# B2B AUTOMATIC TAXONOMY CONSTRUCTION

Ivan Bedini
*Orang Labs, Caen, France*

Benjamin Nguyen, Georges Gardarin
*PRiSM Laboratory, University of Versailles, Versailles, France*

Keywords: Semantic Web, Ontology Learning, B2B, XML Mining.

Abstract: The B2B domain has already been subject to several research experiences, but we believe that the real advantage of introducing semantic technologies within enterprise application integration has not yet been investigated fully. In this paper we provide a new use case for the next generation Semantic Web applications with regards to enterprise application integration. We also present the results of our experience in automatically generating a taxonomy from numerous B2B standards, constructed using *Janus*, a software tool we have developed in order to extract semantic information from XML Schema corpora. The main contribution of this paper is the presentation of the results of our tool.

## 1 INTRODUCTION

One of the most frequently asked questions during exchanges with other colleagues is surely: "*Why introduce* ontologies in the area of enterprise applications integration and interoperability? *What is their contribution* and *what are the new benefits* compared to existing technologies?"

While current solutions work and enterprises are able to exchange electronic information between each other, several experiences show it is practically impossible to connect two or more enterprise applications on the fly. Even when two businesses use standards claiming conformance to the same base and same type of messages, business integration remains difficult.

An example of this is shown by (Anicic, 2005), where authors argue that the integration of two applications, one based on the Standards in Automotive Retail (STAR) and the second on the Automotive Industry Action Group (AIAG), where both of their native interfaces are based on the Open Application Group (OAG) standard, requires the construction of a supplementary external module to connect them. Many other similar examples exist, and form the motivation of this work.

The predominant view of application integration is that it should be completely performed at *design time*, i.e. when deciding on integration rules between applications, rather than being performed at *run time*, i.e. during the business exchange execution.

In this context, advantages of a Semantic Web (SW) based approach for enterprise applications integration has been widely recognised (Perez 1999), (Fensel, 2001a), (Zhao, 2003a). But as clearly presented by (Sabou, 2006) and (Motta, 2006), the problem of the definition of a reference knowledge as base to improve the ontology mapping still remains.

In this context we consider this problem as equivalent to enterprise applications integration.

The aim of this paper is not to resolve the whole problem of business application integration, but to analyse the problem and to present a solution to the reference knowledge generation, starting from existing XML B2B documents.

In Section 2 we present an analysis of the B2B use case in the context of the Semantic Web and show current approaches to business exchanges. Section 3 presents Janus, the tool that we have developed in order to retrieve semantic information from existing XML Schema files and some results obtained by the application of Janus on a collection of 23 B2B XML based standards freely available on the Web. In Section 4, we discuss related works and Section 5 is a conclusion



Bedini I., Nguyen B. and Gardarin G. (2008).
B2B AUTOMATIC TAXONOMY CONSTRUCTION.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - ISAS*, pages 325-330
DOI: 10.5220/0001702903250330

# 2 THE B2B USE CASE

In this section, we present a generic B2B use-case, and advocate the use of ontologies to solve integration problems.

## 2.1 Why we need Semantics?

The book by Gregor Hohpe (Hohpe, 2003) clearly shows that there are many problems with application integration. He provides an exhaustive list composed of 65 patterns to be considered when building a system able to manage the whole process of application integration. In this paper we do not address the whole process of integration, but we focus on the content of messages exchanged between enterprises in B2B applications.

B2B provides an interesting use case for semantic applications because, by its nature, it focuses on the problem of different designs and ways of structuring the same set of concepts. Yet no existing approach implements techniques based on semantics. Currently, applications exchange information on the basis of passing parameters or data, formatted in XML according to strict, pre-defined syntaxes and semantics. We define this approach as the **exactness method**. This method has the advantage of allowing good error management, but leaves no space for data interpretation. In consequence, reasoning on data of this type for integration is virtually impossible, because of the rigidity of data definitions.

