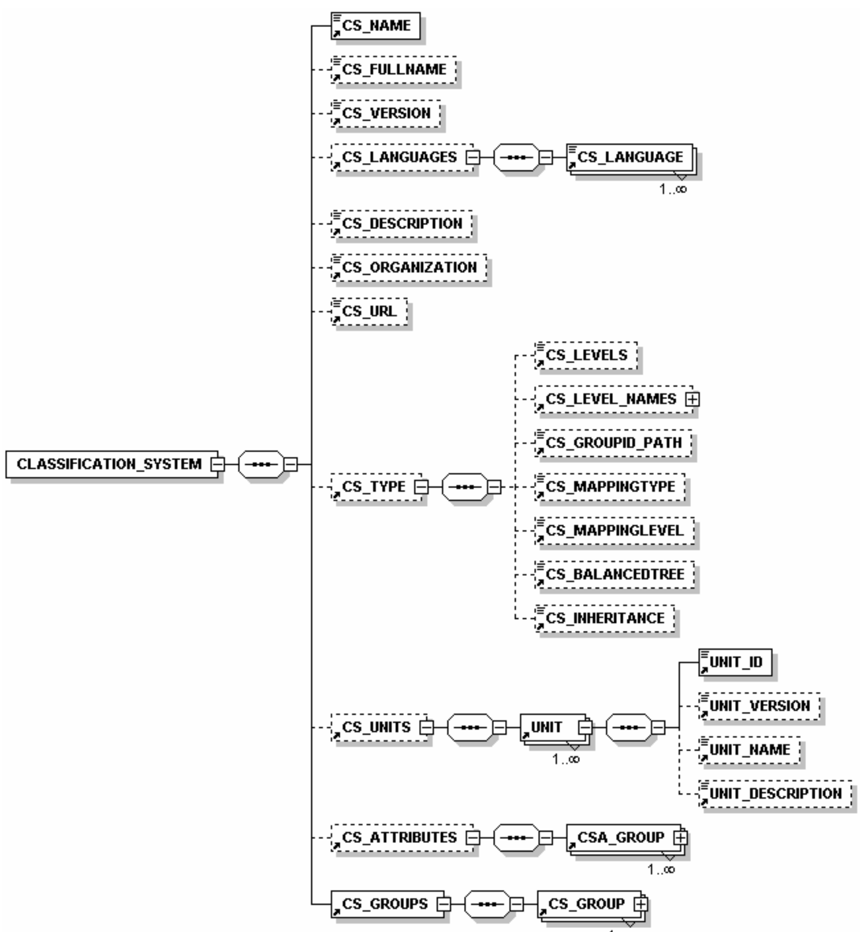

Figure 1. Overview of the Product Classification System Model

versions must be supported. All elements that can be changed independently must add a version number to their identification.

The specification in XML Schema allows to use the model for the exchange of classification systems as a business document, because a transmission format is defined at the same time. The W3C XML Schema Definition Language is an XML language for describing and constraining the content of XML documents [10]. We will state the names of the relevant data elements in angular brackets, e.g. [CLASSIFICATION_SYSTEM] as the root element.

The essential design parameters of classification are components of our model. The model is composed of meta information, units, attributes, product groups and sets of attributes. Figure 1 gives an overview of the model; it shows a graphical representation that is built by the tool XMLSpy.

Among the meta information are those information that describe and characterize the system, but do not define its structure and content. The meta information serves as identification and explanation of the system. The belonging data elements are for the present: an identifying name, e.g. "eCl@ss-4.1-de" [CS_NAME]; version, e.g. "4.1" [CS_VERSION]; language according to ISO, e.g. "DE" [CS_LANGUAGE]; long name, e.g. "eCl@ss, German, version 4.1" [CS_FULLNAME]; describing text, e.g. "Classification system of the Institute of the German Economy" [CS_DESCRIPTION]; number of levels, e.g. "4" [CS_LEVELS]; names of the levels as a list [CS_LEVEL_NAMES]; name of the organization, e.g. "eCl@ss e.V." [CS_ORGANIZATION] and the internet address that provides the system definition, e.g. "http://www.eclass-online.com" [CS_URL].

Afterwards we will introduce additional meta information; their necessity and meaning emerges only from the model description.

