PhD thesis proposal

Data Warehouses in the Path from Databases to Archives

Supervisors: Gabriel David, Cristina Ribeiro

FEUP / INESC-Porto

2008-01-18

Introduction

Organizations are increasingly relying on databases as the main component of their record keeping systems. However, at the same pace the amount and detail of information contained in such systems grows, also grows the concern that in a few years most of it may be lost, when the current hardware, operating systems, database management systems (DBMS) and actual applications become obsolete and turn the data repositories unreadable. The paperless office increases the risk of losing significant chunks of organizational memory and thus harming the cultural heritage.

Significant research addressing this concern has already been conducted. The conclusions discard approaches now considered naive like trying to preserve specimens of the machines, system software and applications, in all their main versions, so that the backups of every significant system could be used whenever needed. A variant of this, instead of preserving the hardware, suggests simulating the older hardware in newer machines. More promising research suggests the conversion of database contents into an open neutral format with a significant amount of semantics associated (XML dialects), so that it becomes independent of the details of the actual DBMS.

The present proposal stems from this principle but tries to go a step further based on the following observation: there is a parallel in the attitude of a data warehouse designer approaching a database-centred operational information system (IS) to specify a data warehouse (DW) and an archivist analysing a document-centred organizational IS to specify an archiving policy and system. Both search an integrated model of the organization, merging information from a diversity of sources, systems and technologies, both have a process-centric methodology, specifying data marts or classifying related series of documents, both have long-term validity and integrity requirements, both have an evaluation attitude, leaving out irrelevant details in the data or in the series of documents to concentrate on the essential, both want to build an archive which remains basically unchanged, except for the addition of newer data or

documents, and both want to expose the respective information contents in a simple and systematic way.

Of course there are differences, first of all in the respective goals. The DW designer usually tries to answer the information needs of the organization management from the point of view of decision support, monitoring, trend analysis and forecast, while the archivist wants to preserve the memory of the organization and its processes, for future generations. So concrete decisions on evaluation and elimination procedures may differ, according to the specific requirements, but the general working framework seems similar.

Thesis

Following this basic intuition, the research question proposed is to explore the adequateness of the DW approach as a target vehicle to perform, with respect to a given IS, the functions considered essential from an archivist viewpoint like appraisal, classification, elimination, description, and access while respecting properties like authenticity and integrity.

Context

Archivist methods have been put under great strain by the growing amount of institutional information that is being stored in digital format. Although the goals and principles of the discipline remain solid, their application to the new supports and information structures is subject of much debate [1,2,3].

The domain of research in this proposal is archiving data records produced as a result of the regular activity of an institution, which are usually kept by a DBMS.

At first, dominant approaches to the preservation of electronic records were either too much biased to the paper-based procedures or unfeasible: the projects of preservation of both the data and the tools to read it in the original format, including hardware and the several layers of software (operating systems, DBMS, applications), in the appropriate versions, despite its value from a museum perspective, were soon recognized as an unfeasible solution to the archivists goals. The same happened to the projects betting on simulating the old machines. The third alternative is collectively labelled as migration, though it has many trends.

Two research areas are essential for this proposal. The first has focused on the description of documents to render them available for retrieval, across the frontiers of domain modelling, document nature and storage technology, and has grown along with the Web. The second has roots on the concerns of archivists with the fragility and opacity of digital materials, and has a broader research agenda fuelled by the needs of organizations and increased awareness of the need for new approaches.

From a pragmatic viewpoint, the first line has originated several standards to add metadata to current digital objects (Dublin Core in the case of generic objects, RDF for Web objects and EAD for digital archive objects) and thus contribute with partial solutions to the description problem. However, the general goal of building digital archives needed a more comprehensive treatment which resulted in the OAI standard[4].

The second line has the more ambitious goal of dealing with the general problem of preserving the wealth of information that is being generated in digital form or converted to it. Several projects have been dealing either with fundamental models for integrating preservation into the management of current records or with solutions to the concrete problems that arise in a specialized domain [5].

One of the basic solutions is to serialize the database, systematically storing the data dictionary (table names, columns, integrity constraints) and the actual values of each column in each table line. This way, one is able to record all the values of every database with a single archiving model. The main problem of this approach is that it forgets that the data is just part of the problem in a database system. Most real information systems are structured in three layers: data + business rules + presentation. If the presentation layer may contain not too much knowledge, both the data and the business rules layers keep their own part of the semantics of the data. In certain cases, the values are meaningless without the code that discloses their interpretation. The solution envisaged is to perform a previous step of eliciting implicit knowledge in the application code and storing it as explicit columns in a new data model. This operation is a typical step in a DW design process.

The proposed approach

The process of specifying a DW can be used for building an information asset that is both faithful to the original data and organized in a way that can be given use in the future without the complexity of the original system. Although the DW primary intent has been to support management decisions with flexible and relevant data (a goal more typical of current archives), the tendency to add more and more data to it turned DW into the most valuable repository of large organizations. Understanding this, recent recommendations on DW design [6] stress the importance of a global, process-centred analysis of the organization, resulting in the clear establishment of a DW bus architecture where dimensions, a kind of authority files, are initially identified and then reused in the several stars that constitute the dimensional model. The data considered relevant includes all basic facts, to enable arbitrary future queries. The knowledge of techniques to deal with changing dimensions without information loss is also central to give the DW an archival value.

Research Goals

The research explores the application of DW technology to the preservation of complex electronic records.

