The E-Commerce Product Classification Challenge

Ellen Schulten, Heloise Ontology Associates
Hans Akkermans, Free University Amsterdam
Nicola Guarino, LADSEB
Guy Botquin, Content Europe
Nelson Lopes, Institut der Deutschen Wirtschaft Köln
Martin Dörr, ICS-FORTH
Norman Sadeh, Carnegie-Mellon

[Final version v1.0, 11 July 2001.
Intended for IEEE Intelligent Systems Magazine, special issue on Intelligent E-business (July/August 2001), as an addition to the Guest Editor Introduction (by Hans A).]

This article launches an international research challenge in the area of intelligent e-business. The challenge is to come up with a generic model and working solution that is able to (semi-)automatically map a given product description between two different e-commerce product classification standards.

B2B E-Commerce: in search for a Lingua Franca or Rosetta Stone

A fundamental premise - and a major economic driver - behind B2B Electronic Commerce is that labor-intensive and time-consuming human interactions can be replaced with (semi-)automated Internet-enabled processes. Looking at actual implementations, we indeed see 'simple' applications, such as product search and selection without the intervention of a sales representative, and more sophisticated solutions, such as server-to-server communication for inventory replenishment of enterprises.

Nevertheless, the hard reality of e-business, and the slower than (some industry watchers) expected adoption of electronic buying, points to the complexity of replacing human interactions by computer. Of course this is not difficult to understand. In the human world, dialog is structured by grammatical, semantic, and syntactic rules that live in a shared context of social and cultural conventions. The young e-commerce world still lacks this rich background, and we are still far from really achieving the vision of a 'Universe of Network Accessible Information' - as the Web is defined by the W3C. The existing protocol and hypertext standards appear to be insufficient for the exchange of unambiguous information between network devices - an essential condition for electronic commerce. This is especially true in business between enterprises, where interoperability between different systems, back-end system integration, complex business processes, and stringent business rules have to be taken into account.

Notwithstanding the lack of proper language standards, many enterprises already started to populate the web in some sort of community, be it an intranet, an extranet or a marketplace, and as a consequence, almost as many speak each a different language.

Multi-linguality is not a problem in itself, on the contrary, it often allows for creativity and refreshing diversity. However, when the means for translation are absent, things get tricky. And this is exactly the case in the B2B e-commerce. One might compare the situation to the building of the Tower of Babel with everybody crafting his own piece. The results are legendary.

The right time: research and industry communities to join forces

Therefore, industry and research communities currently join forces in multiple, not-for profit, standardization and harmonization initiatives. Interestingly, the analogy with the Tower of Babel returns in the metaphors used by such initiatives. Examples are Lingua Franca (often used to refer to XML) and the Rosetta Stone, from which Rosettanet.org derives its name. Lingua Franca is the trade language used at the beginning of the 18th century by numerous language communities around the Mediterranean in order to enable international trade. The Rosetta Stone is a black basalt slab that was inscribed with the same message in three languages, enabling scholars to crack the code of hieroglyphics. These metaphors are not explained here just for historical interest. They actually indicate two different roads that can be taken to arrive at interoperability: one can either define a new standard or build a layer that provides a map between different standards -or, maybe more realistic, find a path connecting them.

The need for consensus in a trading community arises at many different levels. This is reflected in the different focus areas of these harmonization initiatives. Extensive effort is put in consensus building on business processes and business documents, catalog representation, and business directories. But there are comparatively few initiatives focusing on the apparently simpler harmonization of the basic building blocks of any commercial transaction: the ***product descriptions*** themselves. Basically, two functions can be associated with product description standards: (1) from a set of requirements, the client needs to narrow down the search for the complete set of applicable products; (2) the client needs to comprehend the individual product description to the precision needed for a specific application. It is good practice to use classification systems and standardized attributes for that purpose.

In the ideal world, all electronic commerce between businesses would be utilizing one universal product classification system. But for at least two reasons, this does not look feasible in the real world. Firstly, because products come and go, and hence product classifications will always be under development. Secondly, because enterprises simply can not wait decades for a global standard to ‘arise’. Instead, many enterprises are already populating the Web in some sort of community, and as a consequence, various (sometimes overlapping) product classification systems are currently being developed and implemented. Sometimes initiated from chosen design principles, sometimes ad hoc driven by industry needs, sometimes based on industry conventions in the ‘old economy’, et cetera. The current state of the art has resulted in a rather awkward, user-unfriendly and difficult to maintain Product Dictionary that uses different -unrelated- indices and that contains overlapping sections without suitable cross-references.