4. Modeling of Product Groups

Classification systems are hierarchical structures. Two necessary requirements must be taken into account: Firstly, each product may be associated to one group only [CS_MAPPINGTYPE=single]; secondly, the association has to be done on the lowest level (leaf level) [CS_MAPPINGLEVEL=leaf]. If one of these conditions is not fulfilled, then the system is not a classification system, but a so-called catalog group system [11]. The design aspects of classification systems can be applied to catalog group systems. Hence we extend the purpose of the model to catalog group systems as well. Another, however not necessary, characteristic of classification systems is, that the tree structure shows an identical depth in all part trees; in other words, the tree is balanced [CS_BALANCEDTREE=true]. To sum it up, three additional meta information are serving for the characterization of the tree structure [CS_TYPE], two of them distinguish the system types.

Each group [CS_GROUP] must be addressable by an identifier [CSG_ID]. At least a language-dependent group name is assigned [CSG_NAME]. If the identifier is built in such a manner, that it describes the path leading to the group in the tree, then we call it a composite identifier. For example, this type is used in UNSPSC and eCl@ss. In other cases it is necessary to give an explicit reference to the superior (or parent) group [CSG_PARENT_ID]. Catalog applications need these references to set up the tree structure. The type of the identifier is determined as a meta information for the whole system [CS_GROUPID_PATH=true or false].

A text can be added for the further explanation of the group and its content [CSG_DESCRIPTION]. The same

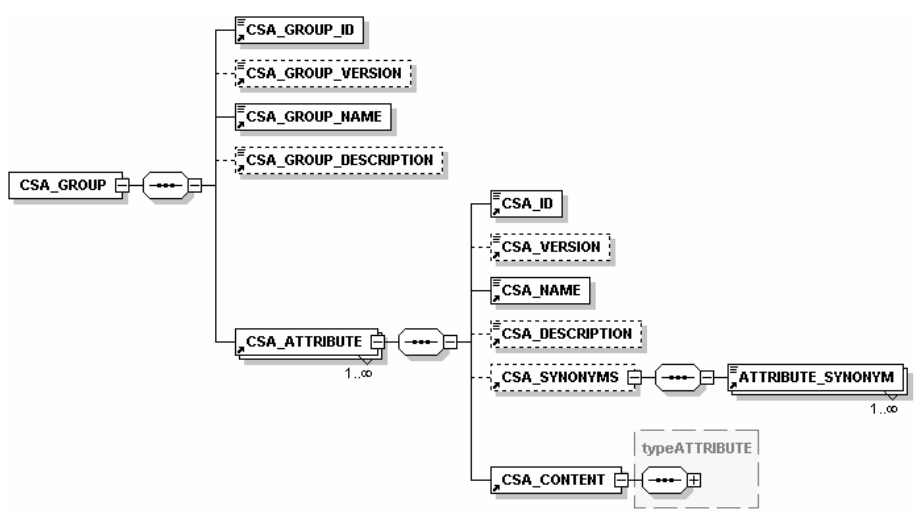

Figure 2. Modeling of Attributes (1)

task can be taken on by an image – or more generally: by a multimedia object [MIME] – that shows typical products of the group or other information. Alternative terms for the group name are filed in a list of synonyms [CSG_SYNONYMS].

Additionally, it is possible to refer to an external definition of certain sub trees instead of specifying the sub trees within the classification system [CSG_ORGANIZATION, CSG_URL]. By it distributed classification systems that consist of several different sub systems are supported. For example a web-service at the given URL provides the system definition.

5. Modeling of Attributes

All attributes are defined group-independently in a dictionary [CS_ATTRIBUTES] to avoid redundant definitions of equal attributes. Sets of attributes can be assembled easily on basis of these definitions.

For the management of attributes and their application in catalog systems a division of the attribute pool into attribute groups can be helpful. An example: A group “Dimension” could contain the attributes “Size”, “Height“ and “Width“. Hence all attributes are grouped, which address similar or related product features [CSA_GROUP]. Each attribute group is characterized by an ID, a name and if necessary a description (see figure 2).

Each attribute of the dictionary [CSA_ATTRIBUTE] has a numeric identifier [CSA_ID] and a meaningful and delimiting designation [CSA_NAME] (e. g. “Material of the surface”). This can be complemented by a text [CSA_DESCRIPTION] and a list of synonyms [CSA_SYNONYMS].

However, a statement belongs to the definition of an attribute, which data content can be expressed [CSA_CONTENT]; therefore a data type, a domain and a unit can be determined (see figure 3). Many domains are already standardized by international norms, therefore the model contains further data elements for the explanation of these standards [AD_NAME, AD_VERSION, AD_DESCRIPTION, AD_URL]. Often attributes have an unit of measurement, e.g. kg, inches, mm. On the other hand the spectrum of units is a limited set that is for the most part described in ISO standards. For this reason it is appropriate to define all units independently from the attributes [CS_UNITS] and to assign them to attributes if needed [ATTRIBUTE_UNIT_IDREF]. Another aspect of attributes is the use of attribute values in formulas. Hence the expression of a common symbol can be useful [ATTRIBUTE_SYMBOL]. Example: attribute “Resistance”, unit “Ohm”, symbol “R”.

