XML and DATABASES

MR VARNIER FRANCOIS
Prof : Mr HERZ David
INT 2003
1) IMPACT OF INTERNET ON DATABASE............................p3

2) THE XML STANDARD AND ITS DERIVATION.........................p4

3) XML AND DATABASE..............................................p5

4) XML AND FEDERAL DATA........................................p7

5) XML AND THE MANAGEMENT OF KNOWLEDGE..................p8
INTRODUCTION

The explosion of internet and the progress of technologies give some new approaches of databases used in information system. It has some consequences on the management of information which becomes more complicated. Firms are some good consumers of big and heterogeneous information. Also in order to answer to their needs, some new standards such as XML: extended marked language trying to make and represent information. This standard integrates structural data and multimedia data and more of mixed heterogeneous data it gives the interoperability with all objects which use different data. So the management of knowledge becomes more automatically with low costs and also that’s why XML will have a good future.
1) IMPACT OF INTERNET ON DATABASE.

Databases were developed in 1990 around the object and the relational object. In order to answer to the needs of firms’ local area network (lan), it has become an obligation to couple databases and internet technologies. The development of internet and firms’ technologies show three kind of application which need the mix between internet technologies and databases. 

First of all there are the application client server web which are built on three states: client->application server->data server. This kind of architecture is a result of a database which must be opened to extern client of firms’ lan and which can access by the web to the system information of the firms and it gives the possibility to the firm’s personal members to access to the system information of their firms.

After the application client server there is the web generation of web site which are very numerous and very heterogeneous. The contents of web site have become very complex for example we can find several HTML files with different links and it has become very difficult to update it when it is necessary that’s why database is a good solution to solve this problem. It gives a method of representing information and organizes a logical hierarchy between the information.

The last application is the web business. The need of this application are for example to show some information to the consumers and more of that it must memorize the contacts and what consumers like or dislike.

In order to explore the links between the products and to show the hierarchical descriptions and prices it must use some high speed connections and some transactional databases.

The new architectural of web data which are coupling the web and the database, needs to extend HTML to take some data and some requests and to make the web servers more dynamic and to increase the bandwidth between database and the application.

Today the architectural are made of three levels:
.A level which controls the user’s interface
.A level which has got the web server and it adds an application server.
.A level which gives to the firm the control of the data in the database.

Each of them must give the possibility to the application’s code to run and to show the data and to treat the application’s logic or to manipulate some complex objects which are stored in the database.

As a consequence of a lack of HTML’s extension new development of data web is going to define a model of exchange who is the information’s content and not the presentation.

So the XML defines a standard model of data exchange.
It is a simple textual model with tags. It gives the possibility to build some complex documents by aggregation and multivalued association. It authorizes the definition of rule’s integrity who imposes a structure. XML must have a powerful impact on database and should make a revolution in the database. It introduces a new kind of database calls datawebhouses.

2) THE XML STANDARD AND ITS DERIVATION.

The XML was born of the SGML(standard generalized markup language) from the W3C consortium. As HTML, xml has got some tags which represent the structure’s definition. figure 1 shows an extract of xml code. The first line defines the xml’s version and the code to use. The second line gives the access of the DTD (document type definition). The DTD defines the tags and the structure of the document. The document describes some heterogeneous bars and restaurants.

Figure 1

```xml
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
<!DOCTYPE Guide SYSTEM "voyage.dtd">
<Guide version="2.0">
  <Restaurant>
    <Nom> Aubergeade </Nom>
    <Telephone> 0148152256 </Telephone>
    <Manager> Dupuis </Manager>
  </Restaurant>
  <Restaurant>
    <Nom> La Licorne </Nom>
    <Rue> Des Moines </Rue>
    <Telephone> 0148253278 </Telephone>
    <Manager> Dupuis </Manager>
  </Restaurant>
  <Bar>
    <Nom> Rose and Crown </Nom>
  </Bar>
</Guide>
```

XML is different from SGML but there are no defining tags. These tags are fixed by the application. They must be in the right order. First of all, the document must begin with a beginning tag and finish with a end tag. The tags frame the elements. The tags allow us to build some elements of the document. Each element is linked with a simple or complex attribute. They can be integrating in other elements deeply as necessary. A document is simply a continuation of elements which are built into other elements.
The DTD (definition type document) specifies the scheme of the documents. It can be seen like a constraint of integrity and it introduces the notion of well formed document, it means to be conformed with the DTD. 

A family of standards are running now around the XML and the W3C:

- XLL: language of definition for the hyper textual links and xpointer.
- XSL: language of presentation which can define the styles with rules presentation integrating cascade sheets.

- DOM: interface of navigator in the documents for object oriented language.

