RODA and Crib
A Service-Oriented Digital Repository

José Carlos Ramalho
DI/UM - CCTC
jcr@di.uminho.pt

Miguel Ferreira
DSI/UM
mferreira@dsi.uminho.pt

Luis Faria
DGARQ
lfaria@iantt.pt

Rui Castro
DGARQ
rcastro@iantt.pt

Francisco Barbedo
DGARQ
frbarbedo@iantt.pt

Luis Corujo
DGARQ
lcorujo@iantt.pt

Abstract

In 2006 the Portuguese National Archives (Directorate-General of the Portuguese Archives) engaged in the development of an OAIS compatible digital repository system for long-term preservation of digital material. Simultaneously, at the University of Minho a project called CRiB was being devised which aimed at the development of a wholesome set of services to aid digital preservation. Among those services were format converters, quality-assessment tools, preservation planning and automatic metadata production for retaining representations? authenticity. This paper provides a detailed description of both projects and discusses how these may be integrated into a complete digital preservation solution based on currently available archiving and preservation standards, e.g. OAIS, EAD, PREMIS, METS and ANSI/NISO Z39.87.

Introduction

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) (SYSTEMS 2002), METS (of Congress 2006) , EAD (of Congress 2002) and PREMIS (Group 2005).

At an higher level the OAIS model is composed by three mega processes (ingest, administration and dissemination). In RODA we have specified the workflows for each one of those. Ingest process takes care of new information packages additions to the repository (Submission Information Packages - SIP): the SIP structure was formal specified. During ingest SIP are transformed into AIP (Archival Information Package): we had to specify a data model for storing AIPs. Dissemination process takes care of consumer requests delivering information packages to them (DIP - Dissemination Information Packages). We had to specify one or more DIP structures for each type of Digital Object stored in RODA repository. Currently RODA is capable of storing and give access to the following types of Digital Objects: Text Documents, Still Images and Relational Databases.

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, each Database Management System has its own datamodel. So we had to take mesures towards format normalization. Every Digital Object being stored in RODA suffers a normalization process: Text Documents are normalized into PDF; Still Images are normalized into uncompressed TIF; Relational Databases are normalized into DBML(Ramalho et al. 2007) (Database Markup Language).

The RODA project is divided in different components, being the base component the Fedora Commons framework. Fedora implements the common digital repository features, as digital object (and metadata) storage abstraction and relationships between objects, and it can be extended by the Fedora's Generic Lucene Search engine. On top of that, the RODA Core Services implements all the base RODA services, which can be accessed programatically. Finally, the RODA Web User Interface allows the end user to easily browse, search, access and administrate all the digital objects, metadata and ingest, 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 consumers 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;
- Conversion of digital objects from their original formats to more up-to-date encodings;
- Quality control assessment on 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 the 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 a RODA component, allowing the use of its features in the ingest normalization and metadata generation tasks, on the preservation planning and events, and even on the dissemination services.

RODA project

Digital archives are complex structures usually composed of human resources, high-end technologies, policies and information. The RODA project set ground for a series of studies on all these axes. Its original goals were rather ambitious, namely:

- to define the functional requisites for a digital archive, its consumers and compliant applications;
- to devise conceptual, logical and content models for a digital archive;
- to identify the set of metadata schemas that are necessary to support all functions of the digital archive (descriptive, technical, structural and preservation metadata);
- to identify technical and organizational requisites;
- to develop a digital repository system capable of storing and preserving digital objects for the amount of time defined in the law;
- to develop software modules that integrate with available records management software applications;
- to develop an acquisition and ingest policy for the digital objects produced by Portuguese public institutions;
- to devise a preservation plan and policy for the digital archive;
- to promote a study on business models capable of financing the digital archive;
- to define taxonomies of significant properties for each class of digital objects to be supported by the archive in order to implement quality control mechanisms.

