Relational Databases Digital Preservation

Ricardo André Pereira Freitas[1] and José Carlos Ramalho[2]

[1] Faculty of Science – University of Porto
Porto – Portugal
[2] Department of Informatics - University of Minho
Braga – Portugal
rfreitas@fc.up.pt, jcr@di.uminho.pt

Abstract. Digital preservation is emerging as an area of work and research that tries to provide answers that will ensure a continued and long-term access to information stored digitally. IT Platforms are constantly changing and evolving and nothing can guarantee the continuity of access to digital artifacts in their absence.
This paper focuses on a specific family of digital objects: Relational Databases; they are the most frequent type of databases used by organizations worldwide. A neutral format that is hardware and software independent is the key to achieve a standard format to use in digital preservation of relational databases. XML for its neutrality was chosen for this representation of the database.
The presented solution offers a possibility to achieve relational databases preservation. The prototype follows the "Reference Model for an Open Archival Information System" (OAIS).

Key words: Digital Preservation, OAIS, Relational Databases, XML, Digital Objects, Significant Properties

1 Introduction

Nowadays, due to the constant evolution in the hardware and software industry more and more of the intellectual and business information are stored in computer platforms. The main issue lies exactly within these platforms. If in the past there was no need of mediators to understand the analogical artifacts today, in order to understand digital objects, we depend on those mediators (computer platforms). Nothing can guarantee the continuity of access to digital artifacts in their absence [12]. A new problem in the digital universe arises: Digital Preservation.

Accessing the information does not mean a simple access to the bits that all digital objects are made of, rather it means an access to the information understanding what's there. Although digital information can be exactly preserved in its original form by only copying (preserving) the bits, the issue appears when we notice the very fast evolution of those platforms (hardware and software) where the bits can be transformed into something human intelligible [9]. Digital archives are complex structures that without the software and hardware –

which they depend on – the human being, or others, will certainly be unable to experience or understand them [8].

Our work addresses this issue of Digital Preservation and focuses on a specific class of digital objects: Relational Databases [9]. Relational databases are a very important piece in the global context of digital information and therefore it is fundamental not to compromise its longevity (life cycle) and also its integrity, liability and authenticity [15]. These kind of archives are especially important to organizations because they can justify their activities and give us a glimpse about the organization itself. What kind of organization does not have its information system based on IT platforms? So the question is, how can we ensure access for a long-term, to the information of a relational database and understand what's there? The information must be interpretable for those who demand.

2 State-of-the-art

There are known approaches to the problematic of digital preservation – technology preservation, emulation, migration, normalization, encapsulation [12] [18] [20], and more. We intend to research and study the problematic within digital preservation but focusing on a specific family of digital objects: Relational Databases. Before going further lets characterize these digital artifacts.

A database can be defined as a set of information that is structured. In computing, a database is supported by particular program or software, usually called the Database Management System (DBMS), which handles the storage and management of the information. In its essence a database involves the existence of a set of records of data. Normally these records give support to the organization information system; either at an operational (transactions) level or at other levels. For example, obtaining knowledge to help in decision support (Data Warehousing Systems).

The structure of relations and relationships between entities within a database depends on the type of the used model. Our study focuses on the relational model, widely available and certainly the most used. However there are other logical models for databases: the flat model, the hierarchical model, object-oriented model, among others [21].

In digital preservation it is fundamental to establish the significant properties that should be preserved for each class of digital objects. We will try to achieve some consensus over these issues and then analyze some of the current projects on this field of research.

2.1 The Significant Properties

In general the significant properties in a digital object are those that are identified by its community of interest.

Information that indicates the original operating system and the DBMS that used to support the database is important to characterize the environment of the original database. The date of creation of the database and identification of

its creator should also be preserved. This information is identified as technical metadata.

The information in a relational database has a particular structure based in relations usually called tables [5]. Lee Buck [2] and Ronald Bourret [1] on their approaches concerning XML and Databases do not mention any information about the database structure. However, the structure may provide a way of interpreting the data in order to work and extract valid information – knowledge. On one hand we have the data stored in the database and on the other hand its structure. The data contained in the records of the database obviously has to be preserved but through this analysis we conclude that it will be necessary to also preserve the structure of the database [15]. Some structure features considered important for preservation are:

- **TABLES** (Name)
- **COLUMNS** (Name, Type, Size, ...)
- **KEYS** (Primary keys, Foreign keys, ...)

