Product Data Integration in B2B E-Commerce

Dieter Fensel, Ying Ding, and Borys Omelayenko, Vrije Universiteit Amsterdam
Ellen Schulten, ***Heloise Ontology Associates***
Guy Botquin, ***Content Europe***
Mike Brown and Alan Flett, ***SemanticEdge***

To overcome current bottlenecks in B2B E-commerce, we need intelligent solutions for mechanizing the process of structuring, standardizing, aligning, and personalizing data. This article surveys the overall content-management process and discusses the requirements for scalable support for it.

The dramatically increased flexibility afforded by the Internet in business-to-business transactions also presents steep challenges in merging information coming from so many sources. B2B marketplaces, which function as an intermediate communications layer, reduce the number of mappings needed for their user community from $n*m$ to $n + m$ (see Figure 1). However, to provide this service, they must deal with the problem of heterogeneity in their customers' product, catalog, and document descriptions. Effectively and efficiently managing different description styles becomes a key task for these marketplaces. In real-world marketplaces, developing a scalable approach for information integration has become the main prerequisite for scaling businesses.

Successful content management for B2B electronic commerce must deal with several challenges: extracting information from rough sources; classifying information to make product data maintainable and accessible; reclassifying product data; personalizing information; and creating mappings between different information presentations.

The lack of standards—really, the inflation and inconsistency of newly arising pseudostandards—makes all these subtasks more difficult. As a benefit to both academics and industrialists who want to provide solutions for this key process in B2B electronic commerce, this article focuses on these challenges for content management and discusses potential solution paths.

Information integration in B2B e-commerce

A successful marketplace must integrate various hardware and software platforms and provide a common protocol for information exchange. However, the real problem is the exchanged content's heterogeneity and openness. This heterogeneity arises in at least three levels: the content, product catalog structure, and document structure.

The content of the exchanged information must be modeled. Historically, many different ways to categorize and describe products have evolved. Often, vendors have their own private way to describe their products. Structuring and standardizing the product descriptions is a significant task in B2B e-commerce, ensuring that the different actors can communicate with each other so that their customers can find the products they want. Here, content-management solution providers can offer added value by helping their vendors build and instantiate an ontology for certain product domains.

E-commerce is about electronically exchanging business information—of which product descriptions are just one element. The product descriptions are an electronic catalog's building blocks, together with information about the vendor, the manufacturer, the lead time required, and numerous other business-related considerations. Furthermore, a catalog provider should include quality control information, such as catalog version, date, and identification number. The total composition of these

1094-7167/01/$10.00 © 2001 IEEE

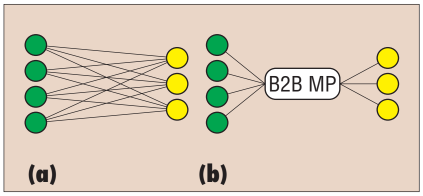


Figure 1. Preventing exponential growth in the number of mappings: (a) Direct communications between *m* suppliers and n customers = *n* * *m* mappings; (b) Mediated communications between *m* suppliers and *n* customers = *n* + *m* mappings.

building blocks is called the *catalog structure*. Where there are two electronic catalogs involved—for example, when two vendors in one marketplace have different catalog providers, or when two different marketplaces want to communicate—the structure of these catalogs must be aligned as well.

Going one step further into the content-management process, we come upon the catalog's actual use. In a marketplace, a buyer will want to send a purchase order after picking up the necessary information from the catalog. The vendor must reply with a confirmation, which starts the buying process. For the buyer and the vendor to read and process each other's business documents again requires a common language. Marketplace software developers such as Commerce One, which developed its structures based on the xCBL language, provide a large collection of document structures that reflect different aspects of a trading process. Aligning these document structures with other document definitions such as those from Ariba (cXML) is far from a trivial task.

Consequently, three types of mismatches can arise. The first type, in content, mainly concerns the real-world semantics of the exchanged information: people describe the same products in different ways. The second and third types, in product catalogs and business documents, more generally concern the exchanged information's syntactical structure.

The overall information integration process must tackle a number of serious problems. In particular, product descriptions must be structured; classified; (re-)classified and (re-)described in various dimensions because no standard product classifications exists; and personalized to let customers find the products they seek. Figure 2 provides a snapshot of the overall process.