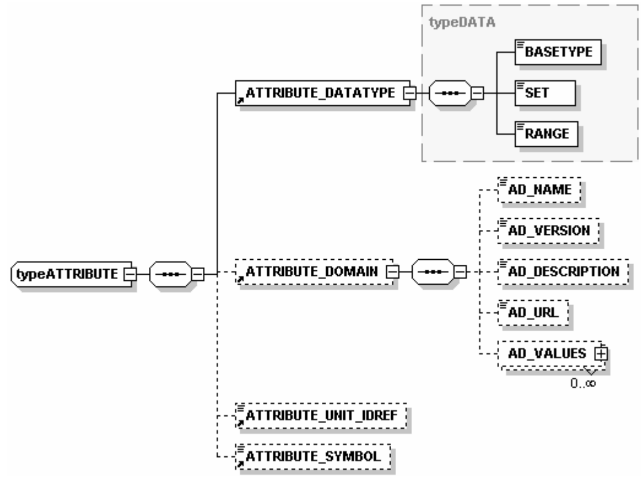

Figure 3. Modeling of Attributes (2)

On basis of the attribute definitions sets of attributes can be assembled [CSG_ATTRIBUTESET]. As we can see in eCl@ss [12], specific sets of attributes are defined only on the lowest group level. Besides, general attributes with relevance to all groups can exist, too. From the view of the management and maintenance and classification systems the concept of inheritance is be very helpful: Attributes can be assigned to groups on every level of the tree structure, then they are passed down to all groups below. The general attributes need to be defined only once at the top level. The use of inheritance is signaled as a meta information [CS_INHERITANCE=true or false].

Compiling attributes to a set of attributes includes both the reference to attributes [CSGA_IDREF] as well as the definition of attribute properties. Thereby it must be possible to overwrite properties that have already been defined in the dictionary. Only in this way the domain of an attribute can be limited or a unit can be selected, which is more meaningful in the group context (e.g. millimeters for the length of nails instead of meters). In practice the attribute properties are set often very generally in the dictionary, e.g. without units and domains, and the specification is detailed at the group level.

Moreover a set of attributes can add properties to its attributes. If the attribute has to be filled for all products of the group, then it is a mandatory attribute [CSGA_MANDATORY=true]; otherwise it is an optional attribute. The attribute groups serve to define the sequence in which attributes are listed in catalog applications. Within an attribute group the sequence can be defined by a numeric value [CSGA_ORDER]; at the same time it represents the only order criterion for those systems that do not support attribute groups.

