RODA: digital preservation for the portuguese public administration

José Carlos Ramalho
jcr@di.uminho.pt

Miguel Ferreira
mferreira@dsi.uminho.pt

Rui Castro
Rcastro@iantt.pt

Luis Faria
lfaria@iantt.pt

Francisco Barbedo
frbarbedo@iantt.pt

Cecília Henriques
chenriques@iantt.pt

Glória Santos
gloria@iantt.pt

Luis Corujo
lcorujo@iantt.pt
Digitarq (2003-now)
• metadata management (EAD based)
• digital object management (NISO MIX)

RODA (2006-2008)
• metadata management (EAD based)
• digital object management (...)
• digital preservation protocols and policies

CRAV: Readers Virtual Room (2006-2007)
• request management
• document workflow
Context

DigitArq (2003-now)
• metadata management (EAD based)
• digital object management (NISO MIX)

RODA (2006-2008)
• metadata management (EAD based)
• digital object management
• digital preservation protocols and policies

CRAV: Readers Virtual Room (2006-2007)
• request management
• document workflow

Partners/Contracters:
• National Directory Board of Archives
• Photography National Archive
• Oporto’s county Archive
• Some city hall archives (can grow exponentially)
RODA: Motivation

• Today History is being made in the digital world;
• Digital Object production grows everyday;
• There are no structures to support incorporation, management and long-term preservation of digital objects;
• We have to preserve the digital memory, heritage and testimonials of public organizations.
  • Example: SGU work
Some Requisites/Questions?

• How do we achieve Authenticity?
• How do we describe and classify DO?
• How can we implement digital preservation?
“O Codex 632” by José Rodrigues dos Santos

Subject: Who really was Cristophoros Colombo? 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?
Authenticity

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
Digital Object Classes

- **PIXAR**
- **JPX**
- **PNG**
- **PICT**
- **PICT-R**
- **JP2**
- **JPEG**
- **EXR**
- **PHOTO**
- **TARGA**
- **EPS**
- **HDR**
- **PBM**
- **SCITEX**
- **WBMP**
- **GIF**
- **TIFF**
- **PDF**
- **XML**
- **LaTeX**
DO Anatomy

Conceptual level

Logical level

Physical level

Database
Text Doc.
Still Image

SQL Server
Access
PDF Doc.
Ms Word Doc.
PNG image

Hard Disc
Tape
...

...
If one of these levels becomes obsolete we loose access to the DO
DO Preservation Strategies

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

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

Emulator: application capable of reproducing the behaviour of an hardware/software platform. Ex: ZX Spectrum, GBA, ...
Emulator: application capable of reproducing the behaviour of an hardware/software platform. Ex: ZX Spectrum, GBA, ...

- Advantages
  - Original technological context recreation
  - Object’s look & feel preservation

- Disadvantages
  - Emulators also become obsolete
  - Users have to operate obsolete systems
  - Creating emulators is a complex task
  - Copyright problems
  - To preserve a complete operating system to be able to visualize a single document may be overwhelming
  - Information reuse is not guaranteed
Encapsulation
Encapsulation

Preserving the original bit stream together with enough metadata capable of ensuring its future interpretation and access.
Preserving the **original bit stream** together with enough metadata capable of ensuring its future interpretation and access

### Advantages
- It allows the postponement of preservation responsibilities
- Targeted for objects that will be accessed in a far future
- Emulator and visualizer development is delayed

### Disadvantages
- Complex objects have complex specifications
- An incomplete specification can have nasty effects
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 longterm perspective costs are high
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

What are the significant properties?

Disadvantages
- Possible loss of information during conversion
- Continued maintenance is needed
- In the longterm perspective costs are high
Preservation Services
CRiB project: http://crib.dsi.uminho.pt
Open Archival Information System

ISO 14721
**OAIS (Functional Components)**

- **Ingestion**
  - Reception, validation, transformation/normalization, description of the whole package submitted by the producer

- **Storage**
  - Ensures information preservation at physical/logical level (e.g. refreshing, migration, integrity checks, disaster recovery, etc.)