Structured product descriptions

Suppliers have product catalogs that describe their products to potential clients. They want to make this information available online through a B2B marketplace. Because so many product catalogs already exist electronically, you might think this would be a simple task. However, these product catalogs are designed for the human reader. Thus, extracting the actual product information, classifying it, normalizing it, and storing it in a structured format is primarily a series of manual tasks. Figure 3 illustrates the task for converting unstructured product descriptions to structured and classified documents. Content-management solution providers often have several hundred employees working in content factories to manually perform this information processing, often starting with only printed catalogs.

The most critical element to converting descriptions, particularly with the view to automating some of the stages, is transforming informal text into a formal—that is, machine readable—format. This process generally has two main subtasks: defining the product categories and their attributes (defining the schema), and extracting the actual values for the defined attributes. Mechanizing the data-extraction step introduces a third subtask: the manual or automatic (machine-learning) derivation of extraction rules or patterns.

Information extraction includes various techniques applied to automatically extract specified information from short newswires, natural language texts, and full or free-text documents. In a nutshell, we can regard

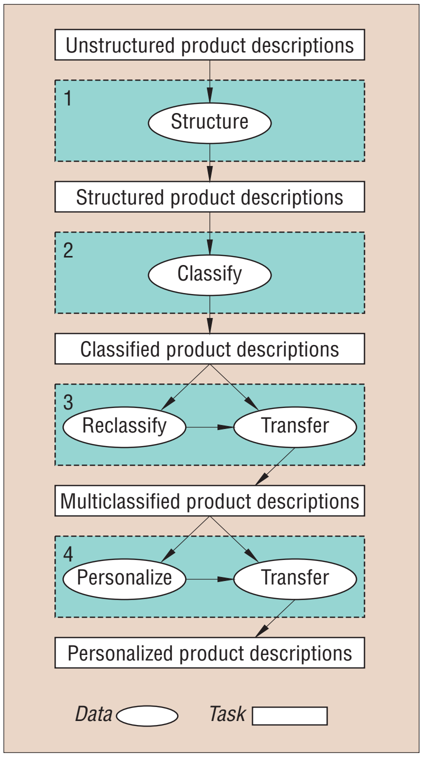
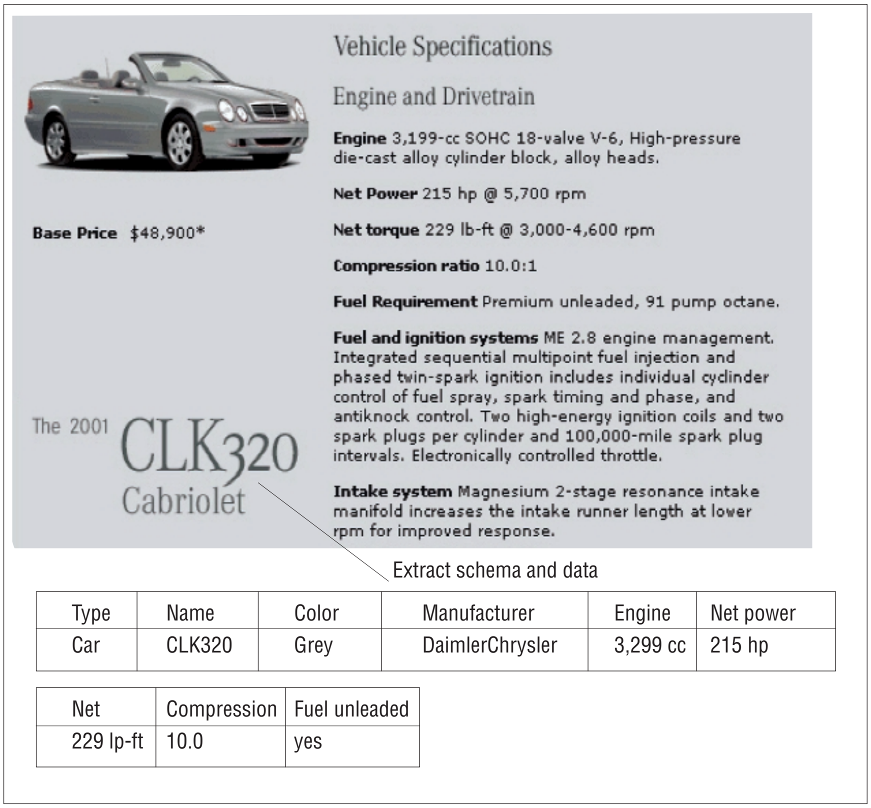

Unstructured product descriptions
1
Structure
Structured product descriptions
2
Classify
Classified product descriptions
3
Reclassify
Transfer
Multiclassified product descriptions
4
Personalize
Transfer
Personalized product descriptions
Data
Task

Figure 2. The main subtasks in content management for B2B e-commerce.