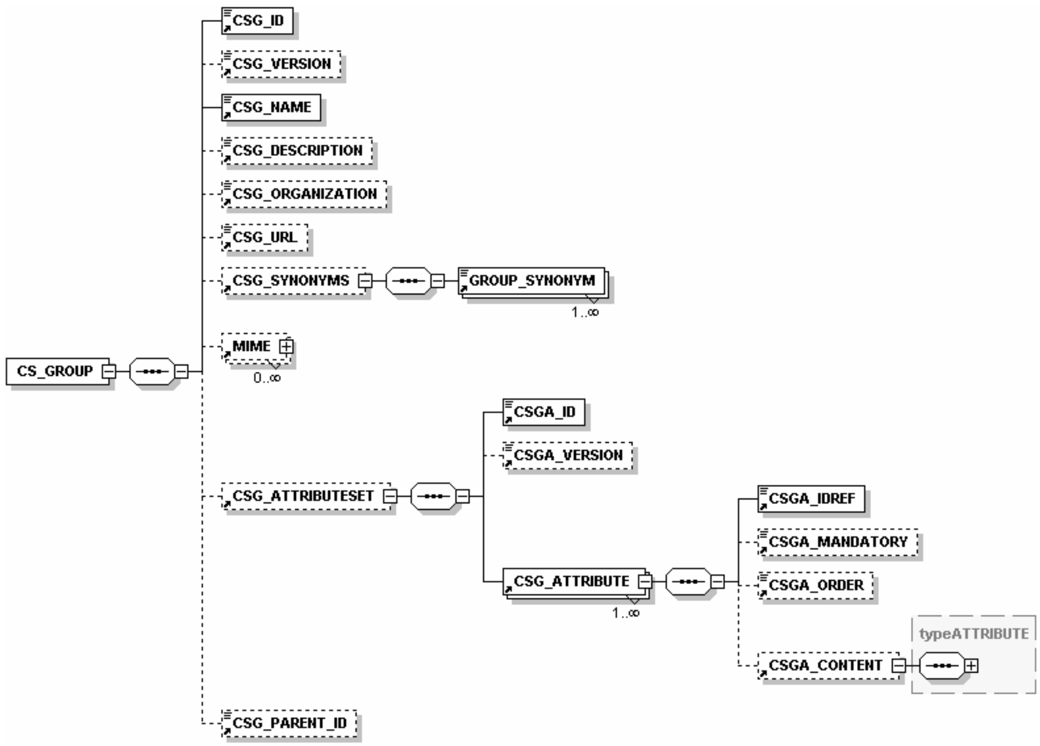

Figure 4. Modeling of Product Groups and Sets of Attributes

6. Application of the Model

The developed model is the result of application requirements and the analysis of classification systems in practice. Now it should serve for a uniform description of industrial classification systems and XML catalog standards.

The table 1 shows, which design parameters and components of the model can be found in four selected product classification systems: eCl@ss, ETIM [13], RosettaNet Technical Dictionary (RNTD) [14] and EGAS [15] as an extension of UNSPSC [16]. ETIM and RNTD are vertical systems developed for the wholesale of electro technical products respectively for electronic and IT components.

The systems itself are documented by non-formal and formal specifications. Though only RNTD is specified by an XML document. All other systems use simple Excel or comma-separated value (CSV) files as containers and provide very few semantics. Since the files differ completely in structure, importing them into target systems is a time-consuming task.

Before we apply the model to XML catalog standards we have to ask, which standards are capable of transmitting classification systems. We observe, that many catalog standards are confined to the classification of products by allowing a reference to the classes and attributes. cXML, eCX and EAN.UCC belong to this group of irrelevant standards. ebXML is a framework and does not specify business documents. In contrast the following standards provide special document types or data elements for classification systems: BMEcat [17], OAGIS [18] and xCBL [19].

The analysis of the three catalog standards was done on data element level. Our model serves as an analysis scheme. For each standard was tested whether the relevant issues can be represented and if so by which data elements. A cut out of the comparing analysis is contained in table 2 (however, the data element names are missing due to limited space).

7. Conclusion and Further Work

In this paper we have discussed the design and components of product classification systems in B2B e-commerce and suggested a data model based on XML. All things considered, none of the four selected industrial classification systems realizes all design parameters of our model. The systems themselves are documented quite differently. System specifications are often provided in proprietary formats; hence their processing in catalog systems is less automated. Especially, the organizations that develop and maintain classification systems provide no XML data (exception: RosettaNet).

The application of XML e-business standards for the transmission of classification systems is hardly possible,

Table 1. Analysis of selected Product Classification Systems

Area	Content	eCl@ss 4.1	ETIM 1.1	RNTD 1.4	EGAS 1.0
Meta Information	Levels	4	2	2	4
	Languages	2 (4)	1	1	multiple
	Attribute Inheritance	-	-	-	-
Units	Definitions	-	+	-	-
Attributes	Descriptions	+	-	+	+
	URL	-	-	+	-
	Synonyms	+	-	+	-
	Groups	-	-	-	-
	Datatype	+	+	+	+
	Value Orders	-	+	-	-
	Unit	+	+	+	-
	Symbol	+	-	+	-
	Mandatory vs. Optional	-	-	-	-
	Attribute Orders	-	+	-	-
Groups	Description	+	-	+	-
	URL	-	-	+	-
	Synonyms	+	+	+	-
	Multimedia Objects	-	-	-	-
	Set of Attributes	+	+	+	+

Table 2. Analysis of selected XML Catalog Standards

Area	Content	BMEcat 1.2	OAGIS 7.2.1	xCBL 3.5
Meta Information	Identifier	+	+	+
	Fullname	+	-	+
	Description	+	+	+
	Languages	+	-	+
	URL	-	-	+
	Tree Type	partially	-	-
	Attribute Inheritance	-	-	+
Units	Identifier	+	-	+
	Description	+	-	-
Attributes	Identifier	+	+	+
	Description	+	+	-
	URL	-	-	-
	Synonyms	-	-	-
	Groups	-	+	-
	Data Type	+	+	+
	Value Orders	-	+	-
	Unit	+	+	+
	Symbol	-	-	-
	Mandatory vs. Optional	+	+	+
	Attribute Orders	+	+	-
Groups	Identifier	+	+	+
	Description	+	+	+
	URL	-	-	-
	Synonyms	+	-	-
	Multimedia Objects	-	+	-
	Set of Attributes	+	+	+

because no one of the analyzed standards is capable to transfer all systems completely. The loss of structural information is in many cases very high. Matching the tables 1 and 2 can proof this.

