RODA and CRiB: A Service-Oriented Digital Repository

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

Luis Corujo
lcorujo@iantt.pt
**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)
**PREMISES**

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

- **CRiB:**
  - Distributed Migration
  - SOA
  - Quality control
  - Migration Adviser
Authenticity

“O Codex 632” by José Rodrigues dos Santos

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
DO Classes: normalization
Open Archival Information System (OAIS)
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
<XML>
  
  
  
  
</XML>
```
SIP Structure (example)

- one still image
- Technical Metadata:
  - color
  - dimensions
  - ...
- creation properties:
  - date
  - hardware
  - ...

**SIP Structure (example)**

One still image

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

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

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

**XML**
SIP Structure (Example)

One still image

Creation properties:
- date
- hardware
- ...

Technical Metadata:
- color
- dimensions
- ...

Descriptive Metadata:
- producer
- collection
- ...

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)

Manifest

DO = Image+

Properties

Technical Metadata

Descriptive Metadata

A

B

C
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)
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

fedora

arno

<i>TOR</i>

<MyCoRe>

Archimède

<i>eJprints</i>

<i>DSpace</i>
Requisites based comparison

- Ingestion
- Management
- Dissemination

- AIP Management
- Dissemination
Matching data models

DSpace
DO Anatomy

Conceptual level
Database
Text Doc.
Still Image

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

Physical level
Hard Disc
Tape
...

...
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 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
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

**What are the significant properties?**
Distributed Migration

- Remote conversion services
  - known APIs
  - descriptive metadata for localization and invocation (UDDI)
- Advantages
  - Platform independency
  - Redundancy/multiple migration paths
  - Compatible with other migration strategies
    - Normalization, migration on request
  - Generalized cost reduction
- Disadvantages
  - Bandwidth requirements
  - Slow
Distributed Preservation Services

Diagram with arrows and numbers.
Distributed Preservation Services

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
RODA+CRiB

CRiB (Preservation)

Ingest
Fedora
Access
Archival Storage
Administration
Data Management
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?
- ...

01/04/2006
Databases

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

First prototype:
- Data
- Structure
Browser
Disseminators
Metadata Editor
Advanced Search

Selecionar os níveis de descrição que incluir na pesquisa:
- todos os níveis
- escolher níveis

Selecionar o intervalo de tempo em que incluir na pesquisa:
- sem restrições temporais
- especificar intervalo

Preencher as palavras que incluir na pesquisa:
- todos os campos
- escolher campos

1 resultados obtidos
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
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_rastra_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
Business model

- CRiB is free: mferreira@keep.pt
- RODA is *free*: property of Portuguese National Archives
- Support, maintenance and internationalization: KEEP Solutions Lda, [www.keep.pt](http://www.keep.pt)
Let’s Preserve Tomorrow’s History...

Questions?