- **Metadata management**
  - Responsible for the management of stored DOs
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

Producer

Negotiation

Archive

Submission Contract
Ingestion

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

one still image
SIP Structure (Example)

one still image

creation

properties:
- date
- hardware
- ...

<XML>

---------------------

---------------------

---------------------

---------------------
SIP Structure (Example)

- one still image

Technical Metadata:
- color
- dimensions
- ...

Creation properties:
- date
- hardware
- ...

XML
SIP Structure (Example)

- one still image

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

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

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

First image: One still image

Second image: XML structure for metadata
SIP Structure (Example)

- one still image
- creation properties:
  - date
  - hardware
  - ...
- Technical Metadata:
  - color
  - dimensions
  - ...
- Descriptive Metadata:
  - producer
  - colection
  - ...
- Manifest
SIP Structure (Example)

Compressed File

one still image

Creation properties:
- date
- hardware
- ...

Technical Metadata:
- color
- dimensions
- ...

Descriptive Metadata:
- producer
- collection
- ...

Manifest
SIP Structure (+complex)
SIP Structure (+complex)
SIP Structure (+complex)

DO = Image+
SIP Structure (+complex)

DO = Image+ Properties
SIP Structure (+complex)

DO = Image+

Properties

Technical Metadata
SIP Structure (+complex)

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

Manifest

DO = Image+
Properties
Technical Metadata
Descriptive Metadata
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)
AIP Storage

metadata

DOs

AIP
Normalization
Stages

• Analysis and Planning
• Prototyping
• Testing and Dissemination
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

fedora

arno

academic publications online

ITOR

<MyCoRe>

Archimède

DSPACE
Requisites based comparison

Ingestion
AIP Management
Dissemination

DSpace
Fedora

Ingestão
Gestão de AIP
Disseminação
Matching data models

DSpace

Diagram showing relationships between Community, Collection, Item, Bundle, Bitstream Format, and Bitstream.
Matching data models

Fedora

| PID | FOXML | Object Properties | Relationship Metadata | DC | EAD | Default Disseminator |

Descrição

Itens

Serviços

Diagram:

- $OD_2$
- $OD_{SR}$
- $OD_{DC}$
- $OD_D$
- $QR_{Original}$
- $QR_{Normalized}$
- $OP_{Rep+Files}$
- $OP_{Event}$
- $OP_{Agent}$
Roda Data Model
Roda Data Model

Description Objects
Roda Data Model

Description Objects

Representation Objects

Preservation Objects
Architecture

- RODA's Web User Interface
- Content adaptor
- RODA Information Services
- Fedora Services
- Fedora Generic Search
- Web services communication interface
RODA Schemas
Preserving Conceptual Object

Conceptual level
- Database
- Text Doc.
- Still Image

Logical level
- SQL Server
- Access
- PDF Doc.
- Ms Word Doc.
- PNG image

Physical level
- Hard Disc
- Tape
- ...

Conceptual level
Logical level
Physical level
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.
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.
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

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

37
Databases

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

First prototype:
- Data
- Structure
DBML

- Platform and RDBMS independent
- Stores the DB structure and information
- BLOBs are exported and preserved as standalone files in the representation
- Transformations to SQL and back are defined
DBML

- Platform and RDBMS independent
- Stores the DB structure and information
- BLOBs are exported and preserved as standalone files in the representation
- Transformations to SQL and back are defined
**DBML**

