Research on Elaboration of an Integrated System Based on Xml Data Analysis

Loredana MOCEAN
Babeş - Bolyai University of Cluj-Napoca,
Business Information Systems Department
mloredana@econ.ubcluj.ro

This paper approach the importance of XML for organizing and managing better the data based on texts. This document provides the specification for a data model for describing information organization structures (metadata) for collections of networked information.
As an important result we propose a new model of an integrated system based on XML and using the data analysis It also provides some steps we must follow for this data model using XML, the Extensible Markup Language.
Keywords. *XML, Integrated System, Database.*

Introduction

Many companies and whole corporations have benefit from using the Extensible Markup Language (XML) for organizing and managing better the data based on texts.
XML puts at command a large used-up standard for labeling of the structures and contents of data. With all these, the using of XML were main limited into the interchange of data between the servers of corporations data and not to the level of each engaged person.
XML is a standard language of data description used in whole world for sharing business informations in Web without taking in discussion of the incompatible programs, of computer networks, of data structures or operating systems. XML makes easier information interchange.
Because all the files contains the same XML labels, it is permitted indexing, searching, combining and reusing efficiently of text based information. XML is text-based, so why permits the information interchange between systems which, in normally way are not compatible.
Many companies have implemented XML as standard for transactions which imply extended systems and have invested in networks and Internet services for offering the necessary support.
However, because it wasn't available the all easier XML instruments easy to use for office programs, the employees rarely was labeling the structures and contents of the own documents.
The opportunity to capture and reuse further the data contained in these was missed, or the companies were due to miss longtime in migrating the data to a system level of enterprise.

Results

As an important result we propose a new model of an integrated system based on XML and using the data analysis. The advantages of our model:

- The work-process is adaptable, the work-process can be modified and adapted to the requirements of the projects and customer;
- The capacity to process in short time a big volum of data;
- The efficient communication with the customer and the quick reaction to this requirements; we obtain them within the risen flixibility within the organizational level;
- The exteriorization of data processing services by dint of multiple languages;
- We need to design differently levels of gave through which process can assure these levels.

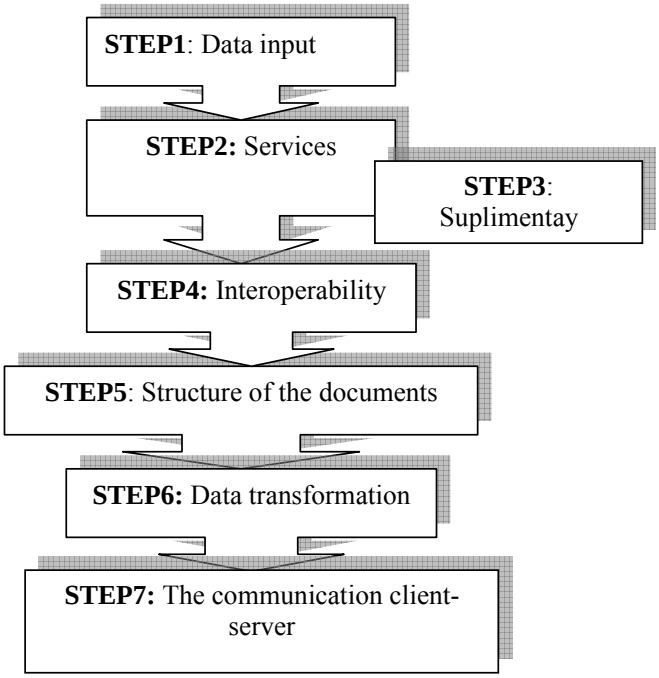

Fig.1. The steps followed in the elaboration of the model

The first step: data input

When we design the module of data input, we have to think to the following problems:

- When we propose the model we think that the data will be introduce by specialized users.
- We will implement some data validation procedures to assure an excellent quality of introduced data.
- The data input will be made with the help of online or offline, data indexes, e-books, databases, market studies, etc; for these services may benefit commercial banks, financial institutions, assurance companies, educations instititutions, universities and other institutions which work with a big volum of data.
- The data input must be done by specialized teams on different formats of data, the data input in different programming environments. The data must be delivered in one of the following formats:
  - XML/SGML/HTML (*.xml, *.htm)
  - Comma Separated Values (*.csv)
  - Word (*.doc, *.rtf, *.txt)
  - Excel (*.xls)
  - Portable Document File (*.pdf)
  - Access (*.mdb)
- The data input must be assure directly by the paper documents, microfilms or scanned images. We will use OCR programs, of last generation, for offering a high speed of processing and a high quality level.
- Within OCR with different languages support, we can convert our data in XML/SGML/HTML or PDF format, which permit data search for text formats.