## 2.2 Business Exchange Approaches

As far as we know, current approaches to message content definition for electronic business exchanges are mainly based on three types of solutions, which are: **Ad-hoc solution,** where the format is defined multilaterally during the design time phase of the application; **Proprietary solution** where the format is decided unilaterally (e.g., by a main contractor in cooperation with small businesses, such as a big retail group and its suppliers) and; **Adoption of standards** where the format is defined by a consortium in some standardization body.

As shown in the European e-business report (E-Business W@tch, 2007), at least three enterprises out of four that realize business exchanges with partners, declare implementing applications based on B2B standards solutions (at least in Europe). Moreover, the authors of this report also state that the broad adoption of XML based standards in combination with web services, could become the key to shape electronic business transactions between enterprises in the future.

## 2.3 The Canonical Data Model

Gregor Hohpe (Hohpe, 2003) suggests building a Canonical data model in order to minimize dependencies from different data formats, but he does not explain *how* to build it. We suggest adopting an ontology based approach when building the canonical data model and using semantic web technologies to improve application integration. This approach is quite different from other experiences in the e-business domain, such as (Corcho, 2001), because it targets global message definition rather than a thesaurus like eCl@ss or UNSPSC. A message is not a well defined hierarchical set of items, but meets a specific request. This practice makes complex the matching of two messages, and therefore application integration, because standards can model them with different pieces of information.

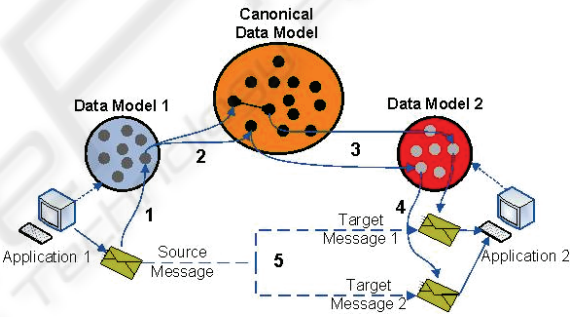


Figure 1: Messages translation procedure.

In other words, we are not able to say beforehand if the sending application ship messages that correspond exactly to the receiver application messages, in a one-to-one association. However, we make the hypothesis that the sender application manages some “concepts” that are similar to those of the receiver application. We propose a procedure to correlate these messages (see Figure 1), based on the following steps: 1) detect what concepts the message conveys; 2) match them with the canonical model; 3) find corresponding concepts in the target application model; 4) chose the message mappings that best fit the requirement and finally; 5) translate the message. As we can see, the main problem is building the canonical model. The difficulty is that the classical development of a domain ontology, typically entirely based on strong human participation, does not adequately fit this use case, because it needs a more dynamic and automatic ontology building system, in order to be able to

integrate new business partners on the fly.

# 3 AUTOMATIC CONSTRUCTION OF THE TAXONOMY

In this section we present Janus, a tool we have developed that manages information extraction from XML schema files. We also present the firsts results obtained from the automatic construction of a B2B taxonomy.

## 3.1 B2B Corpus Source

For this experience we have investigated more than 30 B2B standards, but not all are freely available and require membership fees (these have not been studied during the tests presented here).

Only cXML does not provide an XML Schema (XSD) for messaging content definition, but instead provides a DTD based standard. Moreover no RDF/OWL format is officially provided by any consortium. For this reason we decided to initially consider only standards offering XSD files and to focus our efforts on developing a tool of information retrieval specifically for this format. XML Schema simplifies the defininition of a structure for elements (candidate concepts for the ontology) notably limiting the difficulties of natural language interpretation. However as we show below, these documents introduce some noise at semantic level that needs special attention in order to provide good quality results. Almost all organizations provide a package containing several XSD files, one for each specific message, one for grouping common data, others for grouping common data type definitions and code lists. In the end we get a corpus source composed of a collection of 23 standards (listed in Table 1), with more than 2000 XSD files that has been considered enough in order to have significant information about B2B business message definition practices and semantics. Others standards can be added in future in an incremental way.

