Relational Databases Digital Preservation
Phd Thesis Proposal

para o Programa Doutoral em Informática
das Universidades do Minho, Aveiro e Porto

(MAPI)

José Carlos Ramalho

Departamento de Informática
da Escola de Engenharia da Universidade do Minho

January 2008

1 Context and Background

Digital Archives are complex structures composed of human resources, state of the art technologies, policies and data. Due to the heritage keeping role that archives assume in our society, it is important to make sure that, the data that is produced by our organizations is preserved accordingly in order to document and provide evidence of their activities.

Information stored in an archive must be treated differently than other types of information because it is kept with the purpose of providing evidence of some organizations activity. Due to this fact some properties should be preserved for long periods of time: integrity, liability and authenticity. The preservation of this information extremely complex as digital objects are far from being stable. They are software and hardware dependent. Normally, their auto-preservation period is about 5 years.

In this context digital preservation practices become very important and should be part of the institution's planning. The problem is how to keep digital objects in such a way that their information is accessible long past their auto-preservation period.

RODA (Repository of Authentic Digital Objects) is a joint venture between public administration and academic researchers that aims to become the public administration repository. A repository where users can rely on digital objects authenticity and where digital objects are expected to endure long beyond the 5 years expectation.

For the first prototype three kinds of digital objects were considered: text documents, still images and relational databases.

This proposal target is the relational database type objects. In the next section we propose several problems that should be addressed and worked in order to achieve longterm digital preservation of relational databases.

2 Aims

We start by enumerating a set of questions that should lead this work during the first stages.

First the general big questions:

- How do we keep archived databases readable and usable in the long term (at acceptable cost)?
- How do we separate the data from a specific database management environment?
- How can we preserve the original data semantics and structure?
- How can we preserve authenticity and provenance of databases?
- How can we preserve data while it continues to evolve? (here we can split the problem in two: operational databases and frozen databases)
- How can we have efficient preservation frameworks, while retaining the ability to query different database versions?
- How can multi-user online access be provided to hundreds of archived databases containing terabytes of data?
- Can we move from a centralized model to a distributed, redundant model of database preservation?

Then, the more technical questions:

- What are the salient features of a database that should be preserved?
- What are the different stages in the database preservation's life cycle?
- What documentation is preserved together with a database, and in what format?
- What are the legal encumbrances on database preservation?
- What can be learned from traditional archival appraisal for the selection of databases for preservation?
- To what extent can the preservation strategies, and procedural policies developed by archivists be adapted for databases?

- How can we mesure the quality of preservation strategies when they are applied to databases?

These are the big questions. In order to be confortable when answering them the candidate should acquire knowledge in some areas like: Digital Preservation (policies and strategies), Authenticity (what laws do we have and what is available internationally?), Databases (Normalization, Query, Migration, ...). We predict that a 3 to 4 months work should be done in order to acquire these prerequisites.

3 Supervision and Contacts

This proposal is concluded with the supervisor personnel data:

Research Unit: Centro de Ciência e Tecnologia da Computação (CCTC).

Name: José Carlos Leite Ramalho

Email: jcrdi.uminho.pt

URL: `http://www.di.uminho.pt/~jcr`

And some context links:

RODA Project `http://roda.iantt.pt`

Some related publications :

- José Carlos Ramalho, Miguel Ferreira, Luis Faria, Rui Castro; "Relational database preservation through XML modelling"; Extreme Markup Languages 2007, Montréal - Canada, August 2007.; 2007; Montréal - Canada; 08; URI no RepositoriUM: `http://hdl.handle.net/1822/7120`;
- "An intelligent decision support system for digital preservation"; Miguel Ferreira, Ana Alice Baptista, José Carlos Ramalho; International Journal on Digital Libraries; Springer Berlin / Heidelberg; issn: 1432-5012 (Print) 1432-1300 (Online); 05; 2007; URI no RepositoriUM: `http://hdl.handle.net/1822/6648`;
- "A foundation for automatic digital preservation"; Miguel Ferreira, Ana Alice Baptista, José Carlos Ramalho; Ariadne; issn: 1361-3200; URI no RepositoriUM: `http://hdl.handle.net/1822/5571`; July; 2006;
- Miguel Ferreira, Ana Alice Baptista, José Carlos Ramalho; "CRiB : preservation services for Digital Repositories"; Apresentação efectuada na International Conference Open Repositories, 2, San Antonio, Texas, United States of America, 2007.; 2007; URI no RepositoriUM: `http://hdl.handle.net/1822/6195`;