The step two: services of our model

When we think to the facilities offered, we must think that our model must contains a few modules who offer the following:

- The data processing in structured or unstructured format as XML.
- Data analysis.
- The creation and modification of DTD schema.
- The description of converting specifications.
- The data gathering within OCR.
- The verifying and correcting of data.
- The structure and conversion of data qualities.
- Using the efficiently conversion tools based on the DOM technology our model must offer data conversion in XML, SGML, HTML and another structurated formats, departing from different data formats:
  - RTF/DOC

- TXT
- Textline
- QuarkXPress
- PageMaker
- 3B2
- Mediaview

• Automatically and semi-automatically data conversion

The step three: suplimentary services

We want to implement other modalities of data processing such as:

• Images processing.
• Converstion in/between different data formats.
• Creation and processing the forms.
• Creation and maintain of databases.
• The assurance of Total Quality Management.

The step four: the interoperability enlargement between different platforms

This thing can simplify a heterogeneous infrastructure where we have diverse clients/ servers of databases. As a case study let us consider an client/server infrastructure in which some clients work on Windows platform, other clients work on Linux platform, etc. All these clients communicate with different servers: a mainframe, a SQL server database, an Oracle server, etc. In this moment all these clients must communicate with different types of servers. Each among these servers is exhibited below certain protocols of communication more or less proprietress (ex. TCP/IP, SOAP, etc.).

The integration cost rise considerably with the number of clients/servers because we have a proliferation of communication protocoals, impossible to achieve without a common standard. XML offers this common standard.

A MS SQL Server database can be programmed to return the result of a query in XML format only by means of a stipulation of a request such as "`SELECT abcd FOR XML`". At the database level XML offers a high abstractisation level of the platform and keep in mind that on each machine exists at least a XML parser.

The step five: the structure of XML documents

The XML documents may have a more or less tensely structure, given by a XML schema represented by the XSD document. A XML schema is the analogue of the data structures from DBMS or C. We can define structures with a clearly semantic specification. XML offers from semi-structured data up to data which form predefined structures.

XML Schema and DTD standards can define clearely communication protocoals based on XML. Without these standards, XML is just a structureless format for representing data, maybe with little over HTML.

Furthermore, exists a strong analogy between the structure of relational databases and the structure of the XML document, what make more easier integration of existing applications.

The step six: the data transformation

The XML offers some important services of data transformations (Data Transformation Services). In present, each database defines its own set of data types, in a standard mode. Exist many types of data as integer, real, date, etc which, sometime do not have a clarely equivalent in the translation step.

XML can enforce a set of common data the si therewith facilitates the adoption of a standard, there, where a common convention is difficult to find. Therewith, exist advanced standards for changing of XML documents (as XSLT), and we propose the using of XSLT directly on server.

The step seven: the communication client/server

We already have a wide variety of Java-based XML and HTTP tools to choose from, but you can also take advantage of a prepackaged set of XML-RPC tools. Although the XML-RPC is very useful for debugging and for establishing connections between systems in different environments, you can treat XML-RPC much like you do any other Java feature. There's some setup work to do, especially for XML-RPC servers, but most of

this work is simple and needs to be done only once in the course of a program.
In time, ADO or ADO.net has imposed as a client/server communication technology. A client receive data from server in XML format. This, goes to the simplification of communication between client and server and the indepenedence of technology from the platform (DOT.NET can be used from platform non.NET).

Some important things we must relay on, when we build the system

We must remember as XML, although a very strong technology, can lead to riskiness if it not used proper, as follows:

- XML can bring a penalization of performance, on part of generation/parsing/validation because we can have a big consummation of memory, in the case of a wrong design.
- XSLT technologies deserve many attentions from the viewpoint of the performance.
- XML is not fit to the representation of amorphous systemless date, unstructured, because it contains only a limitated set of characters.
- We can transport XML binary data but only after a previous convertion in a text file format; that means an important restriction especially for a kind of applications like wireless applications.
- XML data fills a space much more than in binary format, therefore must think about the efficiency transportation of the documents XML through network, especially to appariton of strangleholds.
- XML format can be compressed very well with the classic algorithms, data in XML format fit, generally, less than the natural representation in binary format.

Conclusions

In the last few years, the companies were smashed through limits of the relational databases. In their attempt of rise the performances of the systems, they enlarged the complexity, diminished the performance and rise the costs. Even most newest relational-object databases can't efficiently use complex data structures necessary to an application from the last generation.
Moreover, the applications become more complex, as well as the used technologies, slowing down the developmental process and stability. More, it is the necessity of a quick development of applications.
Our system has a performant database for post-relational era, is a new generation of technologies, which combines a multidimensional data server with a versatile applications server. We use an advanced object technique, rapid development for Internet, an advance programming language, a unique data stock technology, etc.
Our system must support all traditional methods for Web pages development, a unique technology named *Cache Server Pages* (CSP) optimized for rapid development of database system.

References

[Buraga04] Buraga, S., *Semantic Web * fundamente şi aplicaţii,* Matrix Rom, Bucuresti, 2004
[Mindruta05] Mindruta, C., *Arhitecturi, tehnologii si programare WEB,* Matrix Rom, Bucuresti, 2005
http://www.pcworld.ro
http://www.devx.com/Java/Article/16407/1763
http://www.softpageinternet.ro/manual_html
http://www.afaceri-online.net