RODA: A Service-Oriented Digital Repository

Database Archiving

José Carlos Ramalho
jcr@keep.pt

José Carlos Ramalho
jcr@di.uminho.pt

Dep. of Informatics
University of Minho
**Context**

**RODA (2006-2009)**
- Metadata management (EAD based)
- Digital object management (...)
- Digital preservation protocols and policies
- National project with public funds (Portuguese National Archives)

**CRiB: Preservation Services for Digital Repositories (2005-2008)**
- Distributed Migration Service
- Migration Adviser Service
- Phd thesis at U.Minho (Miguel Ferreira)
Context

RODA (2006-2009)
• Metadata management (EAD based)
• Digital object management (...)
• Digital preservation protocols and policies
• National project with public funds (Portuguese National Archives)

CRiB: Preservation Services for Digital Repositories (2005-2008)
• Distributed Migration Service
• Migration Adviser Service
• Phd thesis at U.Minho (Miguel Ferreira)
Premisses

- RODA:
  - OAIS as reference model
  - EAD as descriptive metadata
  - Open Source
  - SOA
  - Preservation Policy: Migration
  - Authenticity

- CRiB:
  - Distributed Migration
  - SOA
  - Quality control
  - Migration Adviser
Subject: Who really was Cristophoros Colombus?
Was he italian? Spanish? Or a portuguese belonging to a jewish family?
We must trust our sources: in ancient History there are no direct speech or evidence.

EX: the bible
We must trust our sources: in ancient History there are no direct speech or evidence.

EX: the bible

How do we become trustful?
We must trust our sources: in ancient History there are no direct speech or evidence.

EX: the bible

How do we become trustful?

• Reputation

• Documenting every action taken upon DOs

• Sometimes the law will be an obstacle
DO Classes: normalization
Open Archival Information System

ISO 14721
OAIS (Information Packages)

- Submission Information Package (SIP)
  - Digital Object
  - Metadata created by producer
    - too open...
- Archival Information Package (AIP)
  - Digital Object to be stored
  - Metadata: enough to ensure DO’s preservation and access
    - model defined by PREMIS
- Dissemination Information Package (DIP)
  - DO transformed into the format that will be delivered to the consumer
- Metadata
Ingestion
Ingestion

Submission Contract
• SIP specification
• Ingestion workflow specification
SIP Structure (example)

one still image
SIP Structure (example)

one still image

creation properties:
  - date
  - hardware
  - ...

terça-feira, 7 de Fevereiro de 2012
SIP Structure (example)

-one still image

creation properties:
- date
- hardware
- ...

Technical Metadata:
- color
- dimensions
- ...

<XML>

--

--

--

--
**SIP Structure (Example)**

- **Creation properties:**
  - date
  - hardware
  - ...

- **Technical Metadata:**
  - color
  - dimensions
  - ...

- **Descriptive Metadata:**
  - producer
  - collection
  - ...

one still image
SIP Structure (example)

- one still image
- creation properties:
  - date
  - hardware
  - ...

Technical Metadata:
- color
- dimensions
- ...

Descriptive Metadata:
- producer
- collection
- ...

Manifest
**SIP Structure (Example)**

- **Manifest**
- **Compressed File**
- **one still image**
- **creation properties:**
  - date
  - hardware
  - ...
- **Technical Metadata:**
  - color
  - dimensions
  - ...
- **Descriptive Metadata:**
  - producer
  - collection
  - ...

*terça-feira, 7 de Fevereiro de 2012*
SIP Structure (+complex)
SIP Structure (+complex)
SIP Structure (+Complex)

DO = Image+
SIP Structure (+complex)
SIP Structure (+complex)

DO = Image+  Properties  Technical Metadata
SIP Structure (+complex)

DO = Image+

Properties

Technical Metadata

Descriptive Metadata
SIP Structure (+complex)

DO = Image+ Properties Technical Metadata Descriptive Metadata

Manifest
SIP Structure (+complex)

DO = Image+
Properties
Technical Metadata
Descriptive Metadata

Compressed File
Ingestion Workflow

**SIP** → **AIP**

- integrity check
- virus check
- generation of preservation metadata (PREMIS)
- conversion to a normalized format
- generation of technical metadata
- generation of preservation metadata (PREMIS)
Roda Data Model
Roda Data Model

