ABSTRACT

Data warehouses are used in many application domains, and there is no established method for their preservation. A data warehouse can be implemented in multidimensional structures or in relational databases that represent the dimensional model concepts in the relational model. The focus of this work is on describing the dimensional model of a data warehouse and migrating it to an XML model, in order to achieve a long-term preservation format. This chapter presents the definition of the XML structure that extends the SIARD format used for the description and archive of relational databases, enriching it with a layer of metadata for the data warehouse components. Data Warehouse Extensible Markup Language (DWXML) is the XML language proposed to describe the data warehouse. An application that combines the SIARD format and the DWXML metadata layer supports the XML language and helps to acquire the relevant metadata for the warehouse and to build the archival format.

DOI: 10.4018/978-1-4666-2669-0.ch008
INTRODUCTION

The technological generation in which we live has gradually modified the method to create, process, and store information, using digital means for this purpose. The institutions, enterprises, and governments rely more and more on information systems that increase the availability and accessibility of information. These information systems typically require relational databases, which become valuable assets for those entities.

However, rapid technological changes degenerate into rapid obsolescence of applications, file formats, media storage, and even databases management systems (DBMS) (Date, 2004). If nothing is done, access to large chunks of stored information may become impossible and it will eventually be lost. So, it is important that entities which have major responsibilities in preserving information in digital form become aware of this problem and join initiatives all over the world, seeking for the best methodology for long-term digital preservation, and in particular for database preservation.

The work presented here has been developed in the context of DBPreserve, a research project funded by the Portuguese Foundation for Science and Technology (FCT), in collaboration with INESC Porto, University of Minho, and the Portuguese National Archives (DGARQ). The project goal is to study the feasibility of data warehousing technologies to preserve complex electronic records, such as those constituting databases. The DBPreserve project approaches the long-term preservation of relational databases issue with a new concept, a two-step migration:

- A model migration from the relational model to the dimensional model, using data warehouse concepts to simplify the model simplification and increase efficiency (Rahman, David, & Ribeiro, 2010);
- An XML migration from the dimensional model to an XML (Consortium, 2008) format that represents the data warehouse, to ensure a long-term preservation format.

A data warehouse is structured by star or snowflake representations. A star is made up of a fact table that stores the facts, and dimensional tables that contextualize the facts. There are also bridge tables used to resolve a many to many relationship between a fact table and a dimension table, or to flatten out a hierarchy in a dimension table. A snowflake is similar to a star but the dimension tables have been subject to a partial normalization, resulting in subdimensions. Data marts are subsets of a data warehouse.

We propose the Data Warehouse Extensible Markup Language (DWXML), an XML dialect for describing a Data Warehouse (DW) (Inmon, 2002; Kimball & Ross, 2002; Date, 2004). It has been defined and refined according to data warehouse’s properties and tested using a case study of SiFEUP1. It is used in the project as a complement to the SIARD format (Archives, 2008) used for the description and archive of relational databases. This enrichment leverages past efforts to define an archive format suitable for data tables from databases and adds a layer of metadata for the data warehouse components.

BACKGROUND

Digital preservation concerns sustainable and efficient strategies for the long-term preservation of digital objects (Ferreira, 2006). However, databases and data warehouses are different from conventional digital objects as they have an internal structure, and include schemas and integrity constraints, which are vital for interpreting data.
Related Content

**ROL2: Towards a Real Deductive Object-Oriented Database Language**
[www.igi-global.com/chapter/rol2-towards-real-deductive-object/35859?camid=4v1a](www.igi-global.com/chapter/rol2-towards-real-deductive-object/35859?camid=4v1a)

**Content-Based Publish/Subscribe for XML Data**
[www.igi-global.com/chapter/content-based-publish-subscribe-xml/41506?camid=4v1a](www.igi-global.com/chapter/content-based-publish-subscribe-xml/41506?camid=4v1a)

**Using Rule-Based Concepts as Foundation for Higher-Level Agent Architectures**
[www.igi-global.com/chapter/using-rule-based-concepts-foundation/35872?camid=4v1a](www.igi-global.com/chapter/using-rule-based-concepts-foundation/35872?camid=4v1a)

**Describing and Extending Classes with XMI: An Industrial Experience**
Giacomo Cabri, Marco Iori and Andrea Salvarani (2005). *Software Evolution with UML and XML* (pp. 352-393).
[www.igi-global.com/chapter/describing-extending-classes-xmi/29619?camid=4v1a](www.igi-global.com/chapter/describing-extending-classes-xmi/29619?camid=4v1a)