

XML alone is not sufficient for effective WebEDI

Fábio Ghignatti Beckenkamp, PhD¹ and Prof. Wolfgang Pree, PhD²

¹Telefonica Solutions – Mercador SA
Rua Dona Laura 320, 7º Andar, 90430-090 Porto Alegre, RS, Brazil
Fone: +55.51.3378-9000, Fax: +55.51.3378-9048

²Department of Computer Science, University of Salzburg
Jakob-Haringer-Str. 2, A-5020 Salzburg, Austria

fbeckenkamp@mercador.com, pree@SoftwareResearch.net

WebEDI relies on the Internet infrastructure for exchanging documents among companies. Typically these documents are orders and invoices and can be exchanged by directly integrating the companies' ERP systems or via Web application upload, download or typing of documents. XML is considered as a state-of-the-art way to reduce the complexity of managing the different data formats. Nevertheless, the XML standard alone does not offer the required semantics. When integrating a few companies, it is possible to hard code the semantics in the applications that process the documents. However, when integrating numerous different companies such a solution does not scale. The semantics has to be captured in a more flexible and scalable way. To cope with this necessity, simple ontologies are required to augment the data description.

XML alone is not sufficient for effective WebEDI

WebEDI relies on the Internet infrastructure for exchanging documents among companies. Typically these documents are orders and invoices and can be exchanged by directly integrating the companies' ERP systems or via Web application upload, download or typing of documents. XML is considered as a state-of-the-art way to reduce the complexity of managing the different data formats. Nevertheless, the XML standard alone does not offer the required semantics. When integrating a few companies, it is possible to hard code the semantics in the applications that process the documents. However, when integrating numerous different companies such a solution does not scale. The semantics has to be captured in a more flexible and scalable way. To cope with this necessity, simple ontologies are required to augment the data description.

Introduction - The Mercador B2B solution

Mercador is a B2B company focused on the retail market and has built an infrastructure that allows companies of different sizes to be integrated through the Internet. Mercador implements the concept of a hub. The hub main services are the documents transference, translation and validation. The transference of documents is done from one company to another via the Internet using the Mercador integration servers as the connection hub. When a company sends a document to another, the document is sent to the Mercador hub first, the hub processes the document to validate and translate it to the receiver format. The delivery of the document is also done over the Internet. Therefore the hub implements a set of services that make easier the integration of many with many companies. The solution has tools that support the integration of companies at different levels of automation.

The very basic infrastructure for the companies integration is the Internet. The idea is to bring to the Internet the possibility of exchanging the documents that form the supply process. Those documents, like a protocol, permit the companies to exchange information about their business so as to guarantee the fast execution of the supply process. Mercador has developed a set of Internet functionalities based on the traditional EDI (Electronic Data Interchange) process. Called WebEDI, the system allows companies to exchange business documents using the Internet as the connection infrastructure. So that WebEDI is the transmission of standard electronic documents among companies or business organizations using the Internet.

The transformation and validation of the documents format is an important service that Mercador provides for its customers. One of the most complex tasks of integrating companies is the agreement on the data formats and meaning. Mercador implements, at the hub, services that validate and translate the customer's documents to the Mercador standard for each type of document. From the Mercador standard, the document is then translated to the recipient format. By doing this, Mercador is able to manage the data differences among the different companies. Therefore Mercador implements, at the hub, the necessary knowledge to facilitate the exchange of documents among many with many companies. The hub performs the necessary semantic process of the documents. The transformation and validation of documents is an important service that the hub offers to its customers and which minimizes the difficulties of integrating with many companies.

The Next section explains the XML strengths and limitations when applied to the WebEDI.

The XML use in WebEDI and its limitations

XML has being increasingly used as a standard for data interchange for applications over the Web. Its widespread adoption has particular significance for enterprise application

integration, which is the case of the WebEDI application. It is important, in this context, to understand the strengths and limitations of XML (Yee, A.- eai.ebizq.net).

