Database Preservation Toolkit: a flexible tool to normalize and give access to databases

José Carlos Ramalho
University of Minho
Braga
Portugal
jcr@di.uminho.pt

Luis Faria
KEEP SOLUTIONS Lda
Braga
Portugal
lfaria@keep.pt

Hélder Silva
KEEP SOLUTIONS Lda
Braga
Portugal
hsilva@keep.pt

Miguel Coutada
University of Minho
Braga
Portugal
michaelcoutada@gmail.com

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 specic family of digital objects: Relational Databases; they are the most frequent type of databases used by organizations worldwide.

Database Preservation Toolkit enables the preservation of relational databases holding the structure and content of the the database in a preservation format in order to provide access to the database information in a long term period.

If in one hand there is a need to migrate databases to newer ones that appear with technological evolution, on the other hand there is also the need to preserve the information they hold for a long time period, due to legal duties but also due to archival issues. That being said, that information must be available no matter the database management system where the information came from.

In this area, solutions are still scarce. Main products for relational database preservation include CHRONOS and SIARD. The first one is, in most of the cases, unreachable due to the associated costs. The second one is not really a product but a preservation format.

The main idea behind this work was to explore the main features and limitations of the existing products in order to improve 'db-preservation-toolkit' (`http://keeps.github.io/db-preservation-toolkit/`), an extracted component from the RODA project (`http://www.roda-community.org`).

Therefore, 'db-preservation-toolkit' was improved with respect to performance and also with new features addiction in order to support more database management systems, address some missing features of the other products, support of a new preservation format (SIARD) and provide an interface where it is possible to access and search the information of the archived databases.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Digital Preservation, Databases, Migration, Significant Properties, Digital Object

1. INTRODUCTION

In the current paradigm of information society more than one hundred exabytes of data are already used to support our information systems [6]. The evolution of the hardware and software industry causes that progressively 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). In the eventual absence of appropriate mediators, who can guarantee the preservation of the digital artifacts? In other words, who has the responsibility to support the continuity of access to digital data [1]? Despite the concrete responsibilities and considering that there is no generic solution, several researchers and research projects aim to face this problem.

Although digital information can be exactly preserved in its original form by only copying (preserving) the bits, the problem appears when we notice the very fast evolution of those platforms (hardware and software) where the bits can be transformed into something human intelligible [4]. Digital archives and digital libraries 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 [3].

Our work addresses this issue of Digital Preservation and focuses on a specific class of digital objects: Relational Databases [4]. 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 [8]. These kinds of archives are especially important to organizations because they can justify their activities and characterize the organization itself. Current studies claim that 90% of the information produced in a daily basis is stored in a relational database.

Currently, in this project, we aim to support more database formats on ingestion, more database preservation formats as AIPs and new ways to explore the archived databases. In the following section we will describe the project context and its roots. Next we analyze the relational databases class of objects; we should be able to completely characterize this type of digital objects so that one may choose what are the issues (the things) important/valid/necessary for preservation. Following section establishes the significant properties for relational databases digital preservation. The significant properties are addressed, individually and globally, over different levels of abstraction. At the end we will draw some conclusions, specify the future work to be done and also enumerate some questions that emerge from the research.

2. RODA: THE BEGINNING...

In mid 2006, the Portuguese National Archives (Directorate-General of the Portuguese Archives) have launched a project called RODA (Repository of Authentic Digital Objects) aiming at identifying and bringing together all the necessary technology, human resources and political support to carry out long-term preservation of digital materials produced by the Portuguese public administration.

As part of the original goals of the RODA project was the development of a digital repository capable of ingesting, managing and providing access to the various types of digital objects produced by national public institutions. The development of such repository was to be supported by open-source technologies and should, as much as possible, be based on existing standards such as the Open Archival Information System (OAIS) [2], METS [12], EAD [11] and PREMIS [7].

The OAIS model is composed by three top processes: ingest, administration and dissemination. In RODA we have specified the workflows for each of these processes. The ingest process takes care of new information added to the repository. This information is delivered by the producer as an Submission Information Package or SIP. The SIP structure had to be formally specified so that third-party institutions were able to communicate with the repository. During ingest SIPs are transformed into AIPs (Archival Information Packages). The dissemination process takes care of consumer requests by transforming AIPs into DIPs (Dissemination Information Packages), a subset of the preserved information more adequate for delivery to end-users. Currently RODA is capable of storing and give access to the following types of digital objects: text-documents, still images, relational databases, video, audio and emails.