In more detail, the research will study:

- **properties** that the DW must possess, according to the standards already established (like the Open Archives Initiative);
- **rules of transformation** from operational systems into DW, adopting a process-centric but integrated view, which guarantees those properties;

- a **XML version** of the DW, though the DW model is already very simple and thus fulfils the requirement of platform independence required by long-term preservation;
- **metadata** needs, beyond the data in the operational database, to preserve the meaning of the processes, most often just implicit in the data but present in organizational procedures and in software development documentation;
- **application** to a real system of a concrete institution, playing the role of a case study;
- **assessment** of the results obtained with the help of external experts.

The research is expected to have impact in the area of archiving automated information systems in the same measure of the generality and simplicity of the found solution. An immediate benefit would be the migration of a complex database into an archival format. In a more general setting, the research would contribute to preserve the digital components of organizational memory for future exploration.

Data warehouse for archives model

DW design methodologies were initially derived from the goal of producing decision support systems. However, the maturity of this area along with the increasing computational power available led to more conservative design principles, based on the idea of storing all the basic facts in their finest granularity. The proposed DW bus architecture stresses a global analysis of the organization to identify in the first place all the relevant dimensions, i.e. the authority files and other entities, and then the main processes that will correspond to the facts tables. This very simple star schema, with a central fact table surrounded by a set of dimensions, is at the heart of the dimensional model and of the whole idea of DW. The goal is to organize the information in a simple and systematic model, so that optimized search and processing is possible.

In this research direction, additions to the basic model should be investigated, in particular which dimensions and attributes are mandatory or recommended from the viewpoint of an archivist, which classes of attributes are required in the fact tables, which mechanisms of evolution for dimensions are acceptable, and in general which properties must the DW possess, according to the standards already established (like the Open Archives Initiative).

A set of typical organizational situations should be gathered and example solutions to the corresponding modelling problems provided.

Transformation process

After the definition of the target DW model and its implementation in a concrete DBMS, it is necessary to define the whole process of populating it with the data coming from the available sources. Conceptually, the central problem of this research proposal is to get data from the operational IS in the organization and transform it into the DW.

At this level, the problem is a mapping between models. The conclusion of this study will give rise to a set of mapping rules.

Implementing the rules will correspond to concrete extraction, transformation and loading processes where data will be selected, perhaps eliminated, cleaned, formatted according to the standards, checked for referential integrity against dimensions fulfilling the role of authority files, and transferred. These process definitions will also be gathered.

Taking into account the experience of archivists in this task is essential as it goes through the phases of appraisal, selection, elimination and the automated part of description, typical of the archivist process. The system must support the addition of possible bits of manual description.

This task is the main step in the migration approach proposed. Data is not only subject to a technological conversion, as in many migration proposals, but also to a controlled and documented reorganization that is necessary for the survival of the information. The XML version of the data results from a second step of migration, this time to ensure portability and long time preservation of the information but performed essentially at a technological level and keeping its structure unchanged.

XML model

Although the DW model is already very simple and thus fulfilling the requirement of platform independence required by long-term preservation, a XML version of the DW will be established, less suited to the arbitrary querying the DW provides, but more adequate to information exchange with other systems. XML is becoming a lingua franca for automatic but still human readable information exchange, and so the ability to bidirectional conversion between XML and the DW is unavoidable. Some will even prefer XML for long-term preservation for the sake of uniformity with other kinds of digital objects preservation, but with reduced ability to offer access services to the users.

Metadata requirements

The design process of a DW typically includes the definition of metadata describing the exact process of obtaining a certain fact, the source, the loading date, the tool version used for it, etc. These elements will probably be kept in the extended dimensional model. However, a lot more contextual information will be found mandatory to add more semantics to the data. For instance, the operational system UML or E-R models may be considered important, as well as a data dictionary, explaining the meaning of each attribute in each dimension and fact. Other contextual aspects concerning the organization around the IS and its evolution, the implicit or explicit definition of processes, forms, data sheets etc. may also be important to improve the understandability of the data collected.

A second set of metadata elements will derive from the adaptation of the would be archivistic processing of the electronic records. This includes information on the corresponding appraisal, selection, elimination, and description activities. In particular, one will investigate the implications of the ISAD(G) and ISAAR(CPF) standards on the description of such records and devise ways of extracting parts of that metadata during

the ETL processing. It will, for instance, explicitly state the institutional producer of a certain record, that stands for the corresponding fact, and under which mandate he is responsible for the fact.

Conclusion

To carry on the agenda of this proposal interdisciplinary knowledge must be gathered, with know-how in the domains of archives, information systems and data modelling. Previous experience in developing models and prototypes of databases with a strong component of metadata, according to the International Standard of Archival Description (ISAD)[7] may prove useful. This work will fit well in the current project PTDC/CCI/73166/2006 *DBPreserve - Data Warehouses for the Long-term Preservation of Institutional Electronic Records and Databases*.

REFERENCES

[1] *INTERPARES Project - Preservation Task Force Final Report*. University of British Columbia. 2001. <http://www.interpares.org>

[2] *Records Continuum Research Group*. Recordkeeping Metadata Project,

http://www.sims.monash.edu.au/research/rcrg/research/spirt/about.html

[3] Digital Preservation Coalition. http://www.dpconline.org/graphics/

[4] OAI- Open Archives Initiative. http://www.openarchives.org/

[5] Margaret Hedstrom. The Digital Preservation Research Agenda. Proceedings of the Conference on The State of Digital Preservation: An International Perspective.[2002]

[6] Ralph Kimball, Margy Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition). John Wiley & Sons, 2002.

[7] Cristina Ribeiro, Gabriel David and Catalin Calistru. A Multimedia Database Workbench for Content and Context Retrieval. Proceedings of the 2004 IEEE International Workshop on Multimedia Signal Processing. IEEE Press 2004.