XML's Strengths

1. *Powerful meta language.* Provides a mechanism by which other markup languages can be developed for specialized needs or business domains. Examples are: Chemical Markup Language (CML) and Artificial Intelligence Markup Language (AIML).
2. *Simplicity.* The document structure and contained data are usually readable and easy to understand.
3. *Separation of content and presentation format.* The document separates its content and presentation information clearly.
4. *Common open standard.* It has no adherence to proprietary technologies such as browser, editor or interpreter. In contrast, it has been adopted across multiple industries.

XML has limitations with regard to application integration that are relevant to enumerate.

XML's Limitations

1. *Inefficiencies of text-based documents.* The document representation in text files can be very large due to repetitions of data structures. This may be inefficient when transferring the document over the network.
2. *No data transformation facilities.* XML standards have been developed for business-specific exchange formats applied in the B2B space, such as RosettaNet (www.rosettanet.org). However, the adoption of such standards is still not a reality and the transformation of data among disparate systems is necessary. XML is simply one of many data formats that exist; the transformation of data from one format to another is still a primary challenge.
3. *Absence of content-based routing.* The automation of the integrated business process requires content-based routing and rules. The XML file does not contain the required content and business rules to automatically deliver the document to the proper destination. This work has to be done by an extra system.
4. *Limited semantic interpretation.* XML provides the ability to create specialized tags that describe a particular entity or behavior; however, the semantic interpretation of what the data represents is outside the scope of the XML document. The meaning of tags and data within the document has to be agreed among the integrated parties.

The use of the XML can be tricky even when applying its strengths. For instance, the use of an important characteristic, such as flexibility, can lead to problems when uncoordinated efforts to design schemas using XML produce incompatible results. It is easy to create XML schemas that are only usable in a single application. This is typical when competing industry groups develop different schemas for all members of that industry.

Integration brokers are specialized in implementing a coordinated integration effort and to complement and compensate the given XML limitations listed above. The cited limitation 1 (*Inefficiencies of text-based documents*) can be overcome by the use of binary files. The integration broker implements a parser based on the XML Document Type Definition (DTD) that is able to read the XML document, parses and compiles it into binary format that is transferred. The processing (parse and transformation) of those large documents may also be very CPU intensive.

Limitation 2 (*No data transformation facilities*) is not an exclusivity of the XML standard. The real world is full of different standards provided by different industry sectors, integration brokers or simple system developers. Using XML helps minimize these differences but does not eliminate them. The integration broker implements, at the hub, a transformation process that always includes its XML standard for the given transferred documents. The hub has services that implement the transformation of each given format to the broker and from the broker to the given format. This allows the hub to apply necessary data transformations that guarantee that the document, once in the broker XML format, has reached a certain level of standardization of its data structures.

In the case of limitation 3 above (*Absence of content-based routing*), the integration brokers are required to implement the document delivery by reading the document content and applying the appropriate routing business rules. The simple association among the involved exchange parties can be represented by the description of knowledge like: supplier X supplies the retail A and B. This information is not contained in the XML document; it is implicit to the document existence. But in the case of a many to many exchange, not all participants are willing to exchange data with all the participants. The deliver of a document has to be confirmed by the existence of a valid relationship among the companies contained in the document. Therefore, the broker shall represent this information somehow and somewhere.

The limitation 4 (*Limited semantic interpretation*) has direct impact on WebEDI documents exchange that requires knowledge about the partner's business rules such as:

1. The correct interpretation of data attributes.
2. The application of partners business rules when processing documents.

Case 1 suggests that the information contained in a document has to be interpreted in the same way by both sides, the sender and the receiver. This requires the representation of the abstractions that complement the information about the document-contained data in order to assure its proper process. This is called the semantics of attributes. Examples are:

- Tax – For both companies it is a numeric value but the sender considers Tax a value from 0 to 100 and represents a percentage. However, the receiver considers Tax a value that represents a monetary quantity (the already calculated tax value).
- Package type – Both companies have the same meaning for it, but the domain each one accepts is different. There should be an agreement among the companies to reduce the domain to a unique set.
- Discount – Again both companies have the same meaning for the attribute but the range for each one is different. There should be an agreement on the range or a formula to appropriately convert the values.