Vehicle Specifications

Engine and Drivetrain

Engine 3,199-cc SOHC 18-valve V-6, High-pressure die-cast alloy cylinder block, alloy heads.

Net Power 215 hp @ 5,700 rpm

Net torque 229 lb-ft @ 3,000-4,600 rpm

Compression ratio 10.0:1

Fuel Requirement Premium unleaded, 91 pump octane.

Fuel and ignition systems ME 2.8 engine management. Integrated sequential multipoint fuel injection and phased twin-spark ignition includes individual cylinder control of fuel spray, spark timing and phase, and antiknock control. Two high-energy ignition coils and two spark plugs per cylinder and 100,000-mile spark plug intervals. Electronically controlled throttle.

Intake system Magnesium 2-stage resonance intake manifold increases the intake runner length at lower rpm for improved response.

Base Price $48,900*

The 2001 CLK320 Cabriolet

Extract schema and data

Type	Name	Color	Manufacturer	Engine	Net power
Car	CLK320	Grey	DaimlerChrysler	3,299 cc	215 hp

Net	Compression	Fuel unleaded
229 lp-ft	10.0	yes

Figure 3. Information extraction.

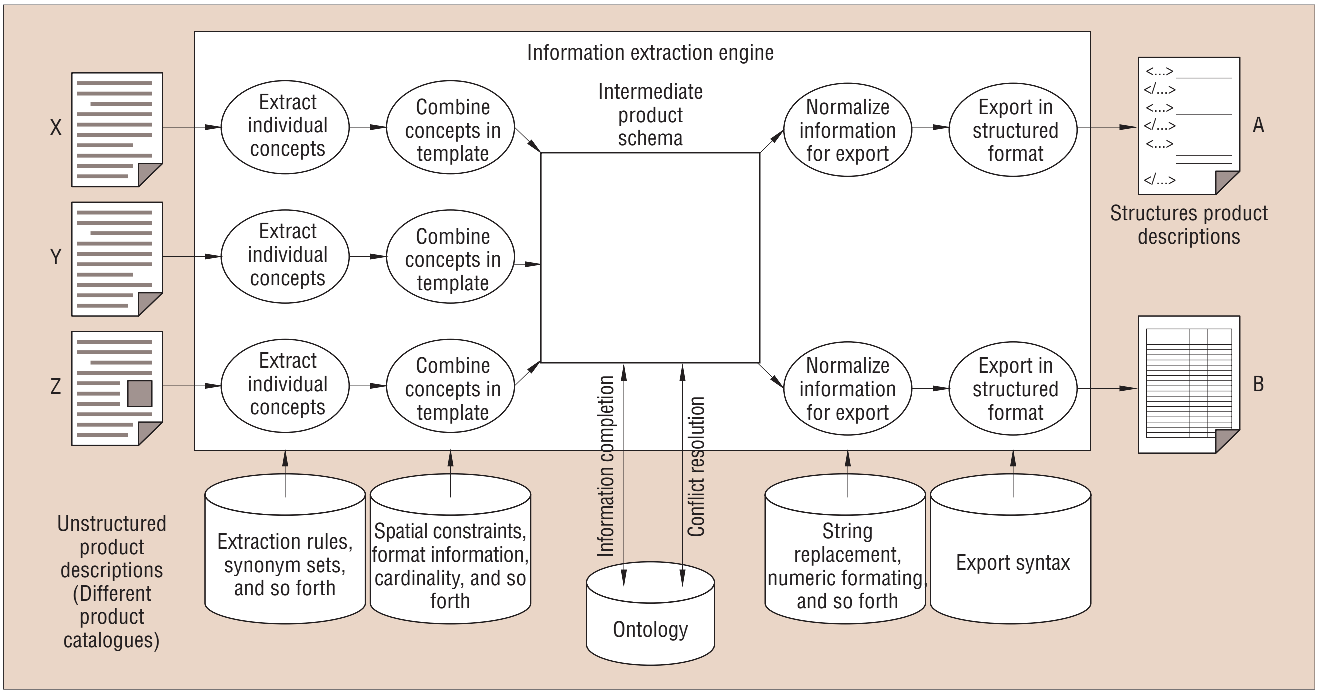

Information extraction engine
X
Y
Z
Extract individual concepts
Extract individual concepts
Extract individual concepts
Combine concepts in template
Combine concepts in template
Combine concepts in template
Intermediate product schema
Normalize information for export
Normalize information for export
Export in structured format
Export in structured format
<...>
</...>
<...>
</...>
<...>
</...>
A
Structures product descriptions
B
Unstructured product descriptions (Different product catalogues)
Extraction rules, synonym sets, and so forth
Spatial constraints, format information, cardinality, and so forth
Information completion
Conflict resolution
Ontology
String replacement, numeric formating and so forth
Export syntax

Figure 4. A sketch of the information-extraction engine.

information extraction as an activity of populating a structured information source (such as database or worksheet) from an unstructured information sources. However, many significant problems for this mechanization arise, such as an inadequate representation format of the information and high irregularities in the information's layout. Each product catalog differs substantially in presentation style, while multimedia information is difficult to extract automatically (for example, extracting a product's color from its picture in the catalog).

SemanticEdge is among a growing number of companies that offer specialized technology for executing this information-extraction task. Figure 4 shows an overview of the solution this company offers. A single information extraction engine encapsulates several trainable and self-learning AI technologies. Users can configure these AI technologies to map different product catalog formats onto a single intermediate, predefined product schema. From this schema, they can export information into one or more formalized representations, which later stages of the content-management process can further process.