By preserving these elements we are able to preserve all the database structure – relations (tables) and the relationships between them.

There are other features in a database, such as triggers, stored procedures, forms, or other application issues that we should consider whether or not to preserve. These elements differ from the previous ones since they represent the database semantics. Depending on what is considered significant to preserve we may choose to preserve these features or not. If we choose to preserve application issues, such as a form that interacts with the database, it may be enough to preserve its code or it may be necessary to preserve an image of its appearance.

2.2 Current Research

Considering the nature of the digital artifacts that we are addressing – relational databases – there is an European strategy encompassed in the "Planets Project" [13] to enable their long term access. The project adopted the SIARD [17] solution, which is based on the migration of database into a normalized format (XML – eXtensible Markup Language [22]). The SIARD was initially developed by the Swiss Federal Archives (SFA).

Another approach, also based on XML, relies on the main concept of "extensibility" – XML allows the creation of other languages [16] (it can be called as a meta language). The DBML [10] (Database Markup Language) was created in order to enable representation of both **DATA** and **STRUCTURE** of the database. The following diagram (Fig. 1) reflects the schema for this language.

Both approaches (SIARD and DBML) adopt the strategy of Migration of the database to XML, why? A neutral format that is hardware and software (platform) independent is the key to achieve a standard format to use in digital preservation of relational databases. This neutral format should meet all the requirements established by the designated community of interest.

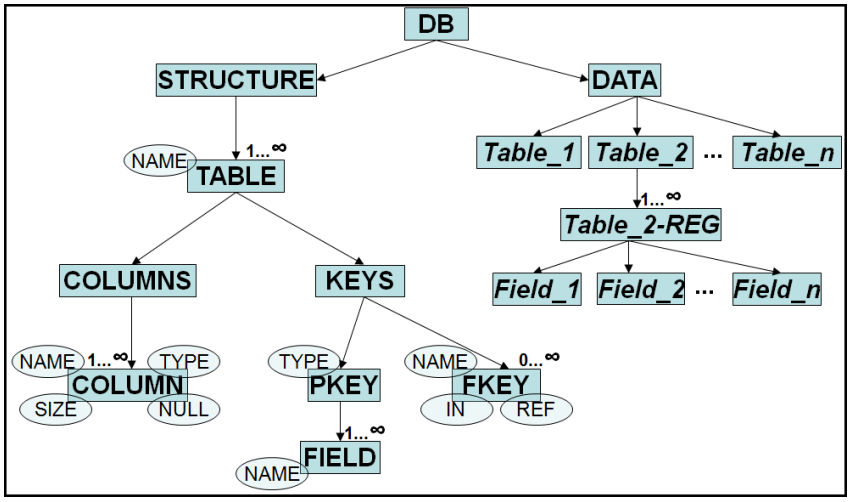

Fig. 1. DBML Schema

2.3 Open Problems

The usage of a normalized format is accepted as being the answer to the problematic of preserving relational databases [10] [17]. By doing this, it is possible to separate the data from its specific database management environment. However, there are still some issues not answered: what significant properties should be preserved? Should the database semantics be preserved? If so, how? How can we ensure authenticity? And how to ensure preservation during the lifecycle of the database (while it evolves)? These are just a few questions that emerge upon this approach.

In order to search for answers on these issues and to pursue other questions or solutions that may emerge, a case study was developed.

3 Possible Solution

Concerning the preservation of relational databases, we adopted an approach that combines two strategies and uses a third technique: migration and normalization with refreshment [9]. The main strategy in our approach is Migration which is carried in order to transform the original database into the new format – DBML [10]. The normalization is crucial to reduce the preservation spectrum to only one format. A third technique (refreshment) will also be needed. The refreshment consists on ensuring that the archive is using media appropriate to the hardware in use throughout preservation [9].

In this case study we used a database on which is not expected any more transactions from the operational point of view. We decided to freeze the database in order to preserve it.

In figure 2 it is presented a portion of code extracted from a DBML document produced by a prototype used in the case study.