There may be many examples like the ones above regarding the attributes meaning. Mercador implements a database where the schema for the attributes of each document from each company is represented. Before starting the companies operation, there is a system that confronts the schemas of both companies finding the unconformities like the ones above. After an agreement, the rules are stored to be evaluated during transformation.

Case 2 requires knowledge representation of business rules about the partner's relation to the specific document type. One example is that different buyers require different invoice information from the supplier. Another example is the negotiation conditions among partners that are given by groups of discounts to different buyers. Below are some

- If attribute A has a valid value, then attribute B does not have to be informed.
- If attribute A has a value less than X, then attribute B must have a value above Y.
- If the buying quantity is above X, apply discount Z to all order items.
- If buyer J belongs to the group L, then apply discount Z.

The integration broker typically solves the limitations of semantics by implementing the business knowledge. In this case, the broker becomes the container of the knowledge. Next session explores the alternatives to include semantics in the XML documents.

Including semantic in the WebEDI

The limitations stated above can be solved in many ways. Depending on the tools and software architecture, the B2B integration company (broker) may solve the necessary semantics by simple custom made solutions. However, the goal is having a solution that scales to an environment of many to many companies exchanging documents and where each company provides the necessary document semantics. The B2B partners should be able to understand the documents doing the integration automatically.

Following there will be an explanation of the pragmatic solution adopted by Mercador to solve the cited limitations. In addition, the new technologies, which are popping up to build a Web containing the semantics, will be presented. These technologies can be extremely useful to the B2B integration.

Mercador actual solution

Document routing

The document routing is easily solved by Mercador once it is a hub. The companies that make part of the Mercador hub always send documents to their partners through the Mercador hub. So the connectivity among the companies is solved by the knowledge the hub holds about the companies connectivity characteristics. Mercador implements in database the necessary information to route the messages from and to the companies. There are no direct connections among the companies to be controlled.

The companies that route documents using the Mercador hub are required to send the documents using one of the three different connectivity tools:

- WebEDI application – Accessing the WebEDI application the user can do:
Document typing – The sender types the document in a special form on the Mercador WebEDI application at the web site; Web site download/upload – The sender uses the Mercador web site to download/upload files over the Internet to the hub;
- Mercador Client – The Client is an application that is installed at the company computing environment and that is used to send/receive files to/from the Mercador Hub using the Internet;
- FTP service – The customers may also send or receive files from/to the Mercador hub using an FTP service.

In the first alternative, the sender is identified by the WebEDI user authentication at the Mercador site. In the second case, the user is identified by an authentication service provided specially for the Client and in the last case, the Mercador system does the FTP connection to the company FTP server.

The document receiver, in the case of document typing, is selected from a list of companies, and in the last cases it is collected from the sent content of the sent document. In any case, the document can only be routed to the addressee if there is a valid commercial relation among the companies. Thus, this is the first business logic that Mercador implements at the hub, the so-called *commercial relation* is maintained in database and identifies which companies are prepared to exchange documents. Figure 1 shows the ER model for the commercial relation. Note that for each type of document there must be a valid relationship. A company is only able to send documents to another one that has an explicit commercial relationship for the given type of document.