Because this may heavily slow down the worldwide development and adoption of e-commerce solutions, it is a good time now to take a closer look at emerging product description standards and associated harmonization issues.

The Challenge: How to arrive at a common ontology between different classification systems?

We invite research groups to participate in a unique B2B E-Commerce Product Classification Challenge. Below we describe a concrete case study: the product item 'writing paper' as classified by two different e-commerce product classification systems. The task of the participating research groups in this Challenge is to design a model (ontology, problem-solving method, implemented working solution) to arrive at a computationally effective, practically useful, but also theoretically principled way of describing the relevant product knowledge and of establishing a (semi-automated) mapping between different e-commerce product classification systems.

We have several reasons to believe that this e-commerce exercise is attractive both from a research and from an industry practice point of view. Firstly, we believe that a small and concrete case example actually uncovers many strategic research issues that are involved. Research advances will yield an important and practically relevant input to the shaping of product classification standard systems in the near future. Secondly, the intelligent systems, knowledge engineering and ontology/Semantic Web research communities have of course already dealt with classification issues and methods in very diverse areas. Cultural communities, notably librarians and museum professionals, have also established long-standing good practices in this field. We are interested to find out in which respects this knowledge can contribute to the area of electronic commerce (the selected case is representative, seems to be simple, but only deceptively so). The third reason for this realistic, case-based contest is that we hope that it will help to bridge the (often wide) gap between industry and science.

An Abridged History of Product Dictionaries for B2B E-Commerce

Although product description systems already emerged in the 'old economy', the need for widely accepted, detailed, unique, and machine-understandable identifiers differs for each industry. Hence, these classification systems are often partly developed, are not suitable for e-commerce solutions, or lack the computational depth that an e-trading community requires. A strategy often used in the starting days of B2B hubs was to take the United Nations Standard Products and Services Code System (UNSPSC) as a starting point. The UNSPSC is a hierarchical classification with five levels (although in practice the 5^{th} level is hardly used). Each level contains a two-character numerical value and a textual description as follows:

XX Segment (The logical aggregation of families for analytical purposes)

XX Family (A commonly recognized group of inter-related commodity categories)
XX Class (A group of commodities sharing a common use or function)
XX Commodity (A group of substitutable products or services)
XX Business Function (The function performed by an organization in support of the commodity)

Major obstacles in using the UNSPSC are that it is rather shallow, not very intuitive, and not descriptive on an attribute level. A further disadvantage is that it is mainly developed in the US, leaving (for example) many European needs behind. In order to overcome these bottlenecks, three different strategies have emerged[1]:

1. Initiatives to enhance the UNSPSC with local attributes. Examples are the Universal Content Extended Classification, managed by the UCEC.org, or the Eccma Global Attribute Schema (EGAS), managed by the ECCMA. Both initiatives take the first four levels of the UNSPSC as a starting point. The EGAS schema is not yet published. The UCEC classification utilizes a standard set of attributes that can be distributed on every level, and are inherited at the commodity level.

2. Initiatives to develop industry specific extensions of the UNSPSC, such as the Rossettanet library (www.rosettanet.org) for the Information Technology and Electronic Components Industry, and recently announced initiatives for similar projects in the Chemical Industry (CIDX) and the Petroleum industry (PIDX).

3. Initiatives that build a new classification scheme from scratch, thereby replacing the UNSPCS. An example is Ecl@ss. Ecl@ss utilizes a four-level hierarchy. Each level contains a two-character numerical value. The last level is enriched with a standard set of attributes.

The Exercise: mapping a sample case of UNSPSC/UCEC and Ecl@ss

A crucial question for the growth of B2B e-commerce is whether a buyer, vendor or trading community that opts for an UNSPSC-related classification scheme (such as in the above options 1 or 2) can communicate with a vendor, buyer, or trading community that opted for an UNSPSC replacement or competitor.