Description Objects
Implementation Requisites

- Graphical Interface for Ingestion process
- Producer registry
- SIP production tool
- Ingestion feedback
- Partial Ingestion
- “Quarantine” zone: cache, ingestion buffer
- SIP validation
- Error reporting
- Persistent identifiers
- PREMIS event generation
- DIP digital signature
- ...
Development framework
Requisites based comparaison

Ingestion Management Dissemination

AIP Management

DSpace Fedora

Ingestion Management Dissemination

terça-feira, 7 de Fevereiro de 2012
Matching data models

DSpace

Diagram showing relationships between community, collection, item, bundle, bitstream format, and bitstream.
Matching data models

Fedora

- PID
- Description
- Items
- Services
If one of these levels becomes obsolete we loose access to the DO.
DO Preservation Strategies

• Focusing the physical/logical object
  o Centered in preserving information in her logical format or/and physical support
  o Uses original technology associated to these objects to ensure the access to them
  o Technology preservation

• Focusing the conceptual object
  o Centered in preserving the object core properties in a way that is independent from hardware and software
  o Conceptual object preservation
Conceptual object preservation

**Migration:** periodic DO transfer from one hw/sw configuration into an updated one (centered in preserving significant properties other than preserving the original bit stream).

**Advantages**
- DO are disseminated in formats known to users
- No need to preserve the original hw/sw platform
- Most used strategy and the only that has worked so far

**Disadvantages**
- Possible loss of information during conversion
- Continued maintenance is needed
- In the long-term perspective costs are high
Migration: periodic DO transfer from one hw/sw configuration into an updated one (centered in preserving significant properties other then preserving the original bit stream).

Advantages
- DO are disseminated in formats known to users
- No need to preserve the original hw/sw platform
- Most used strategy and the only that has worked so far

Disadvantages
- Possible loss of information during conversion
- Continued maintenance is needed
- In the longterm perspective costs are high

What are the significant properties?
Distributed Migration

• Remote conversion services
  o known APIs
  o descriptive metadata for localization and invocation (UDDI)

• Advantages
  o Platform independency
  o Redundancy/multiple migration paths
  o Compatible with other migration strategies
    ▪ Normalization, migration on request
  o Generalized cost reduction

• Disadvantages
  o Bandwidth requirements
  o Slow
Distributed Preservation Services

t_1

<.9, .8, .95, .1>

<.7, .5, .65, .1>

<.9, .8, .6, .1>

<.3, .6, .95, .1>

<.9, .6, .9, .7>

<.9, .6, .95, .6>

<.5, .3, .95, .6>

<.5, .3, .95, 1>

<.5, .3, .95, .1>
CRiB project: http://crib.dsi.uminho.pt
CRiB: architecture
Migration Broker

- Carries out format conversions
  - Invokes the necessary conversion services
- Measures the performance of the conversion process
  - Availability
  - Stability
  - Throughput
  - Scalability
  - Cost
  - Size ratio
  - File count ratio
Object Evaluator

- Determines the amount of data loss involved in migration
- Detects similarity between significant properties of digital objects
  - Object class dependent;
  - Different significant properties for bitmap images, text documents, relational databases, etc.
- Produces evaluation reports in PREMIS format
  - Datetime of intervention
  - Description of involved agents
  - Type of event (i.e. Migration)
  - Outcome of intervention
Object Evaluator

- Determines the amount of data loss involved in migration
- Detects similarity between significant properties of digital objects
  - Object class dependent;
  - Different significant properties for bitmap images, text documents, relational databases, etc.
- Produces evaluation reports in PREMIS format
  - Datetime of intervention
  - Description of involved agents
  - Type of event (i.e. Migration)
  - Outcome of intervention

Not implemented for databases
Text Documents and Still Images
Text Documents and Still Images

- EAD elements capture most of the significant properties: provenance, producer history, context, ...
- Content is kept in a normalized format: PDF and uncompressed TIFF.
EAD elements capture most of the significant properties: provenance, producer history, context, ...