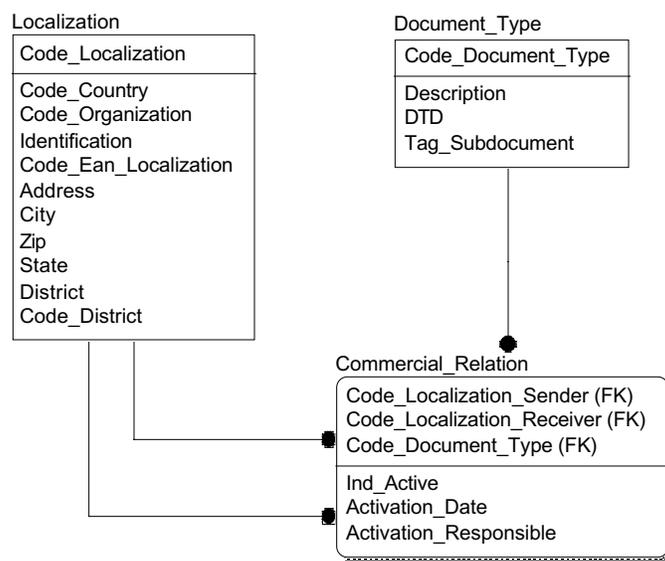


Figure 1 - Commercial relation

The connection using the Mercador Client requires no extra information at the hub. However, the Client has to be configured to connect to the hub. The Client is responsible for requiring the connection to the hub and sending new documents or requiring the available documents at the hub. First, the Client has to be configured with the appropriate IP address (or DNS) of the Mercador server hub and has to be configured to authenticate at the hub. Second, the Client runs at the company computer environment behind firewalls and it must be appropriately configured to cope with the necessary security restrictions of the company network. Usually the Client connects to the Mercador hub using the HTTP protocol on port 80 or the HTTPS protocol on port 443 not being a security problem for the companies.

The opposite occurs for the FTP service. The Mercador hub is responsible for having the login information on the company FTP server. This information is typically FTP server IP address, user and password. Besides this, the hub shall implement a schedule to connect to the FTP server according to the business necessities of the company. For instance, it must connect each 30 minutes from 6:00 to 15:00. Such kind of information belongs to the semantics of the customer business and has to be maintained from the Mercador hub. Once again it is maintained at the Mercador WebEDI database.

Semantic interpretation

The Mercador hub implements the integration among the companies by transforming the document from the sender format to the addressee format and by providing the connectivity among them. To accomplish the document transformation, Mercador specifies a standard XML format for each type of document. Whenever a document is

sent using the hub, it is first transformed for the Mercador XML standard format. By doing this, Mercador reduces the complexity of mapping the divergences among many companies. The premise is that if a document can be transformed for the Mercador standard, it is possible to transform it to the format of the addressees that are already connected to the hub. To complete the job, the document already transformed for the Mercador standard is then transformed to the addressee format.

The WebEDI application server is responsible for coordinating the document transformation process. The integration server is responsible for performing the transformations and for the connection services for the Mercador Client and the FTP services.

Below are the steps during the B2B transformation of the document:

- Sender schema validation – Used to validate the document structure. For each company and document type, it is implemented at the integration server a document schema that is used to validate the incoming document.
- Sender transformation service – Transforms the document for the Mercador XML standard. Any rules regarding correct transformation of the sender document and the Mercador Standard is evaluated at this moment. The rules regarding the business relation between the sender and the receiver are also applied at this moment (for instance, the commercial relation is evaluated). Therefore, the document is transformed for the Mercador XML standard only if it applies all the necessary validation.
- Receiver transformation validation – Transforms the document from the Mercador XML standard for the receiver format. Once in the Mercador XML format, the document is ready to be transformed for the receiver format. There should exist no restrictions or rules to be applied any longer.

A previous work on the details about the document attributes and semantics has to be done in order to assure that this two-step transformation will work. It is necessary to guarantee that the information transformed from one side to the other keeps the same meaning. An integration expert does this work in a careful mapping process supported by a visual tool for documents layout description.

When a new company joins the hub, it is supposed to be connected to hundreds of companies for exchanging documents. A complete mapping of its documents has to be done and their information has to be in agreement with all the companies it will integrate. It is not practical to do the complete mapping of each type of document it supports to each of the related companies by hand. The goal should be to reduce as much as possible the dependence of human interference on the mapping of divergences among the integrated companies.

This description of the document structure is done in a tool developed by Mercador called DSD – Document Structure Descriptor - where the layout of each company is described in a visual manner. This mapping starts by describing each document attribute in order to model the document structure, the attributes types and possible values associated to the document. Figure 3 shows the ER diagram for the DSD.