Normalization plays an important role in RODA. It was not possible to archive every kind of text-document or every kind of still image. Even with databases normalization was necessary as each Database Management System (DBMS) had its own data model. So we had to take mesures towards format normalization. Every digital object being stored in RODA is subjected to a normalization process: text documents are normalized as PDF files; Still Images are converted to uncompressed TIFFs; Relational databases are converted to DBML[8] (Database Markup Language).

The RODA project is divided into many different components and services having the Fedora Commons at the core of its framework. Fedora implements the common digital repository features, as digital objects and metadata storage and the ability to create relationships between objects. Fedora Commons also provides search capabilities by using the Lucene search engine under the hood. On top of that, we have developed the RODA Core Services, i.e. the basic RODA services, which can be accessed programatically. Finally, the RODA Web User Interface allows the end user to easily browse, search, access and administrate stored information, metadata, execute ingest procedures, preservation and dissemination tasks.

In spite of all the efforts invested in the development of RODA, there was still no support for real active digital preservation. Once the materials got into the archival storage they remained untouched and, therefore, susceptible to technological obsolescence, especially at the format level. At the same time, at the University of Minho, a project called CRiB (Conversion and Recommendation of Digital Object Formats) was being devised. This project aimed at assisting cultural heritage institutions as well as normal users in the implementation of migration-based preservation interventions. Among those services were format converters, quality-assessment tools, preservation planning and automatic metadata production for retaining representations' authenticity.

The CRiB system was developed as a Service Oriented Architecture (SOA) and is capable of providing the following set of services:

- File format identification;
- Recommendation of optimal migration options taking into consideration the individual preservation requirements of each client institution or user;
- Conversion of digital objects from their original formats to more up-to-date encodings;
- Quality-control assessment of the overall migration process - data-loss, performance and format suitability for long-term preservation;

- Generation of preservation metadata in PREMIS format to adequately document the preservation intervention and retain the objects' authenticity.

After obtaining supplementary funding to continue the development of RODA, the team decided to use CRiB as its preservation planning and execution unit.

The RODA project follows a service-oriented architecture to facilitate the parallel development and update and allow heterogeneous technology and platform independence between its various components. The CRiB project is also service-oriented, to allow the implementation of services that are only possible in specific platforms and technologies. This paper provides a description of both projects and about the integration of CRiB as on of RODA's components, allowing the use of its features for normalization processes during ingest, metadata generation, preservation planning and format migrations, and even dissemination services.

In this paper we are going to focus on the digital preservation of databases. We will raise the relevant questions on this topic and we are going to discuss the decisions we took in the past and the ongoing work.

3. PAST

Preserving digital data is a complex technological puzzle. Databases are one of the most complex digital object types to deal with. To simplify the problem we decided to address the problem by layers: data, structure and semantics. These layers match database significant properties and tell us what to preserve and how to measure the quality of the digital preservation strategy being followed. The data layer extracts data and migrates it to the preservation format. Structure layer does the same with the database structure. The semantics layer will deal with all the remaining database features that should be preserved.

Our first approach was to deal with the first two layers, the preservation of the database data and structure, i.e., the preservation of the database logical model. We developed a RODA component that extracts the first two layers from its specific database management environment (DBMS). Its first version used DBML[8] neutral format for the representation of both data and structure (schema) of the database.

This component was presented and demonstrated at the Open Planet workshop, "Database Archiving", held at Danmark National Archives in 2012. During the workshop it became clear that more formats should be supported and we should also change the preservation format. Although there is no standard for a database preservation format, SIARD[10] is being adopted in several european institutions and projects and when compared to DBML it already supports part of the semantics layer and had some scalability properties. So we decided to have it also as our preservation format. Back then we also decided to support other DBMS formats like DB2, and other preservation formats like AADL (used by Sweden, Norway and Finland) as input and output formats of our toolkit. This way our toolkit will become a real interoperability tool.