One of the stages of this project consisted in the development of a repository system capable of preserving digital information and making sure that that information remained accessible to its potential consumers without ever compromising its authenticity. This repository would serve as a basis for the development of a fully functional digital archive capable of ingesting and managing large quantities of digital objects at the national level. In its first version, the repository was expected to handle a small range of object classes, namely, text documents, raster images and relational databases.

Architecture

RODA follows the Open Archival Information System Reference Model (OAIS) (SYSTEMS 2002). 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).

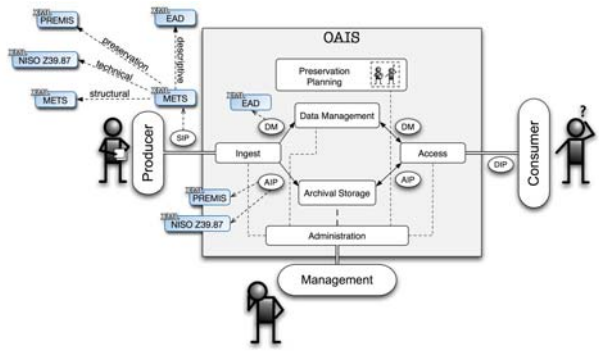

Figure 1: **RODA general architecture**

Before engaging in any technical developments, a collection of functional requisites was assembled by RODA?s archival team (Barbedo 2006) and a study on currently available repository platforms was conducted. In this study, DSpace (COMPANY and LIBRARIES) and Fedora (Lagoze et al. 2005) were compared against this collection of requisites.

DSpace outperforms Fedora on most of the requisites. Nevertheless, the project team ended up choosing Fedora as its development platform. Even though DSpace, as it comes out of the box, combines a broader range of ready-to-use features and user-friendly interfaces, it lacks flexibility and expansibility. One very pragmatic example of this is the support metadata schemas other than Dublin Core. One would

have to go through a tremendous amount of work to make DSpace compatible with more complex descriptive metadata structures such as EAD (of Congress 2002).

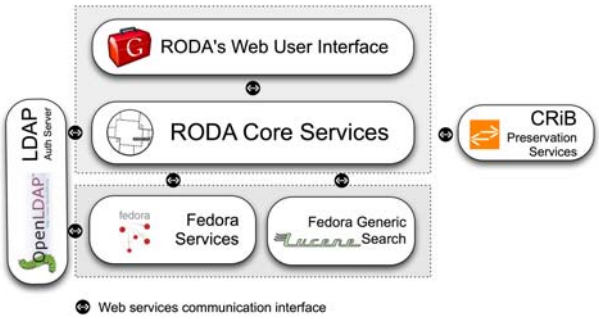

Figure 2: **RODA service oriented architecture**

Figure 2 depicts the overall architecture of services that compose RODA?s repository. On the bottom one may find the basic services provided by Fedora. These account for elementary tasks at the Data Management and Archival Storage level. Examples of such services are ingest, add a data stream to an object, get data stream, purge object, find objects and list data streams. For a complete list of the services provided by Fedora (Project). Fedora search capability is supported by Apache Lucene and its authentication procedures go through a LDAP server (Lightweight Directory Access Protocol).

RODA Core Services are responsible for carrying out more complex tasks such as implementing the complete set of actions that compose the ingest workflow, querying the repository in more advanced and abstract ways and carrying out administrative functions on the repository. The same LDAP server previously described is used by RODA?s Core Services for authenticating repository users.

On top of the RODA?s Core Services lays the RODA?s Web User Interface (RODA-WUI). This layer handles all the aspects of the graphic user interface for producers, consumers, archivists, system administrators and preservation experts. The RODA-WUI components are supported by the Google Web Toolkit and all communication is done via AJAX and Web services technologies.

Ingest process As previously described, Fedora only provides a set of very basic services that developers are expected to extend in order to create a fully working repository system. This includes the development of graphical user interfaces and the characterization of most of the OAIS functional units outlined at the beginning of this section. The ingest process was the first of the units to be developed.

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.

