RODA and CRiB: A Service-Oriented Digital Repository

José Carlos Ramalho
jcr@di.uminho.pt

Francisco Barbedo
frbarbedo@iantt.pt

Miguel Ferreira
mferreira@dsi.uminho.pt

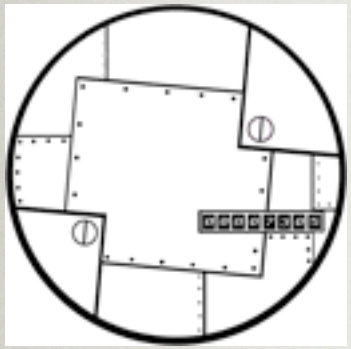
Cecília Henriques
chenriques@iantt.pt

Rui Castro
Rcastro@iantt.pt

Luis Corujo
lcorujo@iantt.pt

Luis Faria
lfaria@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)

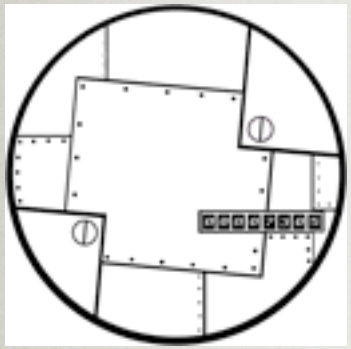
CRiB

Conversion and Recommendation of Digital Object Formats

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

Conversion and Recommendation of Digital Object Formats

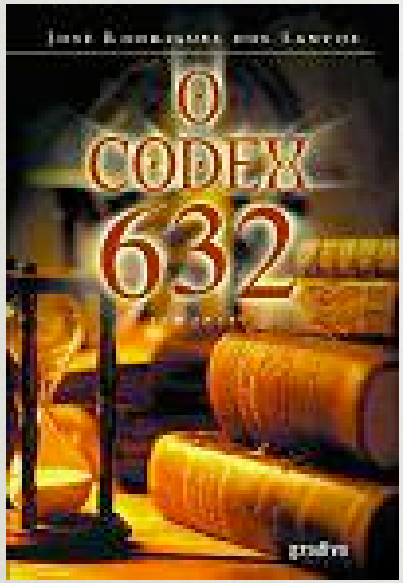
CRiB: Preservation Services for Digital Repositories (2005-2008)

- 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

Authenticity



"O Codex 632" by José Rodrigues dos Santos

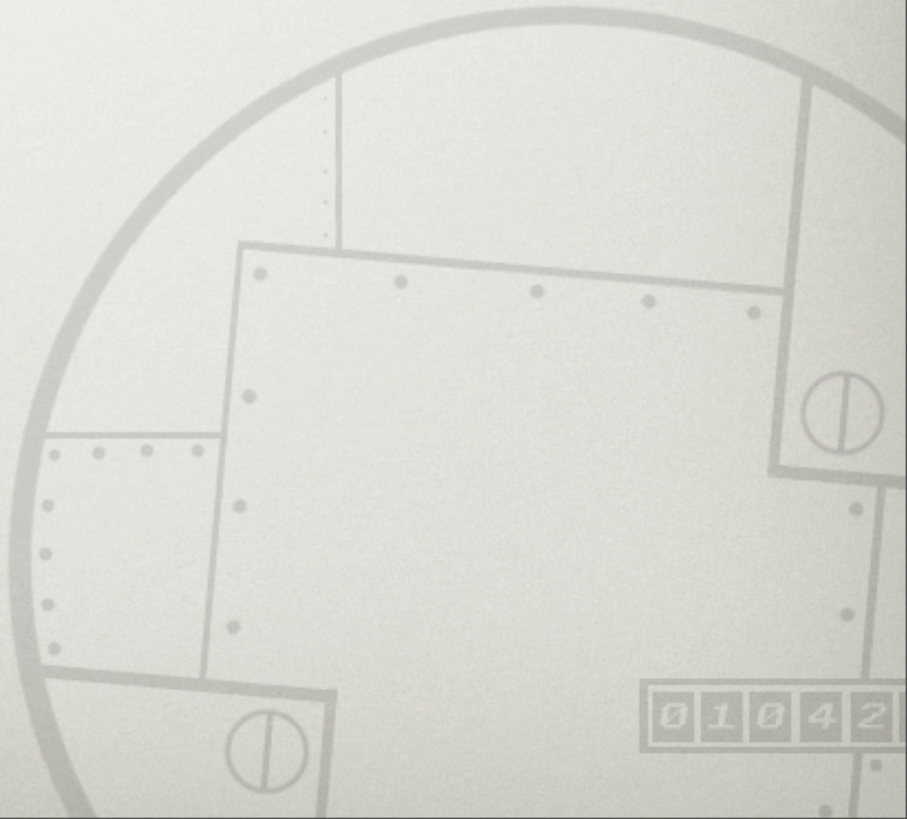
Subject: Who really was Cristophoros Colombus?