The approach detailed in Figure 4 lets us unify information from multiple sources (and hence formats) of unstructured product descriptions into a single format. Hence, we can export the information in any of the product catalogs *X*, *Y*, or *Z* in the single export format *A* or *B*. This format unification is also crucial for the remaining stages of the content-management process.

The combination of multiple AI techniques can deliver extremely accurate information-extraction performance. Nevertheless, two fundamental problems persist: incompleteness—that is, only information stated in the unstructured product descriptions can be extracted, and false values—inevitably some small degree of inaccuracy in the extracted information will exist, so some false values for product features might be extracted.

Domain-specific ontologies can play an important role in significantly reducing these problems. They help identify likely causes for choosing between different options and help users infer additional knowledge that the data source doesn't explicitly provided.

Classified product descriptions

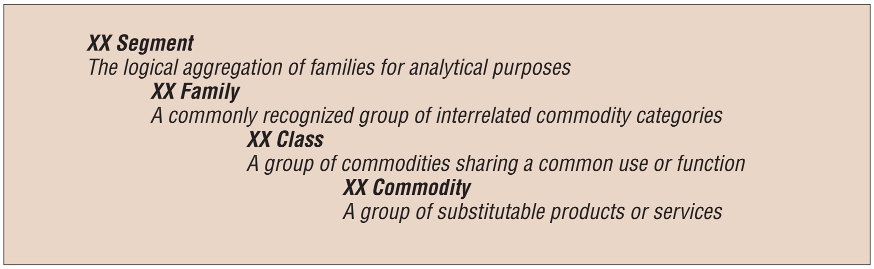
At this stage of the content-management process, we can assume that our product information is structured in a tabular way. Each supplier might use different structures and vocabularies to describe its products, but that might not cause a problem for a one-to-one relationship where the buyer could well get used to the supplier's private terminology. B2B marketplaces that enable *n*-*m* commerce cannot rely on such an assumption. They must classify all products according to a standard classification schema that helps buyers and suppliers communicate their product information. The Universal Standard Products and Services Classification is a widely used classification schema in the US.

UNSPSC was created when the United Nations Development Program and Dun & Bradstreet merged their separate commodity classification codes into a single open system. Currently maintained by the Electronic Commerce Code Management Association (http://eccma.org), a not-for-profit membership organization, the UNSPSC is a hierarchical classification with four levels: segment, family, class, and commodity. Each level contains a two-character numerical value and a textual description (see Figure 5).

Again, classifying the products according to a classification schema like UNSPSC is a difficult and largely manual task. It requires domain expertise and knowledge about the product domain, which makes the process costly. High quality is important for ensuring maintainability and visibility of product information.

Support in mechanizing this process is important for content management in B2B e-commerce. ProCat, a software environment for automated product cataloging, offers such tool support. Developed at Vrije Univeriteit

Figure 5. Example of the UNSPSC hierarchical classification schema.


