Database preservation toolkit:

a flexible tool to normalize and give access to databases
KEEP SOLUTIONS: Projects

- DigitArq, CRAV (2003..[2008-2012])
- RODA (2006..[2008-...[]
- RCAAP (2008-...)
- PPA (2009)
- Open source: RODA, KOHA, DSpace, Moodle, etc.
- Scientific research (european projects)
  - **SCAPE**: Large scale preservation
  - **4C**: the cost for curation
  - **e-arK**: presented by Kuldar

[http://www.keep.pt](http://www.keep.pt)
Developed within RODA project

Now stand-alone open-source project

http://keeps.github.io/db-preservation-toolkit/

Imports and exports between DBMS and DB formats

Supports preservation formats: DBML, SIARD
The Past: 2006 - 2009

RODA:
Repositório de Objectos Digitais Autênticos

José Carlos Ramalho
jcr@di.uminho.pt

Miguel Ferreira
mferreira@dsi.uminho.pt

Rui Castro
rcastro@ianit.pt

Luis Faria
lfaria@iantt.pt

Francisco Barbedo
frbarbedo@ianit.pt

Cecília Henriques
chenriques@ianit.pt

Luis Corujo
lcorujo@ianit.pt
What is RODA?

A digital repository created for archives, with the following features:

- Long term preservation and Autenticity
- Standards based (OAIS, EAD, PREMIS, METS, etc.)
- Following TRAC demands (Trustworthy Repositories Audit & Certification)
- Security policies
- Service Oriented Architecture (SOA)
- Nice and simple design
- Open source
- DGArq and University of Minho
**Context**

RODA (2006-2009)
- Metadata management (EAD)
- Digital Object management (...)
- Digital Preservation Policies and protocols
- National Project (Archives National Board)

CRiB: Preservation Services Digital Repositories (2005-2008)
- Distributed Migration Service
- Migration Service supported by knowledge base
- phd Thesis / U. Minho (Miguel Ferreira)
**Context**

**RODA (2006-2009)**
- Metadata management (EAD)
- Digital Object management (...)
- Digital Preservation Policies and protocols
- National Project (Archives National Board)

**CRiB: Preservation Services Digital Repositories (2005-2008)**
- Distributed Migration Service
- Migration Service supported by knowledge base
- phd Thesis / U. Minho (Miguel Ferreira)
Architecture
Architecture

Being replaced
Data Model
Data Model
Data Model
Digital Object Classes
Digital Object Classes
Digital Object Classes

Normalization
Digital Object Classes

Normalization

Significant properties
Ingest
SIP structure

- Compressed ZIP file
  - `<METS>` envelope
  - Descriptive metadata `<EAD-Component>`
  - Preservation metadata `<PREMIS>`
  - Representation
    - File 1
    - File 2
    - ...
    - File n
SIP structure

- Object class and format
SIP structure

- Object class and format
- Place in classification plan
SIP structure

- Object class and format
- Place in classification plan
- Descriptive metadata
SIP structure

- Object class and format
- Place in classification plan
- Descriptive metadata
- Preservation metadata
SIP structure

- Object class and format
- Place in classification plan
- Descriptive metadata
- Preservation metadata
- Technical metadata
SIP structure

- Object class and format
- Place in classification plan
- Descriptive metadata
- Preservation metadata
- Technical metadata
- Representation files
SIP structure

- Object class and format
- Place in classification plan
- Descriptive metadata
- Preservation metadata
- Technical metadata
- Representation files
- Identification of root file
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?
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?
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?
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?
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 data-bases? What are DB significant properties? (*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 *data-*bases? What are *DB significant properties*? *(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
Databases

Normal evolution path:
Data => Structure => Semantics
Databases

Normal evolution path:
Data => Structure => Semantics

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

• Data?

Normal evolution path:
Data => Structure => Semantics

First prototype:
• Data
• Structure
• Only “frozen” DBs
Databases

- Data?
- Structure?

Normal evolution path:
Data $\Rightarrow$ Structure $\Rightarrow$ Semantics

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

- Data?
- Structure?
- Views?

Normal evolution path:
Data => Structure => Semantics

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

- Data?
- Structure?
- Views?
- Reports?

Normal evolution path:
Data => Structure => Semantics

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

- Data?
- Structure?
- Views?
- Reports?
- Stored Procedure?

Normal evolution path:
Data => Structure => Semantics

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

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

Normal evolution path:
Data => Structure => Semantics

First prototype:
- Data
- Structure
- Only “frozen” DBs
The need for an intermediate representation

Input Formats (M)  Output Formats (N)

MySQL  PostgreSQL  ORACLE
Microsoft SQL Server  ORACLE
Excel

Encoded Archival Description
Official Web Site
Network Development & MARC Standards Office
Library of Congress
The need for an intermediate representation

Input Formats (M)  Output Formats (N)

M*N migrators

HOW?
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)

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

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

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

Manifest
SIP Structure (DB example)

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

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

- Binaries

- Manifest

KEEPS Employees
SIP Structure (DB example)

Technical Metadata:
- color
- dimensions
- ...

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

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

Technical Metadata:
- color
- dimensions
- ...

Manifest

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

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

Binaries

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.

