Project Proposal: Semistructured (XML) Data Management

Comparison of four Database Management Systems for Semistructured Data

Claudia Ehrentraut

May 18, 2011

Database systems today are more ubiquitous than ever. As the volume of digital data explodes, new requirements for fast, easy-to-use and reliable data management have emerged.
(Freire 2009)

Semistructured data is becoming ubiquitous. (Deutsch 1)

1 Introduction

The incipiently presented quotes outline the background for the project which is to be proposed in this paper.
The subject of the project is to explicitly and in detail compare four of the existing Database Managment Systems (DBMSs) for semistructured data: STORED, XTASY, Index Fabric and Lore. According to Abiteboul (2000), this field of study was quite new at the end of the 90s, beginning of the 21st century, nevertheless all the systems mentioned above were implemented during that time. Evidence that this research area has not lost its relevance is given by the fact that it was proposed as a possible research topic by Juliana Freire in the course *Research Topics in Databases* given in Spring 2009 at the University of Utah.
The main goal of the project is to yield a structured and all-embracing comparison of those DBMSs. This means in particular that the systems are to be explicitly related to one another, geared to some general aspects like: disc space used when storing the data, the extend to which data fragmentation is reduced, the time complexity when running a query, the 'quality' of the results returned by the query, the performance/power of the query language supported by the respective DBMS in general, how missing attributes are handled, how the system deals with a partially unknown structure etc. (Compare e.g. Suciu 2000:14, Goldman 1ff.).
Such a comparison is regarded indispensable in order to improve future approaches in developing suitable DBMSs for semistructured data. Subgoals of the project are to give an understandable and to each other comparable presentation of STORED, XTASY, Index Fabric and Lore. This implies an outline of their initial implementation as well as possibly applied changes and improvements within the past decade.
A more fundamental subgoal is to expound what the basic characteristics of semistructured data are, in order to decide how they could be handled by possible database management systems, e.g. objects may have multiple occurrences of the same attribute (Deutsch 4).

Another important subgoal with this project is to be able to constitute in the end what the advantages and disadvantages of the systems are, at which point they may overlap in structure or performance and thus what the motivation could be to prefer one DBMS to another.
In this regard it would be interesting to see if the project could reveal whether the existing DBMSs for semistructured data turn out to be applicable to different areas.
Methodically, the project is to a great extend based on theory, i.e. comparing the systems based on the available literature. However, if possible the DBMSs, STORED, XTASY, Index Fabric and Lore, shall be tested pratically in order to obtain more accurate data for the comparison.

2 Literature Review

Database management systems have been around since the earliest days of computing in the 1960s and research within this field can be stated ever since, leading to the development and improvement of various database models and their respective database management systems: Relational Database Management System (RDBMS), Object-Oriented Database Systems or Hierarchical Database Management System, to name some of the most influential.
Parallel, semistructured data has emerged as an important topic in the past (two) decades. Semistructured data, as for instance defined by Abiteboul (2000:11), has no separate description of the structure of the data (e.g. in form of a schema), but rather the data is directly described by using simple syntax. In other words, according to Abiteboul (2000:11) and Cooper et al. (2001:1), semistructured data is often represented as an edge-labeled graph where the nodes represent data elements and the labeled edges their relation. This is called a self-describing relationship structure which replaces and is to some extent opposed to the (clearly defined) schema in a traditional structured database system.
Now based on the studied literature, XML appears to be *the* format (Deutsch et al. page 1), respectively a popular syntax (Cooper et al. 2001:2) for semistructured data. Due to the fact that XML is considered to play an increasing role in the exchange of data (W3C, Goldman et al. 1), Deutsch et al. (page 1) even expect the availability of semistructured data to increase.
This is the point where the two fields of research have started to converge. According to Buneman (1997:1) database management engineers would like to treat data sources like the Web as a database. The problem, however, is that the electronic data to a great extent lies outside of the scope of the traditional database management sytems, since it is mostly not controlled by a schema, having an irregular, unknown or frequently changing structure (Suciu 2000:10, Cooper 2001:1). That is what the research area in semistructured (XML) data management builds upon.
Subsequently an overview is given about those database management systems that were specifically designed to express, store, respectively manage semistructured data. While sharing the same development goal, the approaches and build-ups are quite different, as shall be outlined briefly:

- XML database model:
  - **Xtasy**: ”Xtasy is a database management system for XML semistructured data [. . .] where each data source is associated with an XML document [. . .] The target query language supported by Xtasy is the long-awaited W3C standard query language XQuery.” (SARTIANI 2003:35ff.)
- Semistructured data(base) model: Object Exchange Model (OEM) (ABITEBOUL (2000:19) is a model for exchanging semistructured data between object-oriented databases. It serves as the basic data model in numerous projects of the Stanford University Database Group:
  - **Lore** (GOLDMAN et al. and DEUTSCH et al.): stores the schema with the data, i.e. the data is stored as a graph while the schema is stored as attributes labeling the graph’s edge. This provides a certain flexibility required by semistructured data, but entails space and time cost because the schema is replicated and processed for each item. RDBMS cannot be used in this approach. Lorel is the analog query language.
- Systems that make use of Relational Database Management Systems:
  - **STORED** terms a declarative query language and is derived from **S**emistructured **TO** **R**elated **D**ata. It provides an approach to include RDBMS by performing a mapping from the semistructured to the relational data model. Compared to other query languages for semistructured data STORED is more restricted in its use (DEUTSCH 3).
  - **Index Fabric** stands for an indexing structure which is implemented on top of a RDBMS (COOPER et al. 2001).

As it was shown, various DBMSs for semistructured data exist. In the course of their development the researchers show in respective papers, that they are very well aware of other proposed implementations (DEUTSCH page 1, COOPER et al. 2001:9). They also compare their experimental results to other systems, e.g. COOPER et al. compared ”the performance of queries using the DBMS’ native B-tree index versus using the Index Fabric implemented on top of the same database system.” (2001:6) However, no extensive comparison[1], but rather partial comparisons with regard to certain aspects, e.g. storage or query language performance, can exclusively be found. Just as little, well-grounded and detailed information about how those systems relate to one another concerning possible advantages or disadvantages are given. The intended project work aims at giving such a comparative presentation.

3 Scope of the Project

It has to be stated that no exhaustive comparison of all DBMSs for semistructured data that may exist can be given. Hence, only those projects were considered which were explicitly declared as database management systems by the developers. Other projects, as for instance Tsimmis, which is ’only’ similar, focusing on integrating heterogeneous data sources (ABITEBOUL 2000:19), can thus not lie within the scope of this project. The DBMSs chosen for this project: STORED, XTASY, Index Fabric and Lore, are

[1] Geared to the aspects mentioned in the Introduction

the ones which were most frequently mentioned in the studied literature. Above that it is seen as a compensation that those systems are depicting quite different approaches, as it was briefly described in the previous section. The chosen systems are therefore considered to give an adequate representation of relevant DBMSs.

4 Anticipated Result

The anticipated result of this project is a detailed and comparative report about four quite different approaches to implement database management systems for semistructured data. By pinpointing to advantages, disadvantages, similarities and differences the paper is meant to serve as a reference for future research which aims at improving state-of-the-art DBMSs for semistructured data.

5 Procedure and Methodology

The proposed project is meant to be carried out in conjunction with the Master's Thesis in Spring 2012. About four months, week 3 to 22, are set as a time span in order to carry out the project. Before the project work is to be carried out, the courses database technology 1 and 2 are planned to be taken to account for an adequate theoretical background.

Subsequently a planned timeline of milestones is given.

5.1 Milestone 1: Planning and Pre-study

Planned: 2012-01-28, duration: 1-2 weeks

The approximate time plan given in the Project Proposal should be reviewed and if necessary adapted. The research groups of the respective DBMSs should be contacted in order to find out if any recent changes were made and if the prototypes could be possibly used for testing. Additionally, meetings with the supervisor should be scheduled.

5.2 Milestone 2: Literature Review

Planned: 2012-02-25, duration: 3-4 weeks

In order to gain a comprehensive knowledge and solid background of the research area, accurate study and review of the literature is crucial. This implies to get acquainted with the key concepts of the project in general, e.g. database management systems, semistructured data and XML. In more detail, the DBMS for semistructured data on which the focus lies need to be embedded in the general context.