## 3.2 Janus: Taxonomy Builder Tool

Our tool implements an adaptation of several techniques originating from the text mining and information retrieval/extraction fields, applied to XML files (that we call **XML Mining**), in order to pre-process simple and compound terms from XML tags. In reality our tool goes further in trying to build a reference ontology, making the hypothesis that each standard's set of files provides enough information to be considered an ontology itself.

Figure 2 shows the overall architecture of Janus.

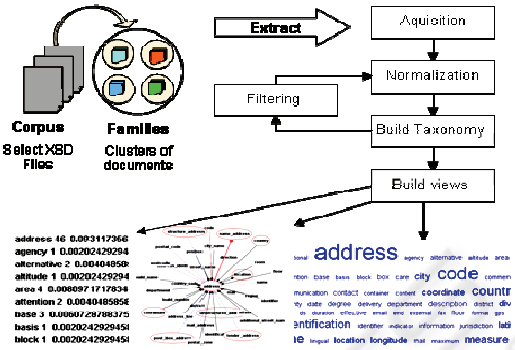


Figure 2: Janus overall architecture.

Let us now detail the algorithm for term extraction and automatic taxonomy construction from XML tags:

**Acquisition Step.** The aim of this step is to organize the corpus source and to select useful terms for the taxonomy. The extraction tasks are: XSD *Parsing* and extraction of XML tag values; *composite words* detection (e.g.: on-line); *detection of "useless" terms* previously identified, like systematic addition of unrelated semantic sense to the tag (e.g.: *CommonData* for *UnitOfMeasureCommonData*); *Splitting* compound terms forming the tag, (e.g.: *PersonIDCode* = person + id + code) and; *abbreviation transformation* (e.g.: Addr = Address, PO = Purchase Order).
As output to this step we produce a set of extracted tags for each family in the form: $Term_1\_...\_Term_X$ (ex.: *ABIEPostalAddressType* that becomes *ABIE_Postal_Address*)

**Normalisation Step.** This step detects those terms that are not useful for the taxonomy and provides the lemmatization of accepted terms. Tasks for this step are: *Case normalisation*, all terms are converted to lower case; *Stop-word normalisation*, removes words like "of", "a", "for",…; *Bad words detection*, terms unknown by the dictionary are cast aside and; *Morphological and semantic normalisation,* which consists in finding the stem and lemma form.

**Build Taxonomy Step.** The aim of this step is to create a first level of semantic relationships and hierarchy between words of the taxonomy.
For this tags are recomposed using their lemma in order to be able to detect similarities between tags (thus between concepts of the ontology that we are building). In second stage, seeing that tags are usually composed by more than one word, a graph

based on Galois lattice is built to relate those tags having the same words (ex. *address* and *postal_address*); we calculate the Term Frequency of graph nodes and; we remove the nodes that are insignificant (values below a threshold). As last synonyms are checked and added to the lattice (just applied to words belonging to the taxonomy itself).

**Filtering Step.** In this step we analyse the words rejected by a first pass and we try to detect false semantics present within a tag. The first task is the *Bad words "reconciliation*", where we try to detect as many abbreviations as possible applying a modified version of the N-Gram algorithm and Levenstain distance. We look for abbreviations of terms already present within the taxonomy and not in a dictionary, because we would detect too many similar terms, most of them out of context. A second task tries to detect *"useless words"*. Using the lattice we automatically detect those words that present disproportionate relationships between graph nodes (like *Type* or *CommonData*), and therefore do not convey any semantics in reality. Finally new terms are integrated.

**Build Views Step.** At this point, we have implemented some visualization methods to view our taxonomy. We have implemented the following views: as list, as tags lattice (with synonyms relationships) and as tag cloud. Others, like "Social Network of Word", are under development.

## 3.3 Results

Table 1 resumes the collection of B2B standards and some information about their declared relationships with other organizations. This table also resumes for each standard body the following information: number of XSD files that they provide (or in some cases, just the files that we have considered), the total number of complex type and element tags, the resulting number of "semantically" different words and; since XML tags can be composed of real dictionary words, mere abbreviations, or simply any sequence of characters, the last column provides the number of words unrecognised by the system.