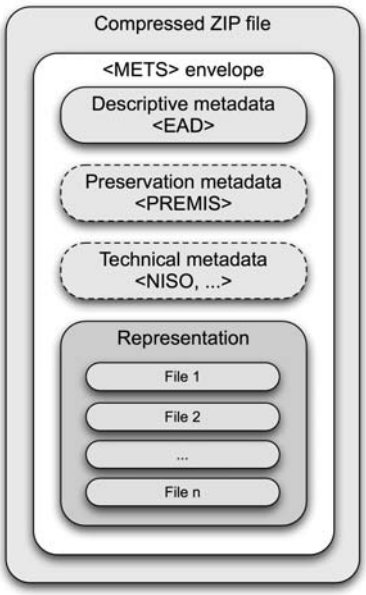

Figure 3: **Submission Information Package structure**

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 3. 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 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.

[1] An EAD record does not describe a single representation. 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 subelements. The team has called this subset of the EAD an EAD-Component.

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.

Access interface The access component establishes an interface between the archive and the end user (i.e. the consumer). This functional unit is able to locate an AIP by querying the data management and retrieve it from the archival storage unit. The AIP is then transformed to a Dissemination Information Package (DIP) and delivered to the consumer.

Data model

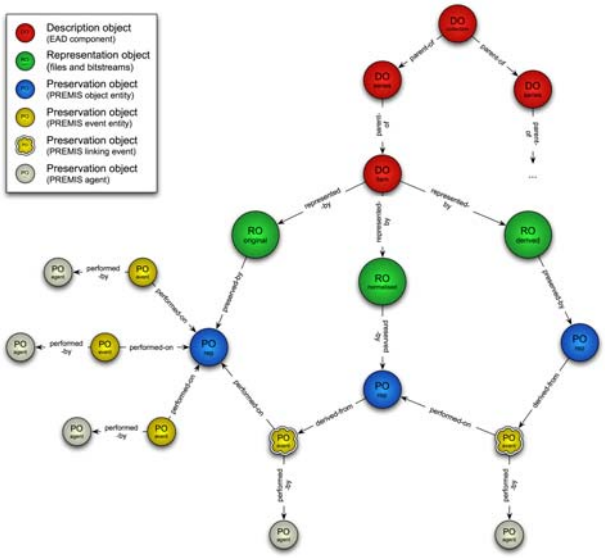

Figure 4: **RODA Data model**

RODA?s data model is atomistic and very much PREMIS-oriented (Figure 4). Each intellectual entity is described by an EAD-component metadata record (DO nodes in Figure 4). These records are organized hierarchically in order to constitute a full archival description of a collection but are kept separately within the Fedora Commons content model. Relationships between EAD-components are created using Fedora?s own RDF linking mechanism.

Additionally, each leaf record of a hierarchical collection (i.e. a file or an item) is linked to a representation object (RO nodes in the figure), i.e. a fedora object that embeds all the files and bit-streams that actually compose the digital representation. Finally, each of these objects are linked together by a set of PREMIS entities that maintain information about the digital object?s provenance and history of events (PO nodes).

Each preservation event that takes place in the repository is recorded as a new preservation-event node (i.e. PO event nodes in the figure). Special events, like format migrations, establish relationships between two preservation-representation nodes. These are called linking events in this context. Each preservation event is executed by an agent, whether this be a system user or an automatically triggered software application. The agent that triggered the event is recorded in PO agent nodes.

CRIB

CRiB[2] is a project being developed at the University of Minho that delivers a set of preservation services intended to aid client institutions in the planning and execution of migration-based preservation interventions (Ferreira, Baptista, and Ramalho 2007; 2006). Preservation planning is supported by a recommendation service that makes educated decisions on the best migration options available and takes into consideration the individual requirements of each client institution. The preservation execution component is handled by a large set of migration services that may be composed together to create more complex migration paths. To better understand how the system works, one should describe each of its constituent components.

Architecture

Figure 5 illustrates CRiB general architecture. The application layer illustrates how client applications may take advantage of the services provided by the CRiB. Examples of such applications may be custom programmes developed by individual users or complex applications such as digital repository systems like DSpace, Fedora, Eprints or RODA.