5.3 Milestone 3: Comparison Guidelines

Planned: 2012-03-04, duration: 1 week

To be able to compare the systems properly it needs to be outlined in detail which criteria need to be considered, respectively are of importance to obtain a valuable comparison. In the introduction, some aspects are already mentioned. Still, by systematically scanning the literature for reference points that are used to scale for example the performance of database management systems, their respective query languages etc., this list may be extended to be as complete as possible.

5.4 Milestone 4: System Analysis (and Testing)

Planned: 2012-04-08, duration: 4-5 weeks
Within this time span a deep and all-embracing understanding of the DBMSs in question needs to be acquired. This implies to outline the systems, possibly in form of UML-diagrams, and to gather information as to how the systems perform according to the defined reference points. Most of the analysis is planned to be based on referring to the description of the systems published by the respective researchers. It is above that tried to procure material about the systems, preferably from the developers themselves, which is as up to date as possible (This is mainly due to the fact that all information on Xtasy, STORED, Index Fabric and Lore found so far, are about 8 to 10 years old, going back to when the systems were implemented). To be able to account for comparable data, it would be optimal to make use of the existing prototypes and test for instance the performance of the query languages on equal data. Whether this will be possible cannot be stated at this point but will be handled within the planning and pre-study phase as mentioned above.

5.5 Milestone 6: Comparison

Planned: 2012-05-06, duration: 3-4 weeks
Within this time span the results of how the systems perform with regard to the reference points shall be compared, by giving detailed descriptions of the advantages and disadvantages of each system, depicting in which respect they may overlap, e.g. in structure, query language performance etc. and which conclusions can be drawn from that.

5.6 Milestone 7: Round-up

Planned: 2012-06-03, duration: 3-4 weeks
Within this last phase the main focus will lie on completing the thesis report, i.e. write out thoughts in full and making sure it is proofread by at least two to three people.

5.7 Milestone 8: Thesis Report and Supervision

Planned: 2012-06-03, duration: 12-13 weeks
After having finished the literature review phase, one should start to collect thoughts, results etc. and write the actual thesis report, while simultaneously performing the above stated steps. Additionally the scheduled meetings with the supervisor are to be attended regularly, for which a short summary of recent preceedings is to be prepared.

6 Budget

Due to the fact that the work on this paper will be mainly based on theoretical research no extraordinary expenses are scheduled.

References

[1] **Abiteboul**, Serge, Peter Buneman and Dan Suciu (2000): Data on the Web. From Relations and Semistructured Data and XML.
<http://books.google.de/books?hl=de&lr=&id=ZaEOesNag24C&oi=fnd&pg=PA9&dq=reading+list+semi-structured+%28xml%29+data+management&ots=Ifdq7uhL22&sig=ECpbzFB6EZ0n_h_eY6LKzwWa23g#v=onepage&q&f=false>. 2011-05-17.

[2] **Buneman**, Peter (1997): Semistructured Data.
<citeseer.ist.psu.edu/viewdoc/download;jsessionid...?doi=10.1.1...>. 2011-05-17.

[3] **Cooper**, Brian F. (2001): A fast index for semistructured data.
<www.scaledb.com/pdfs/FastIndexToSemistructured.pdf>. 2011-05-17.

[4] **Deutsch**, Alin, Mary Fernandez and Dan Suciu: Storing Semistructured Data with STORED.
<nfolab.stanford.edu/lore/pubs/xml.pdf>. 2011-05-16.

[5] **Freire**, Juliana (2009): CS7960 - Research Topics in Databases - Spring 2009.
<http://www.cs.utah.edu/~juliana/rtdb2008/>. 2011-05-15.

[6] **Goldman**, Roy, Jason McHugh, Jennifer Widom: From Semistructured Data to XML: Migrating the Lore Data Model and Query Language.
<nfolab.stanford.edu/lore/pubs/xml.pdf>. 2011-05-16.

[7] **Sartiani**, Carlo (2003): Efficient Management of Semistructured XML Data.
<www.di.unipi.it/~sartiani/papers/thesis.pdf>. 2011-05-16.

[8] **Suciu**, Dan (2000): Semistructured Data and XML. In: Katsumi Tanaka, Shahram Ghandeharizadeh, Yahiko Kambayashi: *Information Organization and databases. Foundations of Data Organization.*

[9] **W3C** (World Wide Web Consortium).
<http://www.w3.org/XML/>. 2011-05-16.