XX Segment
The logical aggregation of families for analytical purposes
XX Family
A commonly recognized group of interrelated commodity categories
XX Class
A group of commodities sharing a common use or function
XX Commodity
A group of substitutable products or services

Amsterdam, it automatically catalogs product descriptions by providing and optimizing various cataloging methods from information retrieval (especially from the text-classification area) and machine learning. The current version applies the information-retrieval metaphor to the product classification task. It views a product description as a query and the classification schema as a document collection; the retrieved classification code corresponds to the retrieved answer document. This metaphor significantly improves the overall productivity in product classification. Future versions will add such features as multistandard classification (such as UNSPSC, UCEC, and ecl@ss), and multilinguality (the product catalog and the product classification standard are described in different languages).

Current case studies show serious problems in achieving high accuracy. Typical problems are the heterogeneity in the vocabulary used in product catalogs and the limited number of words used to describe categories in product classification standards. Existing ontologies such as WordNet are not terribly useful because they lack most of the product specific terms needed in the product classification task.

Reclassified and redescribed product descriptions

Bottlenecks in exchanging information have led to a plethora of standards designed to improve the situation. However, two problems usually arise: there are too many "standards," none of which is an actual standard, and most standards lack important features for various application problems. Not surprisingly, both problems also appear in B2B e-commerce. Only UNSPSC lacks important features for various content management aspects. It is

- undescriptive—it does not define any attributes for describing the products,
- unintuitive—neither suppliers nor buyers find their products easily, and
- shallow—it does not provide enough distinctions for a vertical marketplace that provides numerous products from a certain domain.

UNSPSC is a typical example for a horizontal standard that covers all possible product domains but is not very detailed in any particular domain. Another similar example is the Universal Content Extended Classification, which takes UNSPSC as a starting point and refines it by attributes. RosettaNet is an example of a vertical standard that describes computer hardware and software products in detail. Vertical standards describe a certain product domain in more detail than common horizontal ones.[1]

Because different customers use different classification schemas, the product information must be classified and described according to several schemas. This objective defines three subtasks for successful content management:

- Define links between different classification schemas that relate the various concepts and attributes. Establishing such a connections helps to classify new products in an additional classification schema.
- Reclassify a product. Because there does not need to be a one-to-one correspondence between concepts in different classification schemas, we often require the actual product information to decide about its new classification.
- Transfer the original descriptions into the new description style.

Each subtask is far from being trivial. Take UNSPSC and ecl@ss as examples. Especially in Europe, where UNSPSC is less broadly used, product classification systems arose that did not take UNSPSC as a starting point. The ecl@ss product standard initiative, for example, began in 1997 as a cooperation between leading German industries and the Cologne Institute for Business Research. Ecl@ss features a four-level, hierarchical classification key similar to UNSPSC, with a keyword index containing 14,000 terms. In addition, ecl@ss provides attributes at many levels of the hierarchy, which are inherited top-down in the classification hierarchy. With these attributes, ecl@ss provides a strong alternative solution to the nondescriptiveness of UNSPSC. However, it is a very young standard, mainly used in Germany. The ecl@ss classification scheme broadly resembles UNSPSC, but the population of its structure is quite dissimilar: ecl@ss proposes a more intuitive hierarchy from an end–user's point of view, but the manufacturers' perspective is leading the classification in UNSPSC