Was he italian? Spanish? Or a portuguese belonging to a jewish family?

Authenticity

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

EX: the bible



Authenticity

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

EX: the bible

How do we become trustful?

Authenticity

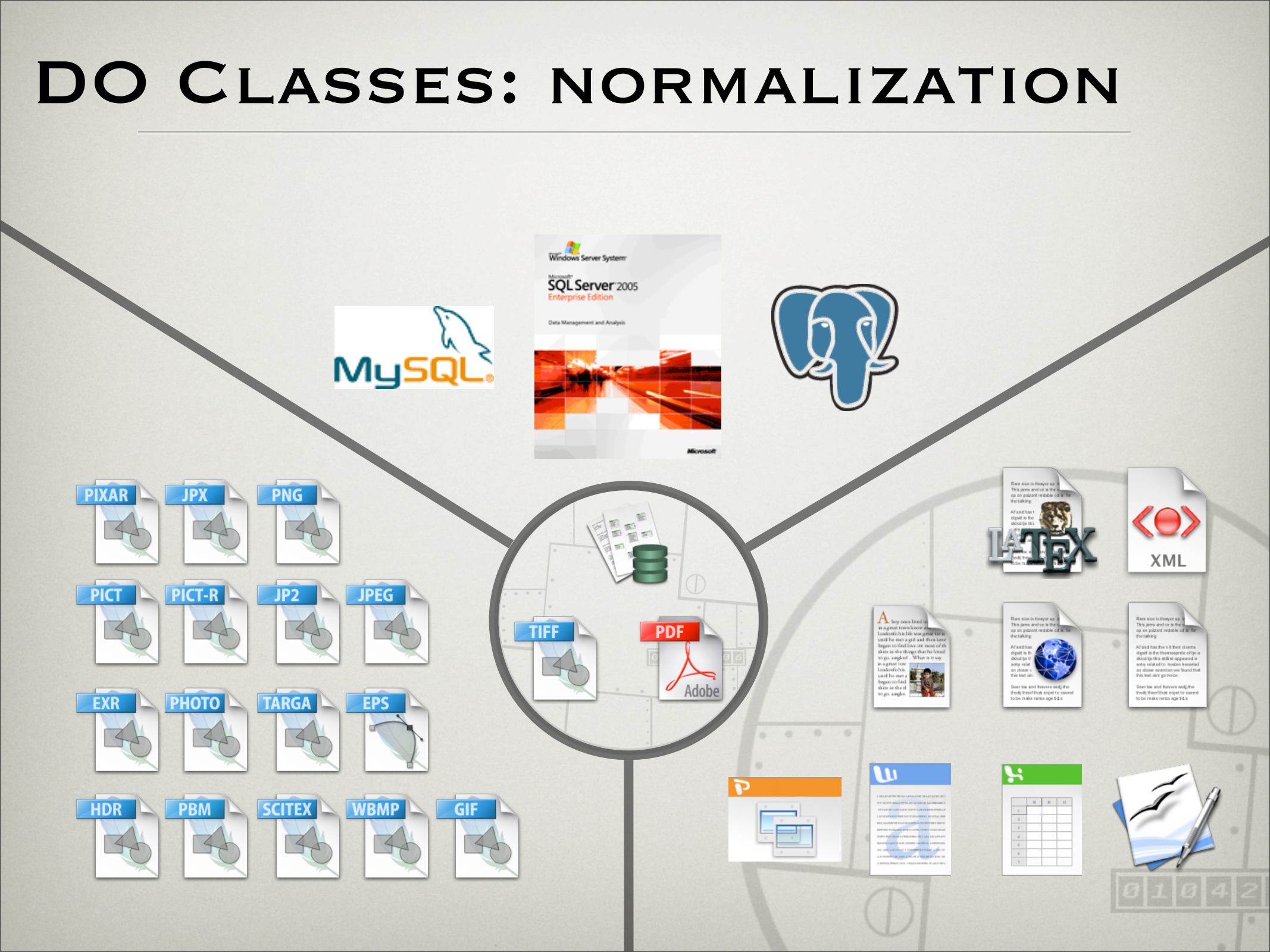
We must trust our sources: in ancient History there are no direct speech or evidence.