4. OAIS, SIPS AND DATABASES

RODA follows the Open Archival Information System Reference Model (OAIS) [2]. OAIS identifies the main functional components that should be present in a archival system capable performing long-term preservation of digital materials. The proposed model is composed of four principal functional units: Ingest, Data management, Archival storage and Access; and two additional units called Preservation planning and Administration. Figure 1 depicts how these functional units interact with each other and with all the stakeholders of the repository (internal and external).

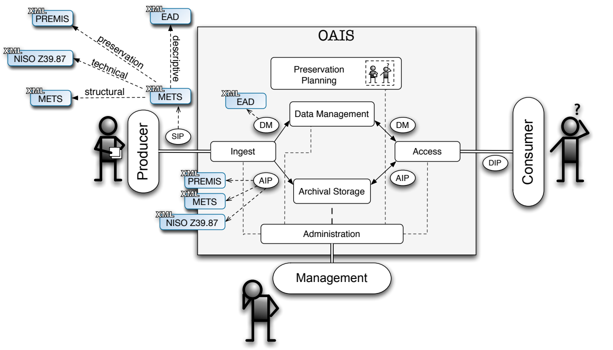

Figure 1: RODA general architecture

4.1 Ingest process

The ingest process is responsible for accommodating new materials into the repository and takes care of every task necessary to adequately describe, index and store those materials. For example, in this stage the repository may transform submitted representations to normalized formats adequate for long-term preservation and request the user to add descriptive metadata to those objects to facilitate their future retrieval using available search mechanisms. It is also common practice to store the original bit-streams of ingested materials together with the normalized version (just in case a more advanced preservation strategy comes along to rescue those old bits of information).

New entries come in packages called Submission Information Packages (SIP). When the ingest process terminates, SIPs are transformed into Archival Information Packages (AIP), i.e. the actual packages that will be kept in the repository. Associated with the AIP is the structural, technical and preservation metadata, as they are essential for carrying out preservation activities.

The SIP is the format used to transfer new content from the producer to the repository. It is composed of one or more digital representations and all of the associated metadata, packaged inside a METS envelope. The structure of a SIP supported by RODA is depicted in Figure 2. The RODA SIP is basically a compressed ZIP file containing a METS document, the set of files that compose the submitted representations and a series of metadata records. Within the SIP there should be at least one record of descriptive metadata in EAD-Component format[1]. However, one may also find preservation and technical metadata inside a submission

[1] An EAD record does not describe a single representation.

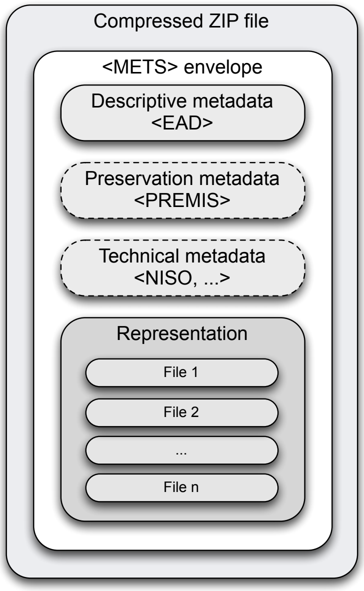

Figure 2: Submission Information Package structure

package, although this last set of metadata is not mandatory as it is seldom created by producers. Nevertheless, it was felt important that RODA should support those additional SIP elements for special situations such as repository succession, i.e. when ingested items belong to another repository that is to be deactivated.

Before SIPs can be fully incorporated into the repository they are submitted to a series of tests to assess its integrity, completeness and conformity to the ingest policy.

If any of the validation steps fails, the SIP is rejected and a report is sent to the archivists group as well as to the producer. The producer may then fix the problem and resubmit a new version of the SIP.

5. DATABASE SIP

Database SIPs are very similar to other SIPs. The difference relies on the representation files. For the other formats we only had to choose one normalization format to use for the representation files: for images we chose TIFF, for text based documents we chose PDF and so on. But for databases there wasn't such a format. Each DBMS supported its own format. Even SQL has some different versions. So, we had to create and specify a new format.

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.