This table shows several aspects regarding current B2B business standards. On one hand they highlight some XML schema definition practices by standardization bodies, such as the use of anonymous types for elements, rather than declared types; the adoption of Upper Camel Case or hyphen for tags to separate compound words (which is what we implement); the trend that financial and related bodies often use abbreviations rather than real terms for tags whereas standardization bodies mainly use common words for tags. Therefore it is possible to define a common taxonomy for the B2B domain. In fact, as shown in Figure 3, by adding one standard at a time, even in a random order, we have observed that after half a dozen of additions, less than 20% of words are really new, and reach only 9% of new words in the last standard added. We have noted that these words usually represent terms characterizing the standard, but that the other, more general terms are already present in the global dictionary. Also we have observed that 60% of the words are shared between standards, 11% of the words are used by more than 10 of them and that this trend increases if measured over tags. So it shows that a dynamic taxonomy like this evolves easily and that a shared vocabulary emerges naturally.

We obtain 70976 tags, of which after normalization about 20000 are distinct. The total number of different words composing them is only 2887. On average, standards share three words over four. For example, *PostalAddress* is a tag, composed of 2 words. *PostalAddressTown* is a tag composed of 3 words. A standard composed of these two tags (normalized elements) would have 5 words, of which 3 are different (Postal, Address and Town). A tag called *PostAddrTwn* would be the same normalized element as *PostalAddressTown*.

## 3.4 Special Concern for "Bad Words"

A considerable number of unrecognised words still remain (see Table 1), at least at first sight.

The analysis shows that these bad words are of the following type: mostly abbreviations (about 50%); about 30% are compound words not split by the system (for example compound words not written in UCC form like *worktime*); about 10% are words not included in the dictionary; and another 10% are acronyms.

Several techniques can be implemented in order to improve the detection of hidden words. Our implementation of abbreviation discovery is able to detect more than 70% of them automatically, which in reality corresponds to 80% of total occurrences (for example *amt* => *amount* has 958 occurrences thus more important than *wvg*=> *waiving* with just one occurrence). Improving these results means (a) adopting a more complex management of abbreviations in order to detect different words having the same abbreviation, (b) implementing NLP techniques in order to mine text documents that often come with XML files and; (c) improving the external dictionary's capabilities.

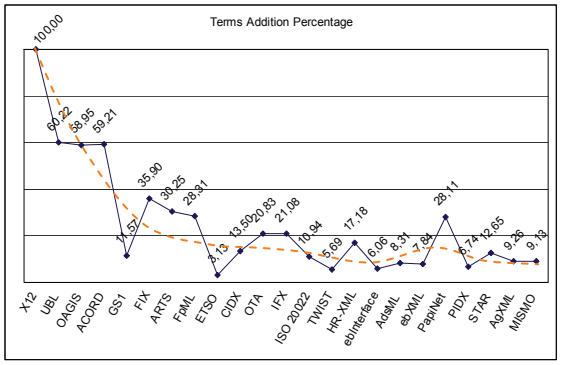


Figure 3: Graph of sequential of terms addition (measures are in percentage).

Therefore solutions that provide good precision and recall exist, but in order to fully exploit the potential of semantic technologies, source document should be somehow **semantically well formed**. No semantic application will be able to understand the sense behind tags such as *AmortMktValDiffPct* or *setr.100.101*.

Another improvement in this direction should be to exploit the structural content of XML files. Rather than using tag name with abbreviations for indicating structural relations like *PostAddrTwn* (11 chars) using simply *Town* (4 chars) as sub-element of *PostalAddress* should be enough for a machine to understand that town is a propriety of the address concept. A positive side effect is the economy of physical space.