EX: the bible

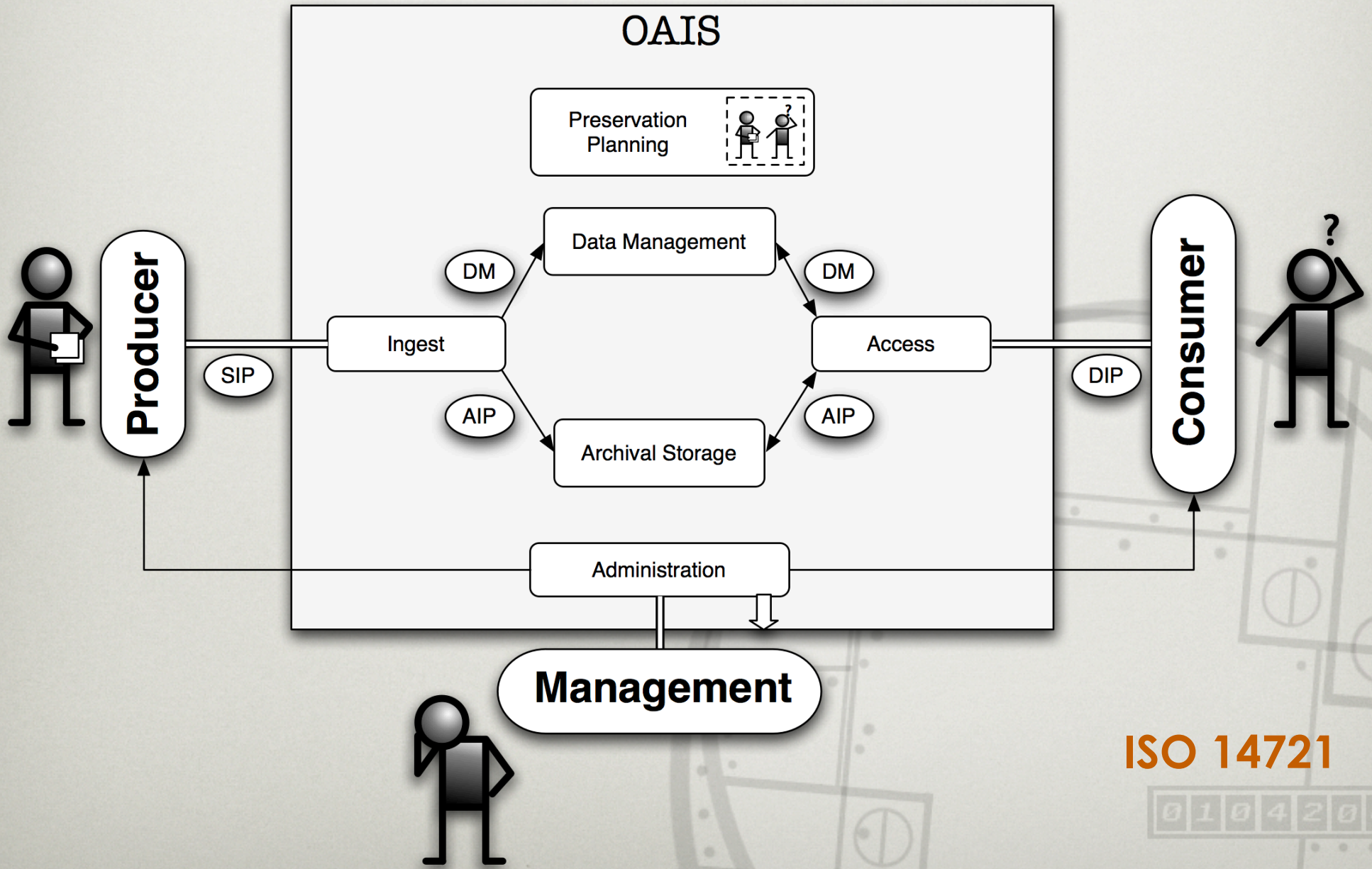
How do we become trustful?