- Platform and RDBMS independent
- Stores the DB structure and information
- BLOBs are exported and preserved as standalone files in the representation
- Transformations to SQL and back are defined

```xml
<TABLE NAME="Districts">
  <COLUMNS>
    <COLUMN NAME="code" TYPE="int" NULL="no"/>
    ...
  </COLUMNS>
  <KEYS>
    <PKEY TYPE="simple">
      <FIELD NAME=""/>
    </PKEY>
    <PKEY TYPE="compound">
      <FIELD NAME=""/>
      <FIELD NAME=""/>
    </PKEY>
  </KEYS>
</TABLE>

<DATA>
  ...
  <products>
    <products-REG>
      <code>a122</code>
      <description>milk</description>
      ...
    </products-REG>
    <products-REG>
      ...
    </products-REG>
  </products>
  ...
</DATA>
```

39
DB SIP composition:

- METS file for packaging and organizing
- EAD file describing intellectual properties
- DBML file(s)
- DO for each found BLOB
- METS file + MIX for each DO
SIP -> AIP

- Check and validation ...
- Generate SQL file
- Generate PREMIS
Dissemination

- Abstract Database Creation: a database of databases... Ingests databases from DBML (DBML-->SQL_{adb});
- Specific Database Creation: execute the SQL file in the selected RDMS
Dissemination

AIP
- Metadata
- DBML
- Files

Access

DIP

SQL

MySQL

Database Browser
## RODA
Repositorio de Objectos Digitais Autênticos

<table>
<thead>
<tr>
<th>Caminho:</th>
<th>Lista de fundos</th>
<th>Navegador</th>
</tr>
</thead>
<tbody>
<tr>
<td>AACC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### PT/TT/AACC/1/2

#### Identificação
- Referência: PT/TT/AACC/1/2
- Título: ILIDIO S.COELHO.2
- Descrição física: extent: 7

#### Conteúdo e Estrutura
- Ámbito e Conteúdo: Generalidades

#### Organização e ordenação:

<table>
<thead>
<tr>
<th>Volume</th>
<th>Página Inicial</th>
<th>Página Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sumário</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Processo</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

#### Disseminações
- Download da representação
- Visualizar base de dados
Browser

Database repository gui

Bibliography:

<table>
<thead>
<tr>
<th>ComponentID</th>
<th>ID</th>
<th>BibRef</th>
</tr>
</thead>
<tbody>
<tr>
<td>790950</td>
<td>80770</td>
<td>null</td>
</tr>
<tr>
<td>790950</td>
<td>80771</td>
<td>null</td>
</tr>
</tbody>
</table>
RODA
Repositório de Objectos Digitais Autênticos

Localizar resultados

com todos os campos:

Título

Nível de descrição

Datas extremas

com pelo menos um dos campos:

Título

com nenhum dos campos:

Título

Encontrados 58 resultados, página 1 de 4, 15 resultados por página:

Pontuação: 100%
Nível de descrição: DC

Pontuação: 100%
Nível de descrição: DC

Pontuação: 100%
Nível de descrição: DC

Pontuação: 100%
Nível de descrição: DC
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.
Look and see how our brothers are working to transfer all our writings into CDROM format.
Society of Archivists 2007 Conference

Submetido por Ifaria em Quarta, 18/10/2006 - 10:57 :: Arquivística

Start: 28/08/2007 - 08:00
End: 31/08/2007 - 17:00

Timezone: Etc/GMT

The Society of Archivists is hosting a major international conference in 2007 - Differing Directions: Challenging Communities

The Conference will be held at Queen’s University Belfast, Northern Ireland on 28 - 31 August 2007. The Society of Archivists (UK and Ireland) covers archivists, records and information management, and archive conservators - and the programme will include three parallel sessions relating to all these professional areas.

For more information see the Call for Papers

---

Participação do RODA no 4ª Congresso Nacional de Administração Pública

Submetido por roatro em Terça, 17/10/2006 - 16:38 :: Arquivística

Start: 02/11/2006 - 07:00
Timezone: Etc/GMT+1

A Administração Pública portuguesa está a viver dias de mudança profunda. Mudança
Let’s Preserve Tomorrow’s History...

Questions?