# 4 RELATED WORK

Our work is related to several research domains. For work closer to B2B we can cite an interesting experience by Zaho and Lövdahl (Zaho 2003b), that provides an approach to develop ontology for Internet commerce by reusing XML-based standards. They also define layers and relationships of the common vocabulary as shared in the following parts: Core, General, Reusable and Special. But they do not go any further and do not provide concretely a taxonomy. Gloria Giraldo and Chantal Reynaud (Giraldo, 2002) have developed a semi-automatic ontology generation software for the tourism industry domain extracting information contained in DTD files. This experiment is really close to our use case but is limited to the sole domain of tourism, which is defined in advance with great precision, and therefore the detection of relevant concepts does not produce conflicts between different representations.

Other experiences that try to mix semantic integration and B2B thesaurus were developed by (Fensel, 2001b) and (Corcho, 2001), but their work was limited to catalogues of products like UNSPSC and eCl@ss, which have hierarchy and semantics well defined. In practice, the goal is the mapping of two thesaurus rather than the construction of an ontology. For more related semantic integration the document by Noy (Noy, 2004) provides an exhaustive list of experiences where our tool should be effective similar in terms of construction techniques, but they mainly target the merging of two input sources at a time. Concerning the automation process of taxonomy and ontology generation in (Bedini, 2007) is shown that solutions implementing an automatic method for such a task are rare. We do not have the room to detail this here.

Finally, for the construction of reference ontologies, the experience of D'Aquin et al. (D'Aquin, 2007) is significant for our work, but they do not consider XML Schema sources.

# 5 CONCLUSIONS AND FUTURE WORK

In this paper we have presented our starting point for building B2B applications in agreement with the "Next Generation Semantic Web Applications" as described in (Motta, 2006).

Despite the great amount of XML files available, current tools and software are only able to extract semantics from text corpora, or ontologies: tools providing the analysis of a consistent group of XML files are rare, and none really exist in the B2B domain. We have therefore developed *Janus*, a tool capable of extracting valuable semantic information from such corpora and have demonstrated its results with the automatic construction of a B2B taxonomy. Although these results are encouraging, it is clear that our system does not yet offer enough to build a canonical data model for the B2B use case, nor does it reduce application integration to an automatic task. We plan on continuing this work with the development of a more complete tool, capable to associate semantic concepts to discovered taxonomy's terms in order to build as automatically as possible a reference ontology for the B2B domain.

Table 1: Presentation of involved B2B standard and of the correspondent extraction of XML semantics.

| Standard Body | Business Area | Alliances | Files | Tags | Dictionary words | Unknown words |
|---|---|---|---|---|---|---|
| **ACORD** | Insurance, reinsurance and related financial service | X12, XBRL, HR-XML | 8 | 5263 | 1162 | 657 |
| **AdsML** | graphics communication | | 14 | 737 | 301 | 10 |
| **AgXML** | Agriculture supply chain | ebXML, CIDX, RAPID | 11 | 808 | 216 | 4 |
| **ARTS** | Retail | | 44 | 5853 | 734 | 44 |
| **CIDX** | Chemical | ebXML, RAPID | 61 | 1881 | 437 | 20 |
| **ebXML** | Cross industry | | 74 | 1401 | 408 | 10 |
| **ebInterface** | Invoice | | 1 | 105 | 66 | 6 |
| **ETSO** | Specific electric transaction | ebXML | 1 | 27 | 32 | 0 |
| **FIX** | Banks, broker-dealers, exchanges and institutional investors | SWIFT (ISO 20022), FpML | 18 | 552 | 117 | 93 |
| **FpML** | Financial | FIX, FIXML | 21 | 2124 | 544 | 34 |
| **GS1** | Supply chain for Healthcare, Defence, Transport & Logistics | ebXML | 289 | 2360 | 216 | 8 |
| **HR-XML** | Human Resource | ACORD | 166 | 12717 | 949 | 71 |
| **IFX** | Financial | | 310 | 4256 | 446 | 249 |
| **ISO20022** | Financial | IFX, OAGIS, TWIST | 74 | 11082 | 256 | 384 |
| **MISMO** | Residential, commercial, eMortgage | IFX, ACORD, ASC X12 | 14 | 1432 | 252 | 26 |
| **OAGIS** | Cross industry | ebXML | 515 | 4584 | 704 | 170 |
| **OTA** | Tourist | | 233 | 3649 | 552 | 67 |
| **PapiNet** | Paper | | 42 | 1394 | 530 | 18 |
| **PIDX** | Petroleum | ebXML, CIDX | 26 | 745 | 341 | 9 |
| **STAR** | Automotive retail | OAGIS, ebXML | 181 | 5518 | 1130 | 88 |
| **TWIST** | Supply chain, payment | FpML, FIX, SWIFT | 18 | 2489 | 457 | 184 |
| **UBL** | Invoicing, ordering | ebXML | 11 | 650 | 274 | 10 |
| **X12** | Cross industry | | 9 | 1349 | 271 | 23 |
| | | Sum*: | 2141 | 70976 | 10395 | 2185 |