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)

Keeps Employees

Technical Metadata:
- color
- dimensions
- ...

Binaries

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

Manifest

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

KEEPs Employees
SIP Structure (DB example)

Compressed File

- Manifest

Technical Metadata:
- color
- dimensions
- ...

Binaries

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

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

KEEPS Employees
DBML

Diagram: 
- `db` connected to `@ attributes`
- `structure` connected to `data`

DBML: Structure
DBML: Data

Diagram:

- data
  - tableData
    - row
      - id
      - cell
        - id

- Container for the row data. A row has a list of cells.
- Container for the cell data. The data corresponds to a simple or composite type. All types are nillable.
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)
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.
Two approaches emerged:

- Preserve the database and the environment allowing authentic access to the information and any accompanying applications

- Extract / migrate the raw data and table structure from the original database
Open Planets: “Database Archiving” (Copenhagen 2012)

Two approaches emerged:

• Preserve the database and the environment allowing authentic access to the information and any accompanying applications
  
  Relies on emulation strategy

• Extract / migrate the raw data and table structure from the original database
Open Planets: “Database Archiving” (Copenhagen 2012)

Two approaches emerged:

- Preserve the database and the environment allowing authentic access to the information and any accompanying applications. **Relies on emulation strategy**

- Extract / migrate the raw data and table structure from the original database. **Relies on migration strategy**
Open Planets: “Database Archiving” (Copenhagen 2012)

Two approaches emerged:

- Preserve the database and the environment allowing authentic access to the information and any accompanying applications

- Extract / migrate the raw data and table structure from the original database

RODA+DBML
Two approaches emerged:

- **Preserve the database and the environment allowing authentic access to the information and any accompanying applications.**
- **Extract / migrate the raw data and table structure from the original database.**

**Relies on emulation strategy**

**Relies on migration strategy**

RODA+DBML

SIARD
Two approaches emerged:

- **Preserve the database and the environment allowing authentic access to the information and any accompanying applications**

  - Relies on emulation strategy

- **Extract / migrate the raw data and table structure from the original database**

  - Relies on migration strategy

  - RODA+DBML
  - SIARD
  - SIARD-DK
Two approaches emerged:

- Preserve the database and the environment allowing authentic access to the information and any accompanying applications (Relies on emulation strategy)

- Extract / migrate the raw data and table structure from the original database (Relies on migration strategy)

- RODA+DBML
- SIARD
- SIARD-DK
- ADDML
<table>
<thead>
<tr>
<th>Feature</th>
<th>DBML</th>
<th>SIARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Yes <em>(no segmentation)</em></td>
<td>Yes <em>(with segmentation)</em></td>
</tr>
<tr>
<td>Structure</td>
<td>Yes <em>(XML)</em></td>
<td>Yes <em>(XSD)</em></td>
</tr>
<tr>
<td>Stored Procedures</td>
<td>No</td>
<td>Yes <em>(XML)</em></td>
</tr>
<tr>
<td>Triggers</td>
<td>No</td>
<td>Yes <em>(XML)</em></td>
</tr>
<tr>
<td>Views</td>
<td>No</td>
<td>Yes <em>(XML)</em></td>
</tr>
</tbody>
</table>
db-preservation-toolkit architecture

Import modules
- MySQL
- Oracle'12
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC

Streaming data model

Export modules
- MySQL
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC
- DBML
- SIARD
- PhpMyAdmin
db-preservation-toolkit architecture

Import modules
- MySQL
- Oracle'12
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC

Streaming data model

Export modules
- MySQL
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC
- DBML
- SIARD
- PhpMyAdmin

ADDML
Pre-ingest / Production / SIP creation

Origin formats:
- MySQL
- Oracle'12
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC

db-preservation-toolkit

Preservation formats:
- DBML
- SIARD
Pre-ingest / Production / SIP creation

Origin formats

MySQL
Oracle'12
SQL Server
PostgreSQL
DB2
MS Access
ODBC

db-preservation-toolkit

Preservation formats

DBML
SIARD
Pre-ingest / Production / SIP creation

![Diagram showing origin formats and preservation formats connected through db-preservation-toolkit. Origin formats include MySQL, Oracle'12, SQL Server, PostgreSQL, DB2, MS Access, and ODBC. Preservation formats include DBML and SIARD.]
Pre-ingest / Production / SIP creation

Origin formats
- MySQL
- Oracle'12
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC
- ADDML

Preservation formats
- DBML
- SIARD
Pre-ingest / Production / SIP creation

Origin formats
- MySQL
- Oracle'12
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC
- ADDML

Preservation formats
- DBML
- SIARD

db-preservation-toolkit

new

to be implemented
Access

Working on RDF viewer, we don’t need to stay stuck in the relational model
New prototype

Msc thesis: “Database Convert Tool: exploring SIARD”

OWL model development

Triple Store Implementation

RDF endpoint implementation

SPARQL based access
Future plan:
Fast database viewer
Recent work

Phd thesis: Graph view dissemination, relational model reverse engineering, algorithm development (from relational to graph model) [http://repositorium.sdum.uminho.pt/handle/1822/25655]

Msc thesis: SIARD support (DBML replacement)

Msc thesis: RDF representation for databases
Questions?