Since late 1990's, XML was accepted as the neutral format for information representation and information interchange. This is due, mainly, to two factors. On one hand, XML documents are purely textual files, structured and independent of any hardware or software platforms. On the other hand, it is widespread and more and more public domain tools are available to help users transforming XML documents.

XML was the obvious choice for the base format of our representation files. Both DBML and SIARD use XML as the base format.

DBML and SIARD are the only XML based database preservation formats. Although they are easy to process by both machines and humans, converting a database into DBML or SIARD is not easy and it is not a task humans can do by hand. So, the next step was to create a tool capable of generating DBML from different DBMS.

We also keep an SQL version of the database information in the version supported by the original DBMS. This has to do with the preservation policy, we always keep the original object or, at least, the closest version of it (we do not know what the future may bring and we can't predict how the actual DBMS will evolve).

6. DATABASE SIP BUILDER

In RODA's context we soon realized that we could not just deliver a format and demand from producers to send us the information packages accordingly. In projects like this one it is important to have wide acceptance from the community of users.

We developed a tool to create these SIPs. This tool was integrated in RODA but due the growing interest it has emancipated as a tool that can be integrated with other systems and tools. Its architecture is presented in figure 3.

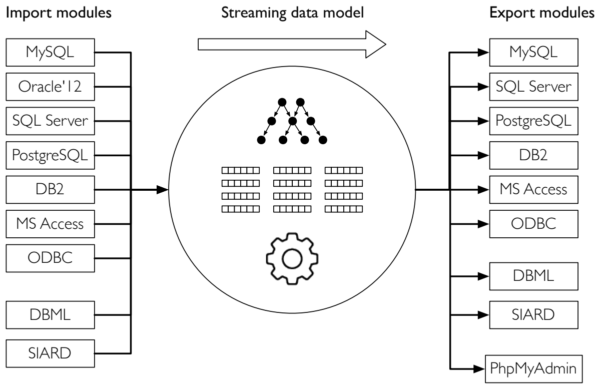

Figure 3: Database SIP builder architecture

We are addressing database significant properties by layers. Each layer raises different problems that have to be solved with appropriate solutions.

6.1 Data layer

Extracting data from a DBMS it is not difficult, we just have to connect to the DBMS and issue an SQL statement like

In fact, EAD is used to describe an entire collection of representations. Our SIP includes only a segment of EAD, sufficient to describe one representation, i.e. a <c> element and all its sub-elements. The team has called this subset of the EAD an EAD-Component.

”SELECT * FROM”. In DBML all the data is dumped in a single XML file. We had the idea to segment the data but SIARD already did that. That was one of the reasons that took us to support SIARD as the preservation format. DBML had to change to be able to take care of real databases and most of the needed changes were already implemented in SIARD.

6.2 Structure layer

Each DBMS stores the structural information in its own specific way and to overcome this situation we had to develop specific connectors, import modules, for each one. For each DBMS we created a connector that connects to the database and knows how to extract its structural information. If we need to support a new DBMS in the future we just need to program a new import module for that DBMS. In the last version, we added support for DB2 creating a new import module for this DBMS.

6.3 Semantics layer

This layer corresponds to the behavioral part of a database and is where the focus of the discussion in this area is. We include in this layer: views, stored procedures, rights, roles, user management, APIs, interfaces, and other feature we can come across.

Currently there are partial solutions for it. DBML does not support it, it only deals with the first two layers.

SIARD enables SIP creators to store views, stored procedures and constrains capturing a significant part of the database behavior. These behavioral components are captured in SQL99 and stored inside an XML envelope.

For some consumers we are still missing many things: forms that the application uses to capture input from users, reports, etc. In most of these cases, we try to capture de knowledge with application metadata and application images/screenshots.

7. DATABASE ACCESS

We can see dbtoolkit as a SIP builder but also as a tool that enables several ways to accessor deploy archived databases. It can deploy a DIP very similar to the original SIP, an SQL based original database or, like figure 3 shows, any other format that has an export module contributed by some community programer. This way, we can look at this tool also as a database converter between different DBMS.