* This sum value does not consider eventual correspondence of common tags or words between different bodies

# REFERENCES


Anicic, N., Ivezic, N., Jones A., 2005. *An Architecture for Semantic Enterprise Application Integration Standards*. In proceedings of INTEROP-ESA 05, Geneva, Switzerland.

Bedini, I., Nguyen, B., 2007. *Automatic Ontology Generation: State of the Art*. Technical report, University of Versailles. (http://bivan.pagespro-orange.fr/Janus/index.html)

Charlet, J., Bachimont, B., Troncy, R., 2004. *Ontologies pour le Web sémantique*. In Revue I3, numéro Hors Série «Web sémantique», 2004.

Corcho, O., Gomez-Perez, A., 2001. *Solving integration problems of e-commerce standards and initiatives through ontological mappings.* In Proceedings of the Workshop on e-business and Intelligent Web, 2001.

D'Aquin, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Sabou, M., Motta, E., 2007. *Watson: A Gateway for Next Generation Semantic Web Applications*. International Semantic Web Conference, ISWC 2007.

E-Business W@tch observatory, 2007. *The European e-Business Report, 2006/07 edition*. 5th Synthesis Report of the e-Business W@tch, on behalf of the European Commission's January 2007. (http://www.ebusiness-watch.org)

Fensel, D., 2001a. *Ontologies: Silver bullet for knowledge management and electronic commerce*. Springer-Verlag, Berlin.

Fensel, D., Ding, Y., Omelayenko, B., Schulten, E., Botquin, G., Brown, M., Flett, A., 2001b. *Product Data Integration in B2B E-Commerce.* IEEE Intelligent Systems, vol. 16, 2001, pp. 54-59.

Giraldo, G., Reynaud, C., 2002. *Construction semi-automatique d'ontologies à partir de DTDs relatives à un même domaine.* 13èmes journées francophones d'Ingénierie des Connaissances, Rouen.

Gomez-Perez, A., Benjamins, V., 1999. *Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods*. IJCAI-1999, Workshop on Ontologies and Problem-Solving Methods.

Hohpe, G., Woolf, B., 2003. *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Addison-Wesley, October 2003

Motta, E., Sabou, M., 2006. *Next Generation Semantic Web Applications*. In Proc. of the 1st Asian Semantic Web Conference (ASWC), Beijing, China 2006.

Noy, N., 2004. *Semantic integration: a survey of ontology-based approaches.* SIGMOD Record, Vol. 33, No. 4, December 2004.

Sabou, M., D'Aquin, M., Motta E. (2006) *Using the Semantic Web as Background Knowledge for Ontology Mapping.* In Proc. of the International Workshop on Ontology Matching, collocated with ISWC'06.

Welty, C., 2003. *Ontology Research*. AI Magazine, 24(3).

Zhao, Y., Sandahl, K., 2003a. *Potential Advantages of Semantic Web for Internet Commerce*. Proceedings of International Conference on Enterprise Information Systems (ICEIS), .Angers, France, April 2003.

Zhao, Y., Lövdahl, J., 2003b. *A Reuse-Based Method of Developing the Ontology for E-Procurement*. Proc Second Nordic Conference on Web Services, 2003.