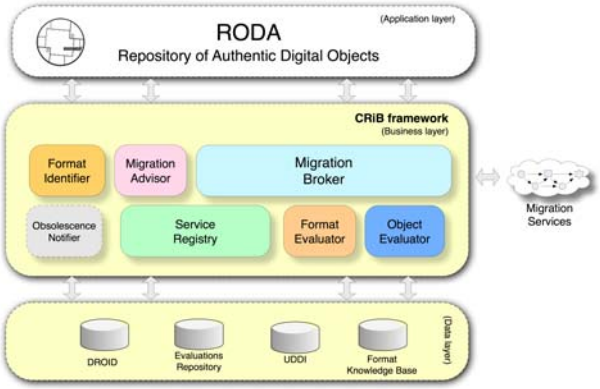

Figure 5: **CRiB architecture**

The middle layer illustrates the set of components that actually constitute the CRiB.

The Format Identifier, as the name suggests, is a service capable of identifying the underlying encoding of a digital representation. Client applications responsible for preserving digital objects must be able to identify, characterise and validate the integrity of its objects, if possible without human intervention. This service is indispensable in accom-

[2]Conversion and Recommendation of Digital Object Formats

plishing this goal. Furthermore, it enables format descriptions to be uniform across all components of the CRiB - format descriptions belong to a controlled vocabulary defined by the PRONOM file format registry developed by the National Archives of the UK (Darlington 2003).

The Obsolescence Notifier is responsible for monitoring the level of disuse of recognised file formats. When a given file format is at risk of becoming obsolete (e.g. when a new version of a format is published), this component will make sure that the adequate preservation events are triggered. CRiB does not actually implement this component as it has largely been the focus of an Australian initiative called AONS (Curtis et al. 2007).

The CRiB also delivers a large set of migration services for converting still-images and text-documents between various formats. The Service Registry component is responsible for storing information about these services. This allows the CRiB to rapidly discover which migration services are available and ready to be used. The metadata elements used within this component are based on the Universal Description, Discovery and Integration (UDDI) standard.

The Migration Broker is an additional component that is responsible for making sure that composite migrations are carried out atomically from the CRiB?s point-of-view. Additionally, the broker is in charge of measuring the performance of each migration service for purpose of assuring quality control. Performance is measured according to multiple criteria, such as: availability, stability, throughput, cost, size of the outcome representation and the number of outcome files in the resulting representation in relation to the original one. The results of these evaluations are then stored in an additional component called the Evaluations Repository. This repository is used by the Migration Advisor to determine the most apt migration services available.

The Object Evaluator is the component accountable for detecting the data-loss that might occur during the conversion process. This assessment is fundamental in determining the success of a migration process and to adequately document the preservation intervention. This component works by comparing the representation submitted to migration with its converted counterparts. Evaluations are performed according to fixed, but extendable, set of criteria, usually known as significant properties. These constitute the set of attributes that are expected to be maintained intact during the preservation intervention. They constitute the range of attributes that characterize the digital representation as a unique intellectual entity, independently of the format in which the representation is encoded.

The evaluations performed by the Object Evaluator are returned to the client application for documentation purposes and stored in the Evaluations Repository (again, to aid the subsequent recommendation process).

The evaluation report sent to the client follows the structure of the Event entity described in the PREMIS Data Dictionary. This entity includes elements for describing the type of event (e.g. Migration), the date and time of occurrence, the agent that carried out the event and detailed information regarding the outcome of the event (e.g. the amount of changes on significant properties that occurred during the migration process).

The Format Evaluator provides information about the current status of file formats. This information enables the Migration Advisor to determine to which formats are more adequate for long-term preservation by looking at its technical characteristics. The Format Evaluator works by questioning the Format Knowledge Base, i.e. a data store of known facts about digital formats. In the future this service could be replaced by other sources of information such as services provided by PRONOM or other external services such as Google Trends.