Content is kept in a normalized format: PDF and uncompressed TIFF.
Text Documents and Still Images

- EAD elements capture most of the significant properties: provenance, producer history, context, ...
- Content is kept in a normalized format: PDF and uncompressed TIFF.
Databases
Databases

• How do we keep archived databases readable and usable in the long term (at acceptable cost)?
Databases

• 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?
Databases

• 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?
Databases

• 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?
Databases

- 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)
Databases

• 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?
Databases

• 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?
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?
Databases

• 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?

General big questions...
Databases
Databases

• What are the salient features of a database that should be preserved? (*significant properties*...*)
Databases

• What are the salient features of a database that should be preserved? *(significant properties...)*

• What are the different stages in the database preservation’s life cycle?
Databases

• What are the salient features of a database that should be preserved? *(significant properties…)*

• What are the different stages in the database preservation’s life cycle?

• What documentation is preserved together with a database, and in what format?
Databases

- What are the salient features of a database that should be preserved? (*significant properties...*)
- 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?
Databases

• What are the salient features of a database that should be preserved? *(significant properties...)*

• 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?
Databases

- What are the salient features of a database that should be preserved? (*significant properties.*)
- 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?
• What are the salient features of a database that should be preserved? \textit{(significant properties...)}

• 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 measure the quality of preservation strategies when they are applied to databases? \textit{(quality assurance...)}
Databases

- What are the salient features of a database that should be preserved? *(significant properties...)*
- 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 measure the quality of preservation strategies when they are applied to databases? *(quality assurance...)*

Technical questions...
Databases: Goals

• How do we store them?
• How do we access them?
Databases: Goals

- How do we store them?
- How do we access them?

RODA questions...
Databases

- Data?
- Structure?
- Views?
- Reports?
- Stored Procedures?
- ...

terça-feira, 7 de Fevereiro de 2012
Databases

Normal evolution path:
Data => Structure => Semantics

- Data?
- Structure?
- Views?
- Reports?
- Stored Procedures?
- ...

terça-feira, 7 de Fevereiro de 2012
Databases

- Data?
- Structure?
- Views?
- Reports?
- Stored Procedures?
- ...

Normal evolution path:
Data => Structure => Semantics

First prototype:
- Data
- Structure
- Only “frozen” DBs

derça-feira, 7 de Fevereiro de 2012
HOW?

The need for an intermediate representation

Input Formats (M)       Output Formats (N)
HOW?

The need for an intermediate representation

Input Formats (M)    Output Formats (N)

M*N migrators
IR: DBML

Input Formats (M)  Output Formats (N)
For each new output or input format you only need to code one new migrator.
DBML design Principles

• Hardware independent;
• Software independent;
• Easy to process;
• Descriptive;
• It should be possible to add metadata;
• It should be possible to add semantics;
DBML design Principles

- Hardware independent;
- Software independent;
- Easy to process;
- Descriptive;
- It should be possible to add metadata;
- It should be possible to add semantics;

XML was the obvious choice
SIP Structure (DB example)
SIP Structure (DB example)
SIP Structure (DB example)

- Manifest
- Descriptive Metadata (EAD):
  - producer
  - collection
  - ...
- DBML:
  - Data;
  - Structure;
  - Other metadata.
SIP Structure (DB example)

Binaries

Manifest

Descriptive Metadata (EAD):
- producer
- collection
- ...

DBML:
- Data;
- Structure;
- Other metadata.

KEEPS Employees
**SIP Structure (DB Example)**

- **Technical Metadata:**
  - color
  - dimensions
  - ...

- **Manifest**

- **Descriptive Metadata (EAD):**
  - producer
  - collection
  - ...

- **Binaries**

- **DBML:**
  - Data;
  - Structure;
  - Other metadata.

**KEEPS Employees**
SIP Structure (DB example)

Technical Metadata:
- color
- dimensions
- ...

Manifest

Descriptive Metadata (EAD):
- producer
- collection
- ...

Binaries

DBML:
- Data;
- Structure;
- Other metadata.
SIP Structure (DB example)

Technical Metadata:
- color
- dimensions
- ...