Let' s make a simple comparison of the basic schemes of UCEC and Ecl@ss that at first sight look very much alike. Taking the concrete example of *writing paper*, we obtain the following two specific cases:

UCEC

[1] In addition, one can distinguish other types of product description and classification initiatives: proprietary systems (which will not be discussed because of the specific commercial interests involved), product description standards mainly focusing on manufacturing engineering such as STEP, and highly industry-specific classification systems such as EPISTLE for the process industry (which will be left aside because of the required industry expertise).

Scroll: Paper
Find: 14-11-15-11

14 Paper Materials and Products
11 Paper products
15 Printing and writing paper
11 Writing paper
0000000303 Type
0000000304 Length
0000000305 Width
0000000306 Weight
0000000307 Color
0000000308 Composition

Ecl@ss

Search: Papier
Find: 24-11-05-34 und Standardmerkmalleisten

24 Kommunikationstechnik, Bürotechnik
11 Büromaterial
05 Büromaterial (sonstiges)
34 Kanzleipapier, Schreibpapier
AAA474001 Alterungsbeständigkeit
AAA889001 EAN Code
DDA081001 Farbe
AAA001001 Hersteller
AAA252001 Hersteller-Artikelnummer
AAA457001 Papierformate
AAA458001 Papiergewicht
AAA003001 Produkt Name
AAA002001 Produkt Typ
AAA475001 Recyclinganteil
AAA215001 Zertifikate
AAA216001 Zulassung

Now we take a look at the classification problem of *writing paper*. We utilize the English translation that Ecl@ss offers, and for the moment leave the attributes out of the comparison. This leads to the following, non-exhaustive, mapping picture (See Figure 1).

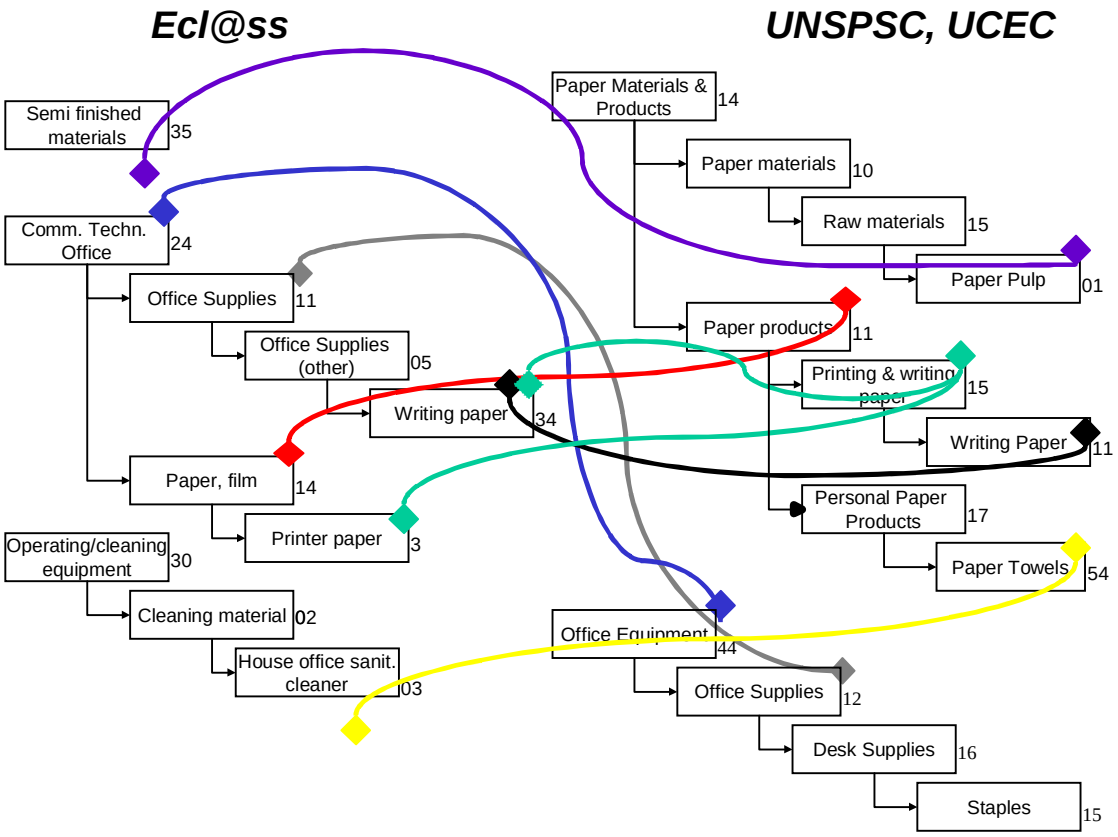

Figure 1: The mapping problem between two different product classification schemes for the case of writing paper.