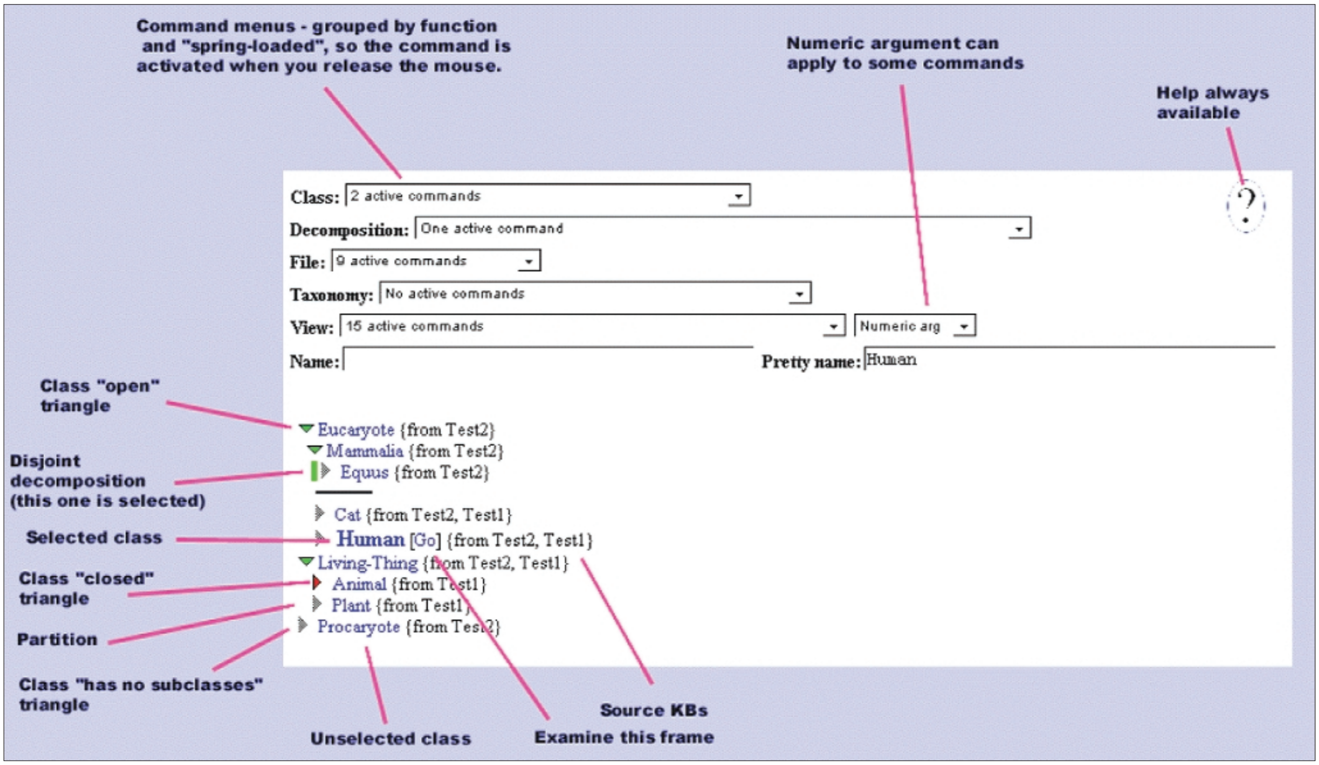
Integrating such descriptions benefits from ontology integration work.[2] Tool support for these tasks are offered by the knowledge engineering community with tools such as Chimaera and Prompt.[3,4] Ontologies provide large taxonomies of concepts enriched by attributes and axioms describing their logical properties. Operations for combining ontologies are inclusion, restriction, and polymorphic refinement. Tools such as Chimaera provide support in merging multiple ontologies and diagnosing individual or multiple ontologies. Chimaera supports such tasks as using ontologies in differing formats, reorganizing taxonomies, resolving name conflicts, browsing ontologies, and editing terms (see Figure 6).

Viewed at an implementational level, the standards are moving towards XML-based representations and require the connection with low-level integration architectures provided by the W3C consortium with XSLT and XPath. Direct catalog mapping with XSLT rules, which appeared to be a partial solution, has raised numerous problems.[5] A direct-mapping approach, even for a simple concept like an address, leads to complex and unmaintainable rules. Various complex XSL-T rules are needed to implement mappings between different styles in describing an address.

To overcome these problems, two of us describe a stepwise approach that assumes that the integration is performed in three intermediate steps and the introduction of an intermediate data model.[6] By clearly separating the aspects of extracting semantics from syntax, mapping at a semantic level, and adding syntactical details to a representation, that approach can achieve simple and reusable transformation patterns. Instead of writing an ad hoc complex XSL-T transformation, the user can select and combine simple and standardized transformation rules.

Personalized product descriptions

Personalization resembles reclassification in part. Different buyers or sellers might want to have different views on the product classification. The content-management process requires semiautomatic support in generating views on product catalogs based on user descriptions. For example, a secretary might


Command menus - grouped by function and "spring-loaded", so the command is activated when you release the mouse.
Numeric argument can apply to some commands
Help always available
Class "open" triangle
Disjoint decomposition (this one is selected)
Selected class
Class "closed" triangle
Partition
Class "has no subclasses" triangle
Unselected class
Source KBs
Examine this frame
?