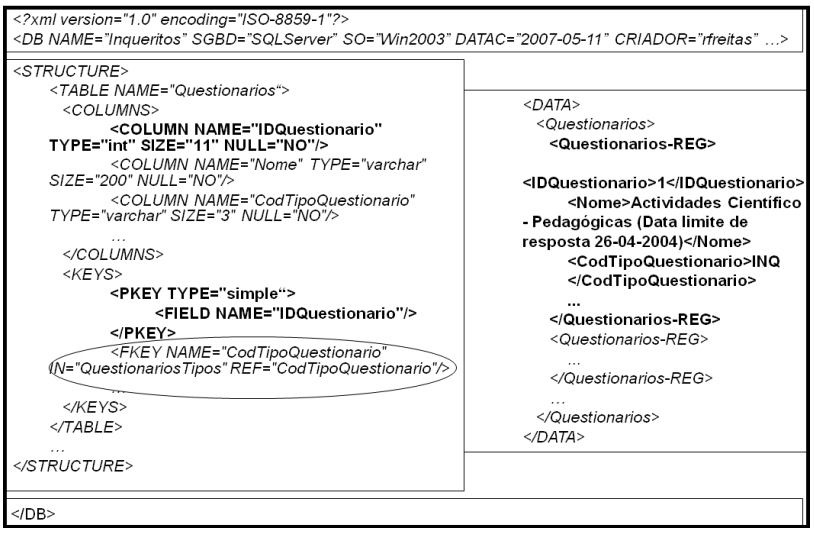

Fig. 2. DBML portion of the document – case study

After this brief example of the document and its format used to archive the database, we are able to analyze the archive focusing on the system architecture and the workflow of the database from its ingestion until its dissemination.

3.1 System Architecture

We will now seek to describe and analyze the architecture of the implemented system. The prototype is based on a web application with multiple interfaces. These interfaces have the mission to take a certain database and ingest it into the archive. The access to the archive in order to do all the necessary interventions on the system will also be done through those web interfaces.

Conceptually, the prototype is based on the OAIS [4] reference model. The OAIS model of reference does not impose rigidity with regard to implementation, rather it defines a series of recommendations.

The OAIS model of reference is concerned about a number of issues related to digital preservation: the process of information Ingestion into the system, the information storage as well as its administration and preservation, and finally information access and dissemination [6] [11].

However, the OAIS model does not impose any computer platforms, development language, database management systems (DBMS), interfaces, i.e., does

not condition the development of the system at the technological level involved. Instead, the model acts as a guide for those who wish to develop digital archives [4]. Figure 3 shows a conceptual design of the OAIS reference model.

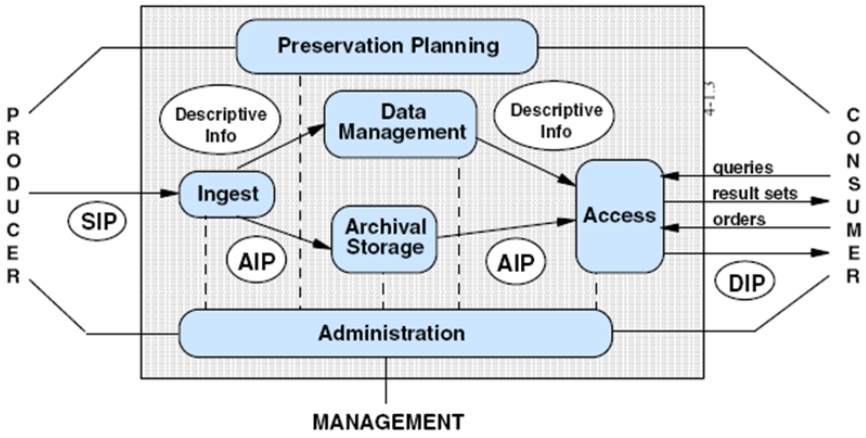

Fig. 3. OAIS Functional Entities. (Courtesy of Consultative Committee for Space Data Systems. "Reference Model for an Open Archival Information System (OAIS) – Blue Book," National Aeronautics and Space Administration, Washington, 2002) [4]

Three information packages are the base of the archival process: Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP). Before ingestion begins, it is necessary to establish a Submission Agreement between the Producer and the Archive. Before dissemination starts, an Order Agreement between the Archive and the Consumer is established. Through these agreements both SIP and DIP constitutions are defined well as the specifications of the sessions for data submission and dissemination. The Administration component is responsible to define the AIP constitution – package that will be stored.