- Reputation
- Documenting every action taken upon DOs

DO Classes: normalization



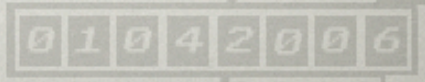
Open Archival Information System



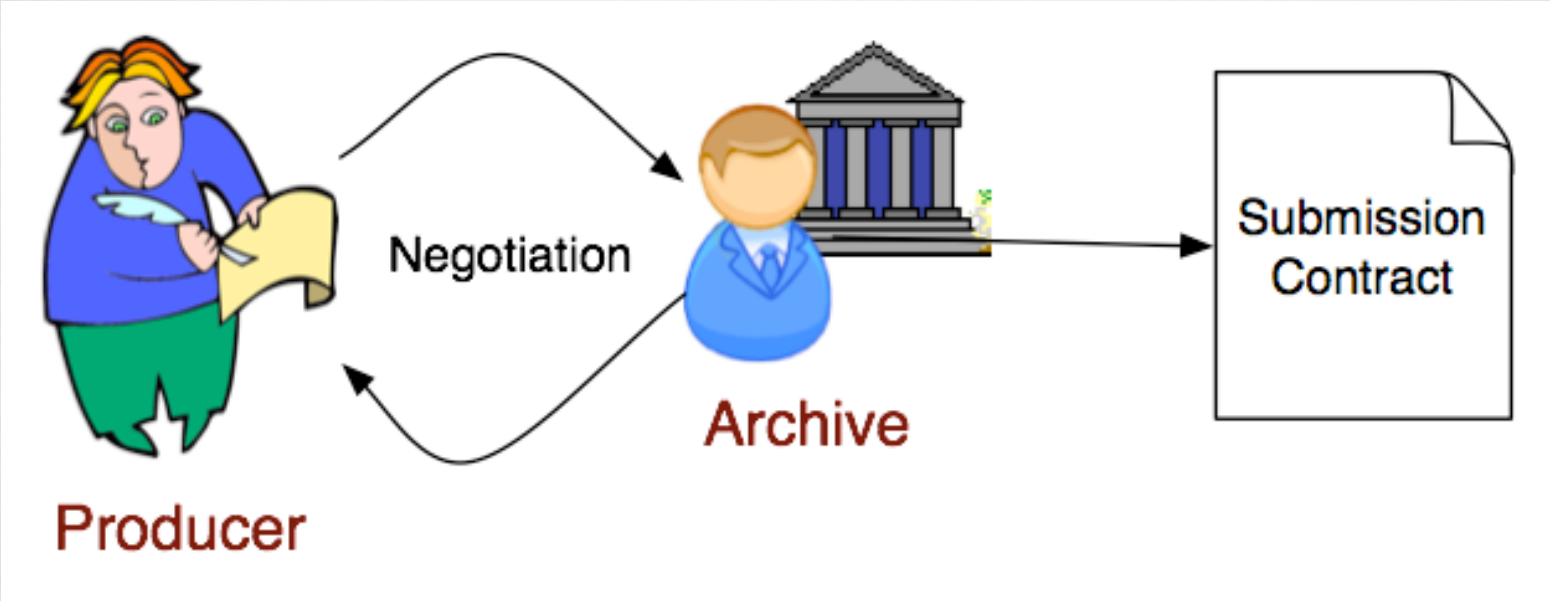