Class: 2 active commands

Decomposition: One active command

File: 9 active commands

Taxonomy: No active commands

View: 15 active commands

Numeric arg

Name:

Pretty name: Human

```
Eucaryote {from Test2}
 Mammalia {from Test2}
  Equus {from Test2}

  Cat {from Test2, Test1}
  Human [Go] {from Test2, Test1}
Living-Thing {from Test2, Test1}
  Animal {from Test1}
  Plant {from Test1}
Procaryote {from Test2}
```

Figure 6. Chimaera.

not want to see the full-fledged product catalog but only the goods that are relevant to his office environment. He might also want to find products classified according to their various business needs or business processes and not according to UNSPSC where he will never find his eraser. Briefly, this task is about defining a view on information based on different users' role descriptions.

Although personalization and view generation have been studied in the information-retrieval and database communities,[7,8] these techniques must be adapted to the specific needs of product catalog adoption. These approaches are static in that the type of a set of users is known in advance and the data views change accordingly at the start of the search process. While this is an improvement, there is no interactive help, in the sense of computer-side suggestions and guidance, in searching the product space. A more sophisticated approach requires a system that will interact with the user in a more pertinent manner, such as for the kind of product being sought (user-profiling), the manner in which the conversation should be phrased (phraseology), and the strategy for searching the product space and optimally meeting the user's requirements (negotiation strategy).

To support this human-oriented, conversational style, we must apply several technologies to the problem. We need to develop both a subjective and an objective information (or product) ontology. The objective ontology models the kind of descriptive elements found in the typical product catalog or database for the information space in question, such as nominal physical attributes of products—the weight of a printer, for example. The subjective ontology models the kind of descriptive elements that human customers typically use when conceptualizing the product space of interest—the quality of a printer's output resolution, for example. We must also develop classification rules to classify the various products as belonging to certain subjective categories—a cheap family printer, for instance. Ontologies then also serve to model the negotiability of each product feature. Dialog models have been built to interact with the user in a manner that the user finds comfortable.

The number of subtasks we discussed do not provide the complete picture. There are at least two more important subtasks we have not yet discussed: enrichment of product descriptions and of product standards. Both are largely complementary. In the first case, a structured product description turns out to be incomplete or nonstandard according to the standard set of attributes that the classification schema is assuming. Then a loop back in the information extraction step is necessary to acquire additional product information. This commonly appearing process requires significant content-management resources. In the second case, a product standard turns out to be incomplete or unsuitable for describing the products appropriately. In this case, life for the content manager begins to get hard. She needs to play an active role in complex standardization bodies in trying to overcome many of the obvious bottlenecks of current B2B standards.

In this work, many concepts and techniques developed in the intelligent information integration and related areas can help significantly.[9–12] However, most of them must be adapted to the specific needs of e-commerce. We must especially ask such approaches whether they can scale up to large volumes of information. ■

References

1. D. Fensel, *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*, Springer-Verlag, Berlin, 2001.

2. H. Sofia Pinto, A. Gomez-Perez, and J. Martins, "Some Issues on Ontology Integration," *Proc. IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5)*, AAAI Press, Menlo Park, Calif., 1999, pp. 7-1–7-12.

3. D. McGuinness et al., "An Environment for Merging and Testing Large Ontologies," *Proc. Seventh Int'l Conf. Principles of Knowledge Representation and Reasoning (KR2000),* Morgan Kaufmann, San Francisco, 2000.

4. N. Noy and M. Musen, "PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment," *Proc. 17th Nat'l Conf. Artificial Intelligence (AAAI-00),* AAAI Press, Menlo Park, Calif., 2000.

5. B. Omelayenko and D. Fensel, "An Analysis of the Integration Problems of XML-Based Catalogs for B2B Electronic Commerce," *Proc. Ninth IFIP 2.6 Working Conference on Database Semantics*, Chapman Hall, Boca Raton, Fla., 2001, pp. 232–246.

6. B. Omelayenko and D. Fensel, "A Two-Layered Integration Approach for Product Information in B2B E-commerce," *Proc. Second Int'l Conf. Electronic Commerce and Web Technologies (EC-WEB 2001)*, 2001, to appear.

7. U. Srinivasan, A.H.H. Ngu, and T. Gedeon, "Managing Heterogeneous Information Systems through Discovery and Retrieval of Generic Concepts," *J. Am. Soc. Information Science*, vol. 51, no. 8, 2000, pp. 707–723.