The Migration Advisor is in charge of preservation planning. It accomplishes this by generating suggestions of migration alternatives and works by confronting the preservation requirements outlined by client applications and its users with all the accumulated knowledge about the quality/performance of each individual migration service (or composition of services). It is important to point out that this component learns from each executed migration. During a migration, the system records its quality/performance in terms of data loss, status of involved formats and migration performance. Using this information, the Migration Advisor is able to rank all the available migration options and produce an appropriate suggestion for a migration intervention. Migration suggestions typically include the target format and the access point(s) of the most optimal migration services and/or migration paths. Additional information on the inner workings of this recommendation process may be found on (Ferreira, Baptista, and Ramalho 2007; 2006).

RODA meets CRiB

As previously described, CRiB offers a large set of preservation services that may be used by any client institution, application or individual user in order to maintain their collections of digital objects in interpretable and in up-to-date encodings making sure that the risk of loosing important representational features is kept to a minimum.

Most of the services delivered by CRiB are relevant to RODA. The following scenario depicts how these services may be used semi-automatically by the repository:

> During the course of its activity, RODA is expected to ingest and archive a large set of digital objects, most of these being submitted by its own producers and very likely, encoded in various formats. After ingesting these objects, RODA typically invokes the file Format Identification service provided by CRiB as to determine whether or not the recently deposited digital object is well formed and recognised as being in one of the preservation formats stated in the preservation policy of the archive.
>
> If the ingested object is not already in an acceptable preservation format, the CRiB may be queried to find available migration services capable of carrying out the correspondent normalisation. After obtaining a list of possible migration services, RODA may opt to invoke one of the suggested access points in order to obtain a novel representation of that object.

Together with the new representation, CRiB returns a migration report that thoroughly describes the outcome of the preservation action, especially in what concerns the effects of the intervention on the significant properties of the original object. This report may then be stored by RODA as preservation metadata to fully document the undertaken intervention. Preservation metadata serves the purpose of providing evidence on all preservation actions applied to any given object in the repository and is considered a fundamental tool in the preservation of authentic digital objects.

The repository also makes use of CRiB?s migration services to create derivative representations of its preserved objects with the goal of making them more adequate for dissemination.

Routinely, the repository consults the Obsolescence Notifier to check if any of its preservation formats is at risk of becoming exceptionally outdated. If so, the repository may request CRiB?s Migration Advisor to provide a recommendation for a new preservation format and engage in the migration of all of its outdated objects.

Preservation management within RODA is handled by a scheduler in which a special user, i.e. the preservation expert, may define the set of rules that trigger specific preservation actions. Preservation actions comply with a common API, so creating and installing new actions in the repository is as easy as copying the programme file to the right directory on the server. These actions may invoke remote services such as the ones provided by CRiB, but must be deployed locally for the sake of conformity. The locally deployed actions must handle all remote service invocations and handle all possible exceptions that might occur.

The scheduler allows the preservation expert to configure the rules that will select relevant objects for a particular preservation intervention as well as scheduling the intervention itself.

Conclusion and Future Work

As previously described, RODA has been planned to be a complete digital repository providing functionality for all the main units that compose the OAIS reference model. RODA fully implements an Ingest workflow that not only validates SIPs, but also takes care of the whole negotiation process between the archive and the producers of information. RODA also accounts for Access providing different ways to search and navigate over available metadata as well as visualizing/downloading stored digital objects. Administration components were also developed allowing archivists to change the descriptive metadata and define rules for preservation interventions such as scheduling integrity checks on all stored digital objects, initiate the migration process of certain representation class/formats, or control which users or groups are authorized to perform certain actions within the repository.

Although RODA covers most of the functional components described in the OAIS reference model, there was still one very important one missing in its design ? Preservation Planning.

According to OAIS, Preservation Planning is responsible for providing ?the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete. Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates to migrate current archive holdings, developing recommendations for archive standards and policies, and monitoring changes in the technology environment and in the Designated Communities? service requirements and Knowledge Base. [...] Preservation Planning also develops detailed Migration plans, software prototypes and test plans to enable implementation of Administration migration goals.? (SYSTEMS 2002).