Manifest

Descriptive Metadata (EAD):
- producer
- colection
- ...

Binaries

DBML:
- Data;
- Structure;
- Other metadata.

KEEPS Employees

terça-feira, 7 de Fevereiro de 2012
**SIP Structure (DB example)**

**Technical Metadata:**
- color
- dimensions
- ...

**Manifest**

**Descriptive Metadata (EAD):**
- producer
- collection
- ...

**Binaries**

**DBML:**
- Data;
- Structure;
- Other metadata.

**KEEPS Employees**
SIP Structure (DB example)

- Technical Metadata:
  - color
  - dimensions
  - ...

- Manifest

- Descriptive Metadata (EAD):
  - producer
  - collection
  - ...

- Binaries

- DBML:
  - Data;
  - Structure;
  - Other metadata.
SIP Structure (DB example)

Technical Metadata:
- color
- dimensions
- ...

Manifest

Descriptive Metadata (EAD):
- producer
- collection
- ...

Binaries

DBML:
- Data;
- Structure;
- Other metadata.
SIP Structure (DB example)

Compressed File

Technical Metadata:
- color
- dimensions
- ...

Manifest

Descriptive Metadata (EAD):
- producer
- collection
- ...

Binaries

DBML:
- Data;
- Structure;
- Other metadata.

KEEPS Employees

terça-feira, 7 de Fevereiro de 2012
DBML
DBML: Structure
DBML: Data
Databases: RODA Answers

- How do we store them?
  ★ DBML + binaries + technical metadata

- How do we access them?
  ★ PhpMyAdmin (hacked version)
Databases: RODA Answers

- How do we store them?
  ★ DBML + binaries + technical metadata

- How do we access them?
  ★ PhpMyAdmin (hacked version)
DBML: input converters

- SQL Server (tested on 2005 version);
- Postgres (tested, demo tomorrow);
- MySQL (tested, demo tomorrow);
- DBML (tested, demo tomorrow);
- MS Access (tested, demo tomorrow);
- ODBC (tested, with problems...).

Input being harvested from...
DBML: output converters

- SQL Server (tested on 2005 version);
- Postgres (tested, demo tomorrow);
- MySQL (tested, demo tomorrow);
- DBML (tested, demo tomorrow);
- MS Access (tested, demo tomorrow);
- PhpMyAdmin optimized SQL (tested, ...);
- Postgres SQL.

Output format being generated...
Some problems

• Extracting data is easy:
  ✦ SELECT * FROM ...

• Extracting the structure is not:
  ✦ DBMS protect this information;
  ✦ Each DBMS stores it differently;
  ✦ Different versions of the same DBMS can also act differently;
  ✦ We have to “prepare/hack” the DBMS.
SIP Builder

Diagram showing the process of creating a SIP (Service Integration Pattern) with mediators for SQL Server, Access, and Oracle, involving metadata, DBML, and files in the SIP Packaging stage.
Real Scenario

• Many archives have ARQBASE/WINISIS databases;
• Others have CALM databases;
• They want to migrate to a newer system;
• Those systems export XML.

With DBML we are one XSLT stylesheet distance from solving the problem...
Disseminators
Metadata Editor
Preservation Metadata Viewer

Integrity Check
Object integrity was successfully checked
Tue, 15 Jan 2008 01:00:00 GMT
Tue, 15 Jan 2008 02:00:01 GMT
Advanced Search
Final thoughts

“Data Preservation is a people problem”

Michael Lesk
Final thoughts

“Data Preservation is a people problem”
Michael Lesk

• People need to be trained to save data in a proper way.
• What to preserve? Data, Structure, Semantics...
• Preservation is for future users but only today users vote on budget
• We need to make data collecting people have preservation concerns
• Preservation is fault tolerance. All systems are imperfect
Business Model

- CRiB is free: mferreira@keep.pt
- RODA is free: property of Portuguese National Archives
- RODA community is being created
- Support, maintenance and internationalization: KEEP Solutions Lda, www.keep.pt
Let’s Preserve Tomorrow’s History...

Questions?

José Carlos Ramalho
jcr@keep.pt