The SIP is composed by both descriptive and technical metadata and the digital content itself. The ingestion process includes the SIP validation delivered by the Producer. If the minimum requirements are achieved the package is ready to be archived. After this, the package is transformed into an AIP. At the other end of the archive the Consumer may query the OAIS trying to find the desired data. When the information is found the Consumer will issue a request to the system that will respond with DIPs – packages used in the dissemination process.

Inside the archive, the administration component manages the AIPs and participates in the ingestion and dissemination process. The preservation component is responsible for implementing preservation policies. Our prototype follows this conceptual model.

The Prototype implementation was a crucial phase of this work. We intend to implement a system capable of ingesting databases, in the form of information packages (DBML + metadata), for preservation. The developed system is based on a Web application and has multiple interfaces that allows not only the ingestion of information, but also its administration, preservation and dissemination. Figure 4 gives an idea of the preservation process over the prototype.

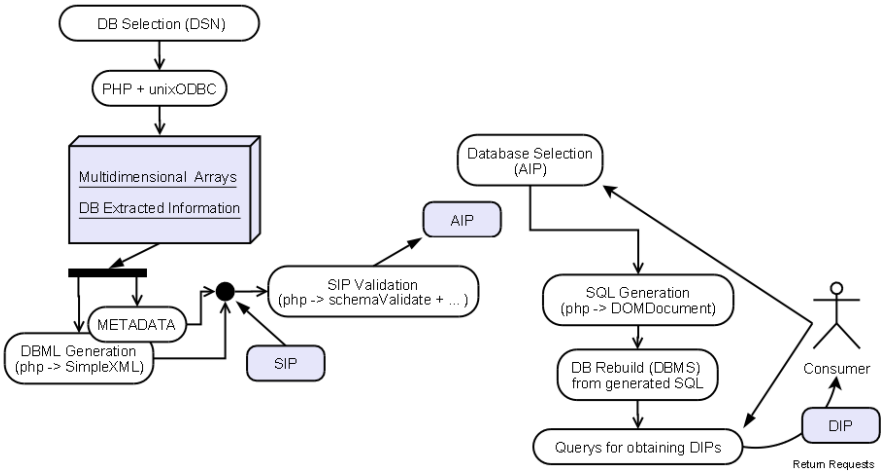

Fig. 4. Database preservation workflow

The several Web interfaces can only be accessible through a previous authentication on the system. The administration component manages these requests, and the various privileges with regard to the handling of information in the archive. The users of the system can be divided into two types, namely two profiles of users: administrators and users. A user has at his disposal the following list of operations:

- Creation of SIPs
- Ingestion of SIPs in the repository (AIPs)
- Consultation of the state of the repository
- Production of SQL [3] from AIPs
- Dissemination of DIPs

The administrator has at its disposal all operations available to users with the addition of operations associated with administration tasks:

- Management of users of the system
- Direct access to the file system (Repository directory)
- Manipulation of drivers (unixODBC) for connection to databases
- Monitoring of the state of obsolescence of information in the repository

– Actions of migration and refreshment of the stored information whenever necessary (preservation policies)

The practical implementation has proven to be very significant because it was possible to obtain interesting results. After the initial assembly of all the tools necessary to perform the tasks required, we quickly started to develop the prototype. Some adjustments were made to obtain better performance of the system.

It is important to refer that this work aimed to test the feasibility of relational database digital preservation using this approach. This was indeed possible, i.e., the objective of converting relational databases (different DBMS) into DBML was achieved. We were also able to rebuild the database in a DBMS from the DBML document in order to achieve the database dissemination.

4 Conclusion and Future Work

Different strategies (refreshment, migration and normalization) were combined to pursue the goal of digital preservation. The main strategy used is migration such as in the European project PLANETS. These strategies were used according to the class of digital objects that we addressed – Relational Databases. If the goal was the implementation of a repository for other family of digital objects the strategies may differ [7] [19].