The Mercador integration expert constructs this description. Once the description is finished, the DSD tool generates a divergence report from the new description and the Mercador standard description. It also generates a divergence report from the new layout and the layout of the companies that the new one will integrate. These reports guarantee that the attributes in both companies have the same type, valid values and the same meaning. The fact that the attribute is related to an attribute in the Mercador standard, guarantees the meaning.

concept of having documents containing semantics on the web and which can be interpreted by Web enabled mechanisms such as agents or web services. The idea is that the information on the web is given explicit meaning. Thus, it is possible to build automatic information processing and integration.

As already explained, XML is useful for defining customized tagging schemes. Complementarily, the RDF's (Resource Description Framework – RDF, 1999 – www.w3c.org) is a flexible approach to representing data giving it semantic meaning. The RDF Schema is a lightweight language to provide basic structures such as classes and properties. RDF is very simple, being very similar to a basic directed graph. It defines how to describe the relationships among resources in terms of named properties and values. The RDF properties represent traditional attribute-value pairs and may also represent relationships among resources. Besides this, RDF is made available on the web by using a serialization syntax based on XML. RDF represents the abstract model and XML provides the concrete textual representation of the model. There are several ways to represent the same RDF data model in XML.

Nevertheless, RDF should include some extra capacities such as data types, consistent expression for enumerations, etc. For the sake of having a more expressive language, a Web ontology language has been defined given the XML and RDF capabilities. The Web Ontology Working Group (WebOnt – www.w3c.org; Clark, K. G. 2002) has been given the task of developing an ontology vocabulary for use in the Semantic Web. This vocabulary should allow the explicit representation of terms and the relationships between entities in this vocabulary. WebOnt intends to create the standardized markup language within which users can formally define specific ontologies for use on the Web. For this purpose, WebOnt is standardizing the DAML+OIL (DAML+OIL, 2001 – www.w3c.org) ontology language.

DAML is the DARPA Agent Markup Language, which is a simple language for expressing more sophisticated RDF class definitions than permitted by RDFS. The DAML group has also added characteristics from the OIL (Ontology Inference Layer), which is a language that uses constructions from the frame-based AI to provide a more sophisticated classification mechanism. (Ouellet R., Ogbuji O., 2002)

Given such knowledge representation language based on XML, the necessary semantics for the WebEDI solution can be completed in a straight-forward way. The SDS tool may still be used to facilitate the building of the knowledge. The whole information about the documents could be stored in a XML file containing the document XML schema to represent the document data structure and the DAML+OIL language definitions to represent the necessary semantics in the same file. In case of storing the complete knowledge into files, the SDS tool would have to be able to do reverse engineering on the persisted files. Storing the knowledge on database would then be optional.

References

Berners-Lee T, Hendler J. and Lassila O. The Semantic Web. Scientific American, May 2001.

Berners-Lee, T. Business Model for the Semantic Web. (www.w3c.com)

Clark, K. G. If Ontology, Then Knowledge: Catching Up With WebOnt. May 01, 2002. (www.xml.com)

DAML+OIL (March 2001) Reference Description W3C Note. (www.w3c.org)

Henry, K. Predicting how Ontologies for the Semantic Web will evolve. Communications of the ACM. February 2002 Volume 45, Number 2.

Ouellet R, Ogbuji O. DAML Reference. May 01, 2002. (www.xml.com)

Ouellet R., Ogbuji O. *Introduction to DAML: Part I*. January 30, 2002. (www.xml.com)

Ouellet R., Ogbuji O. *Introduction to DAML: Part II*. March 13, 2002. (www.xml.com)

Requirements for a Web Ontology Language W3C Working Draft - 07 March 2002.
(www.w3c.org)

Yee, A. Using XML in Application Integration (eai.ebizq.net)

Resource Description Framework (RDF) - Model and Syntax Specification. W3C
Recommendation 22 February 1999.