8. S. Abiteboul et al. "XML Repository and Active Views Demonstration," *Proc. 25th Int'l Conf. Very Large Data Bases (VLDB-99)*, Morgan Kaufmann, San Francisco, 1999, p. 742–745.

9. H. Wache and D. Fensel, "Intelligent Integraton of Information," *Int'l J. Cooperative Information Systems on Intelligent Information Integration*, vol. 9, no. 4, 2000, pp. 257–360.

10. D. Fensel et al., "OnToKnowledge: Ontology-based Tools for Knowledge Management," *Proc. eBusiness and eWork 2000 (EMMSEC 2000) Conf.*, Cheshire Hendbury, Macclesfield, UK, 2000.

11. D. Fensel et al., *J. Data and Knowledge Engineering (DKE) on Intelligent Information Integration*, vol. 36, no. 3, 2001, to appear.

12. S. Navathe, "A Model to Support E-Catalog Integration," *Proc. IFIP 2.6 Working Conf. Database Semantics*, Chapman Hall, Boca Raton, Fla., 2001.

The Authors

Dieter Fensel is an associate professor at the Division of Mathematics and Computer Science, Vrije Universiteit, Amsterdam. After studying mathematics, sociology, and computer science in Berlin, he joined the Institute AIFB at the University of Karlsruhe. Currently, his focus is on using ontologies to mediate access to heterogeneous knowledge sources and to apply them in knowledge management and e-commerce. Contact him at the Division of Mathematics and Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, Netherlands; dieter@cs.vu.nl; www.cs.vu.nl/~dieter.

Ying Ding is a senior researcher in Division of Mathematics and Computer Science, Free University, Amsterdam. She completed her PhD at the School of Computer Engineering, Nanyang Technological University, Singapore. Her research interests include information retrieval, ontology learning and knowledge management. She is currently involved in three running European Union projects: OntoWeb, OntoKnowledge, and IBROW. Contact her at the Division of Mathematics & Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, the Netherlands; ying@cs.vu.nl, http://www.cs.vu.nl/~ying.

Borys Omelayenko, is a PhD student in the Business Informatics Group, Vrije Universiteit Amsterdam. He received his MSc in software development from Kharkov Technical University of Radioelectronics, Ukraine. Currently, he is working on his PhD thesis, targeted at development and application of Semantic Web technologies to resolve various information integration problems of business-to-business e-commerce. Contact him at the Dept. of Mathematics and Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081hv, The Netherlands; borys@cs.vu.nl; www.cs.vu.nl/~borys.

Guy Botquin is Vice President and Chief Technology Officer of Content Europe. After graduating from UCL as a software engineer, he worked as a reseacher at the Louvain Polytechnics Faculty on problematics of machine learning and automatic classification. He is a functional expert in retail modernization. Contact him at gbotquin@alexsys.be; www.contenteurope.com.

Ellen Schulten is an independent consultant in B2B e-commerce. She has a background in change management consultancy at E&Y in the Netherlands and Cap Gemini PBS in the US. She is also active in several e-commerce start-ups and is cofounder of Heloise Ontologies Associates (www.heloisenet.com), a network of ontology experts from leading universities that offers a senior business consultancy on ontologies. She has degrees in philosophy from the University of Utrecht and educational sciences from the Technical University of Enschede, the Netherlands. Contact her at ellen@heloisenet.com.

Mike Brown is Vice President of Knowledge Technology at SemanticEdge, GmbH. He holds a BSc in physics and computer science and a PhD in artificial intelligence from the University of Manchester. He has been involved in developing a wide range of industrial applications of AI and machine learning technology. The Knowledge department of SemanticEdge, which he heads, covers activities including knowledge management and querying, ontology representation and construction and information extraction and machine learning. Contact him at: SemanticEdge, Kaiserin-Augusta-Allee 10-11, 10553 Berlin; michael.brown@semanticedge.com; www.semanticedge.com.

Alan Flettis Head of Ontology Technology at SemanticEdge, GmbH. He holds a Beng in electronic and electrical engineering from the University of Strathclyde and an MSc in applied artificial intelligence from the University of Aberdeen. His main responsibilities at SemanticEdge are providing ontology solutions, including both the development of domain ontologies and the development of the supporting technology. Contact him at SemanticEdge, Kaiserin-Augusta-Allee 10-11, 10553 Berlin; alan.flett@semanticedge.com; www.semanticedge.com.