To solve the described problems we developed an XML Schema that covers all design parameters and is able to describe all classification systems. The benefit using XML Schema language instead of ERM, UML or RDF is providing a format immediately, which can transfer real classification systems in all details.

Additionally, the transmission of classification systems is already an important task in B2B relationships, since suppliers and marketplaces need the classification system data as a part of their core data while creating and maintaining catalogs. Since classification systems are changing in content, the developed document-oriented model has to be extended to a request-response-model that also covers messages for requesting and transferring partial updates of classification systems.

8. References

[1] M. Ginsburg, J. Gebauer and A. Segev, "Multi-Vendor Electronic Catalogs to Support Procurement: Current Practice and Future Directions", *Proceedings of 12th International Bled Electronic Commerce Conference*, Bled, Slovenia, 1999.

[2] M. Stonebraker and J.M. Hellerstein, "Content Integration for E-Business", *Proceedings of ACM SIGMOD/PODS 2001*, Santa Barbara, California, 2001.

[3] A.M. Fairchild and B. Vuyst, "Coding Standards benefiting Product and Service Information in E-Commerce", *Proceedings of ICEIS 2001*, Setúbal, Portugal, 2001, pp. 991-995.

[4] B. Otto and H. Beckmann, "Klassifizierung und Austausch von Produktdaten auf elektronischen Marktplaetzen", *Wirtschaftsinformatik*, 2001, pp. 351-362.

[5] W. Ng, G. Yan and E.-P. Lim, "Heterogeneous Product Description in Electronic Commerce", *ACM SIGeCom Exchanges*, 2000, pp. 7-13.

[6] D. Fensel, Y. Ding, B. Omelayenko et al., "Product Data Integration in B2B E-commerce", *IEEE Intelligent Systems*, 2001, pp. 54-59.

[7] O. Corcho and A. Gómez-Pérez, "Solving Integration Problems of E-Commerce Standards and Initiatives through Ontological Mappings", *Proceedings of IJCAI 2001 Workshop on E-Business & the Intelligent Web*, Seattle, USA, 2001.

[8] C. Quix, M. Schoop and M. Jeusfeld, "Business Data Management for Business-to-Business Electronic Commerce", *SIGMOD Record*, 2002, pp. 49-54.

[9] Y. Ding, M. Korotkiy, B. Omelayenko et al., "GoldenBullet: Automated Classification of Product Data in E-commerce", *Proceedings of BIS 2002*, Posen, Poland, 2002.

[10] W3C, *XML Schema Part 0: Primer. W3C Recommendation*, 2 May 2001. URL: http://www.w3.org/TR/xmlschema-0, 2001.

[11] C. Hümpel and V. Schmitz, "BMEcat - an XML standard for electronic product data interchange". *Proceeding of the 1st German Conference XML 2000*, Heidelberg, Germany, pp. 1-11.

[12] eCl@ss e.V., *eCl@ss – Standardized Material and Service Classification*, Version 4.1. URL: http://www.eclass-online.com, 2002.

[13] ETIM e.V., *ETIM Klassifikationsmodell*, Version 1.1. URL: http://www.etim.de, 2001.

[14] RosettaNet, *RosettaNet Technical Dictionary*, V 1.4. URL: http://www.rosettanet.org, 2001.

[15] ECCMA, *EGAS – ECCMA Global Attribute Schema*, Technical Manual 1.1. URL: http://www.eccma.org/ega, 2001.

[16] UNDP, *United Nations Standard Products and Services Code (UNSPSC)*, Version 5.0301. URL: http://www.un-spsc.net, 2002.

[17] V. Schmitz, O. Kelkar and T. Pastoors, *Specification BMEcat*, Version 1.2. URL: http://www.bmecat.org, 2001.

[18] Open Applications Group, *Open Applications Group Integration Specification*, Release 7.2.1. URL: http://www.openapplications.org, 2001.

[19] CommerceOne, Inc., *XML Common Business Library (xCBL)*, version 3.5. URL: http://www.xcbl.org/xcbl35/xcbl35.html, 2001.