Already from Figure 1, we see several interesting design and mapping issues:

- Some counter-intuitive categories appear, such as *office supplies (other)* as a subclass of *office supplies* in Ecl@ss, and *paper pulp* as a raw, rather than a semi-finished material in UNSPSC/UCEC.
- Even in this restricted case example, mappings over different parts of the classification are required. In UNSPSC/UCEC, all paper products and materials are organized in a single tree, whereas in Ecl@ss paper products are more functionally grouped. For example: *paper pulp* needs a connection (probably) under semi-finished materials in Ecl@ss, *paper towels* would (probably) fall under *operation/ cleaning equipment* in Ecl@ss.
- Some categories have either no or more than one, 'equivalent' class in different product classification schemes. For example, *printing & writing paper* constitutes a single category in UNSPSC/UCEC, but two separate categories in Ecl@ss.

These remarks only serve to emphasize the non-exemplary and preliminary character of these classification scheme mapping conclusions. However, it is quite clear that the Rosetta Stone that translates between the complete Ecl@ss and UNSPSC/UCEC

classification schemes will not be a pebble stone! And we note that this just represents a specific e-commerce case.

The Challenge Research Rules

We solicit research paper submissions that show how to solve this B2B e-commerce product classification challenge. Adequate solutions show not only how to solve this specific case, but they should be able to explain how to solve these multiple classification problems in e-commerce in a *generic* way fitting to the envisioned Semantic World-Wide Web. This Challenge is sponsored by OntoWeb, the EU Thematic Network on Ontology-Based Information Exchange for Knowledge Management ands Electronic Commerce (the co-authors are members of OntoWeb, particularly of its Special Interest Groups on Industry Applications and Content Standardization).

The task set for this research challenge is:

> **To design a *generic* model (ontology, method, sample implementation, experimental computational results) to arrive at an (automated or semi-automated) mapping between Ecl@ss and UCEC as two *sample* e-commerce product classification standards.**

Needed data can be taken from the following sources: Ecl@ss at www.eclass.de, the first four levels of the UNSPSC that UCEC utilizes at www.eccma.org/unspsc/. Further relevant information can be found for example in recent issues of IEEE Intelligent Systems, several conferences and workshops (SWWS, K-CAP, IJCAI), and www.daml.org and www.ontoweb.org.

Submissions will be evaluated by a committee formed by the two above-mentioned OntoWeb SIGs, according to the following criteria:

1. The proposed model should aim at a working solution that is both conceptually and computationally adequate for the case (UCEC and Ecl@ss schemes; sample products such as writing paper). Epistemological adequacy, complexity, as well as achieved precision of the proposed mapping and classification approach are relevant issues.
2. The proposed solution should have a *generic* value by showing how to handle such e-commerce product classification situations *in general*. A related issue is to consider through what mechanisms once achieved mappings can be effectively maintained.
3. If so desired, it may offer recommendations and/or requirements on how future e-product classification standards should look like.
4. Also, it may offer recommendations and/or requirements on Semantic Web languages and standards (e.g. RDF(S), DAML+OIL).
5. Any upfront assumptions to be made for a working solution as well as lessons learned should be made explicit.

Submissions (in PDF format, max. 20 pages) can be sent to the OntoWeb secretariat, c/o Ms. Elly Lammers at elly@cs.vu.nl on a continuous basis up to January 31, 2002. OntoWeb will publish all submissions on its website (www.ontoweb.org) and stimulate their public discussion and evaluation by means of a moderated mailing list. At the end of

this discussion period, the best proposals will be selected for presentation as part of upcoming OntoWeb and/or Semantic Web Workshops that are currently planned for Summer 2002 in a beautiful and sunny place. Sponsorships for a small contest prize are being investigated. We will also make arrangements for appropriate archival scientific publication after the workshop, and help make industrial contacts and connections for interesting solutions that are proposed.