- OTP: a protocol for the web commerce based on XML.

- XML/EDI: definition of exchange data which is defined in XML.

- XML/SCHEMA: a standard which is used to produce xml document much more similar to a database than the DTD.

XML is a good language if it is used to define some hierarchical documents. With XML schema it is possible to make some database schema like an object model.

3) XML AND DATABASE.

The relational data or object need to have a rigid definition scheme which is difficult to modernize. The data must follow these schemes. As a consequence of that, three problems arise:

- The data can be not conform to the schema.
- The structure of the data evolves and needs to be modified frequently.
- The data which can be weak structure such as the free text implements the use of technical interrogation which is complex for example the search by key word.

That’s why the research implements a new data model which has got a semi structure. In this model all the documents are represented as a labeled graph. Each rainbow corresponds with a xml tag and the knot leafs are the data elements figure 2.
But which language can be used for finding the documents in a semi structure database? The W3C work on it. Some propositions are done: XML-QL of AT&T, XQL of Microsoft. It is a possibility to extend the sql with navigability expression and filter with graph builders, it is the XSQL language. This new query language should have an important impact on the search engine.

Today the relational DB as oracle or DB2 are the leaders of the market. The semi structural object can be integrated as an instance of a particular type in a relational object system. Oracle uses his type "<free text>" (cartridge intermedia) to make it possible for storing documents in xml format and using the interrogation by keyword.

Figure 2: An example of a semi structural database.

Figure 3: An example of a storage XML document in a DBR. (source from eXMLmedia).
4) XML AND FEDERAL DATA.

A federal database is an heterogeneous database which is made of different data such as text files, HTML documents XML…..The objective is to give an integrate view to the users of the different data from the enterprise dynamically on request (Technology PULL), or materializing periodically in a data store (Technology PUSH).

Figure 4: An example of federal database (source from eXMLmedia).

Historically this kind of database is developed around the relational model and sql.
Corba makes it possible to integrate the applications. But corba hasn’t got a real standard interface with database and it is not very easy to evolve it.
On a point of view about the architecture of the federal database, there is a standard from DARPA and GIO Widerhold.

Figure 5: Architecture DARPA 13 source from eXML-media
- The central level is a mediator which integrates the data from different sources for composing the responses to the users requests.
- The most internal level is made of several adaptators or wrappers which must transform the local data in order to be compatible with the mediator and to filter the data after a request.
- The external level is made of facilities, these tools must locate the pertinent data and order it for the applications.

The mediator and its interface are an exchange bus called bus EAI (enterprise application bus). But the object mediators have failed because they are too complex if they must be evolved.

XML is a new exchange standard and it can be a solution to take the place of the above architecture.

Figure 6 shows how it can be used. It is an example of a component which were made by the eXMLmedia enterprise:

All the XML-QL requests are received by the mediator which converts them into small requests compatible with the sources which is described in a model referenced for example SQL, pure XML, XSQL.

5) XML AND THE MANAGEMENT OF KNOWLEDGE.

The technical federal databases make the collect of information from different sources possible but it needs to precise and to integrate more information. It is the domain of application technical data mining which represents the future generation of intelligence tools. Data mining consists to extract a part of information from a big source of information and make it exploitable.
It is interesting to take all the fragments which have the same structure and to store it in a relational table in order to have more free volume and more time when it needs to be extracted from the based documents. Finding the repeating fragments is the domain of the data mining.

Another direction of data mining is the automatic tags. A lot of documents from firms need to be transformed into XML files but we must find what kind of important information in the documents should appear. It must find an automatic intelligent tag.

Another way is to used intelligent dictionaries which describe the speech domain, and these concepts and their relations.

**Conclusion**

XML and the semi structural model make easier the integration between web and database. This new kind of database is different that the others because its structure has got more flexibility schema and these are mixed with data. This database takes the links between inter documents and supports hypermedia, also it means the search and the structure of hypertext documents or hypermedia. XML has got a promise future generation. With the evolution of the java language, the XML unifies document formats and java unifies the language of programmation in the three states architecture. However a lot of problems appear around the semi structural database. We have told a little bit of optimization of the storage and of data mining which opens new horizons.
Bibliography:

Online resources

http://www.e-xmlmedia.fr/site_francais/documentation_livres_blancs.htm
Georges,Gardarin Les datawebhouses arrivent

Text resources

Lessons and photocopies from INT.
Glossary:

DOM : Document Object Model
DTD: document type definition
DB: data base
DBR :data base relational
EDI : Electronic Data Interchange
HTML: HyperText Markup Language
XLL :Extensible language link
XML :Extensible Markup Language
XSQL:Extensible Structured Query language
XSL : Extensible Stylesheet Language