It was obvious to the development team that the missing functionality in RODA would easily be fulfilled by CRiB and its set of components. CRiB, because of its service-oriented nature, would integrate seamlessly with the rest of the components and services developed around Fedora.

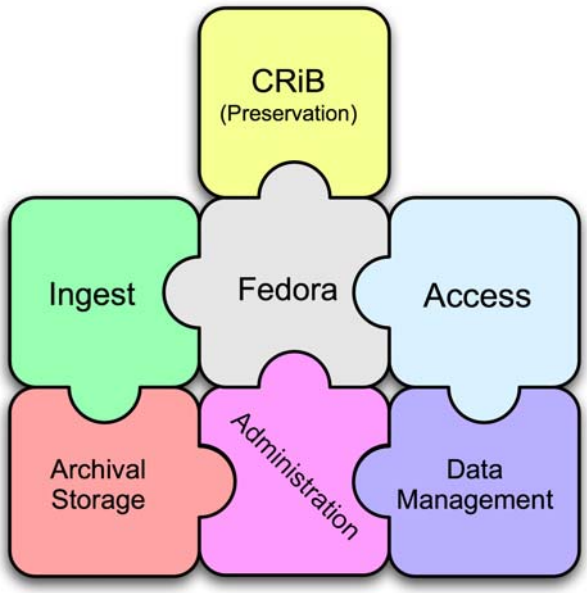

Figure 6: **RODA with the new component**

In order to fully satisfy the initial requirements of RODA, CRiB would have to be able to handle at one additional class of digital representations, i.e. relational databases. Future work will focus on the development of a taxonomy of significant properties for relational databases, the specification of a long-term preservation format/schema for this class of objects, the development of migration services for distinct database products (e.g. Oracle, SQL Server, Postgresql, MySQL and others). Some groundwork on this subject has already been initiated and may be consulted at (Ramalho et al. 2007; Henriques et al. 2002).

References

Barbedo, F. 2006. Especificação de requisitos. Technical Report 41012-005, IAN/TT.

COMPANY, H.-P., and LIBRARIES, M. Dspace web site. http://www.dspace.org. http://www.dspace.org.

Curtis, J.; Koerbin, P.; Raftos, P.; Berriman, D.; and Hunter, J. 2007. Aons - an obsolescence detection and notification service for web archives and digital repositories. *New Review of Hypermedia and Multimedia 13*.

Darlington, J. 2003. Pronom - a practical online compendium of file formats. *RLG DigiNews 7*.

Ferreira, M.; Baptista, A. A.; and Ramalho, J. C. 2006. A foundation for automatic digital preservation. *Ariadne*.

Ferreira, M.; Baptista, A. A.; and Ramalho, J. C. 2007. An intelligent decision support system for digital preservation. *International Journal on Digital Libraries*.

Group, P. W. G. O. O. C. L. C. . R. L. 2005. 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.

Henriques, M.; Libreotto, G.; Ramalho, J.; and Henriques, P. 2002. Bidirectional conversion between xml documents and relational data bases. *International conference on CSCW in design*.

Lagoze, C.; Payette, S.; Shin, E.; and Wilper, C. 2005. Fedora - an architecture for complex objects and their relationships. *Journal of Digital Libraries*.

of Congress, T. L. 2002. Página oficial do ead versão de 2002. http://www.loc.gov/ead/. http://www.loc.gov/ead/.

of Congress, T. L. 2006. Mets webpage. http://www.loc.gov/standards/mets. http://www.loc.gov/standards/mets.

Project, F. The fedora digital object model. http://www.fedora.info/download/2.0/userdocs/digitalobjects/objectModel.html. http://www.fedora.info/download/2.0/userdocs/digitalobjects/objectModel.html.

Ramalho, J. C.; Ferreira, M.; Faria, L.; and Castro, R. 2007. Relational database preservation through xml modelling. In *Extreme Markup Languages 2007, Montral - Canada*.

SYSTEMS, C. C. F. S. D. 2002. National Aeronautics and Space Administration.