Considering the properties (significant properties) of the database that we establish for preservation, the presented solution provides good results. However, in future work, we should be able to establish a consensus on several issues. Some of them are:

– walk further to determine the significant properties of a relational database and for each one of them or globally define the strategy that should be adopted;
– compare alternative strategies;
– how is it possible to ensure authenticity?
– how can we ensure preservation during the lifecycle of the database (while it evolves)?

We should study these issues by evaluating if the significant properties of a database are well preserved. We will also need to test if a preservation framework offers the possibility to query different versions of the database. Another important issue is, will the preservation framework be able to deal with hundreds of archived databases containing terabytes of data? During the research and framework improvement, we should be able to redefine concepts, if needed, and contribute effectively to the preservation of relational databases.

In conclusion, we can say that digital preservation is essential to ensure a future access to digital information legacy. There is no solution to completely solve this problem, and we do not know if it will ever occur. However, the fact that this issue has become the central subject of a scientific study, will probably contribute to solve the problem in the future.

References

1. Ronald Bourret, "XML and Databases, " Copyright 1999-2005 by Ronald Bourret. Last updated September, 2005
2. Lee Buck. "Data models as an XML Schema development method", XML 99, Phyladelphia, Dec. 1999.
3. Donald D. Chamberlin, Raymond F. Boyce, "SEQUEL: A Structured English Query Language," IBM, 1970
4. Consultative Committee for Space Data Systems. "Reference Model for an Open Archival Information System (OAIS) - Blue Book," National Aeronautics and Space Administration, Washington, 2002.
5. Edgar Codd, "A Relational Model of Data for Large Shared Data Banks," in Communications of the ACM, 1970.
6. Michael Day, "The OAIS Reference Model," Digital Curation Centre UKOLN, University of Bath, 2006
7. Claire Eager, "The State of Preservation Metadata Practices in North Carolina Repositories," Chapel Hill, North Carolina, 2003
8. Miguel Ferreira, "Introdução à preservacao digital - Conceitos, estratégias e actuais consensos," Escola de Engenharia da Universidade do Minho, Guimarães, Portugal, 2006.
9. Ricardo Freitas, "Preservação Digital de Bases de Dados Relacionais," Escola de Engenharia, Universidade do Minho, Portugal, 2008
10. M. Jacinto, G. Librelotto, J. Ramalho, P. Henriques, "Bidirectional Conversion between Documents and Relational Data Bases," 7th International Conference on CSCW in Design, Rio de Janeiro, Brasil, 2002.
11. B. F. Lavoie, "The Open Archival Information System Reference Model: Introductory Guide," Digital Preservation Coalition, Dublin, USA, Technology Watch Report Watch Series Report, 2004.
12. K.-H. Lee, O. Slattery, R. Lu, X. Tang and V. McCrary, "The State of the Art and Practice in Digital Preservation," Journal of Research of the National Institute of Standards and Technology, vol. 107, no. 1, pp. 93-106, 2002.
13. "PLANETS - Preservation and Long-term Access through NETworked Services" [Online]. Available: http://www.planets-project.eu/
14. J. Ramalho, M. Ferreira, R. Castro, L. Faria, F. Barbedo, L. Corujo, "XML e Preservação Digital," Dep. Informática, Universidade do Minho e Instituto dos Arquivos Nacionais, Torre do Tombo, 2007
15. J. Ramalho, M. Ferreira, L. Faria, R. Castro "Relational Database Preservation through XML modelling, " Extreme Markup Languages 2007, Montréal, Québec, 2007
16. J. Ramalho, P. Henriques, "XML and XSL - Da Teoria à Prática," FCA - Editora Informática, 2002.
17. "SIARD - Format Description," Swiss Federal Archives - SFA, 2008.
18. K. Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," presented at The State of Digital Preservation: An International Perspective, Washington D.C., 2002.
19. D. Waters, "Good Archives Make Good Scholars: Reflections on Recent Steps Toward the Archiving of Digital Information," 2002
20. C. Webb, "Guidelines for the Preservation of Digital Heritage," United Nations Educational Scientific and Cultural Organization - Information Society Division, 2003.

21. Wikipedia contributors, "Database models," in Wikipedia, The Free Encyclopedia, 2008. [Online]. Available: http://en.wikipedia.org/wiki/Database_models/
22. XML, "Extensible Markup Language", in W3C - The World Wide Web Consortium [Online]. Available: http://www.w3.org/XML/