Back in RODA, we needed a nice user interface that would enable users to explore the archived databases. We took *phpMyAdmin*, simplified it and we end up with a tool that allows users to browse databases, to access data, to access structure information and to execute some SQL queries. This new access component works with MySQL export module and uses a local MySQL DBMS to cache the database. The problem with this approach is that it does not scale for large databases or a large database quantity.

Pursuing the scalability idea we are launching a new project to create a faster viewer with simpler interfaces. The project is illustrated in figure 4. The idea is to dump and index the data on a search engine like Lucene and having that engine as the interface to access data. This way we won’t need an external DBMS or an external database cache to access the data making the access functionality simpler and faster.

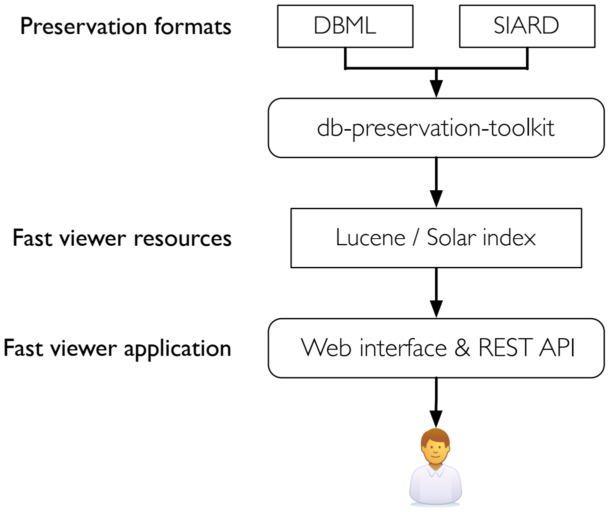

Figure 4: New Database Viewer

8. FUTURE WORK

As future work we still have to improve some features and to run some tests.

We are working on new small projects pursuing the idea of reverse engineering the relational model. Since we are free from the DBMS why shall we stick with the relational model? Relational model is optimized for transactions. If we have an archived *frozen* database we won’t be executing transactions. If we don’t need the relational model we can undo the database normalization going towards the original conceptual database model.

During a phd thesis we have been working to create algorithms to migrate data from a relational model into an ontological model close to the database conceptual model [5]. In a more recente work we created a SIARD to RDF converter and implemented a simple RDF navigator for databases [9].

9. ACKNOWLEDGMENTS

This work is supported by the European Commission under FP7 CIP PSP grant agreement number 620998 - E-ARK.

10. REFERENCES

[1] F. Berman. Surviving the data deluge. *Communications of the ACM*, 51(12), 2008.

[2] Consultative Committee for Space Data Systems. National Aeronautics and Space Administration, 2002.

[3] M. Ferreira. *Introdução à preservacao digital - Conceitos, estratégias e actuais consensos*. Escola de Engenharia da Universidade do Minho, Guimarães, Portugal, 2006.

[4] R. Freitas. Preservação digital de bases de dados relacionais. Master’s thesis, Escola de Engenharia,

Universidade do Minho, Portugal, 2008.

[5] R. A. P. Freitas. *Relational databases digital preservation*. PhD thesis, Engineering School, University of Minho, Portugal, 2013.

[6] P. Manson. Digital preservation research: An evolving landscape. *European Research Consortium for Informatics and Mathematics - NEWS*, 2010.

[7] PREMIS Working Group OCLC Online Computer Library Center & Research Libraries Group. Data dictionary for preservation metadata: final report of the premis working group oclc online computer library center & research libraries group. Technical report, Dublin, Ohio, USA, 2005.

[8] J. Ramalho, M. Ferreira, L. Faria, and R. Castro. Relational database preservation through xml modelling. In *Extreme Markup Languages 2007*, Montréal, Québec, 2007.

[9] F. Rocha. Preservação de Bases de Dados com SIARD. Master's thesis, Engineering School, University of Minho, Portugal, 2014.

[10] Swiss Federal Archives - SFA. Siard - format description. http://www.bar.admin.ch/themen/00876/00878/, 2008.

[11] The Library of Congress. Página oficial do ead versão de 2002. http://www.loc.gov/ead/, 2002. http://www.loc.gov/ead/.

[12] The Library of Congress. Mets webpage. http://www.loc.gov/standards/mets, 2006.