ISO 14721

OAIS (Information Packages

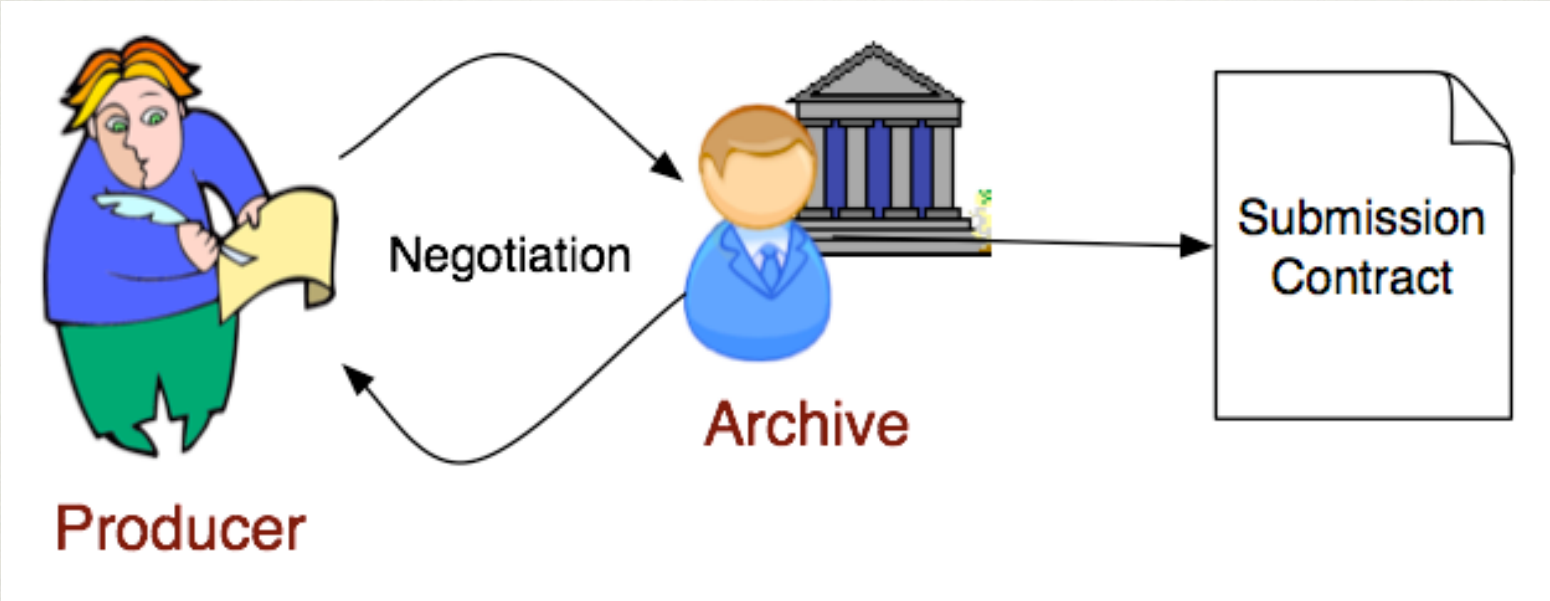
- Submission Information Package (SIP)
  - **Digital Object**
  - **Metadata created by producer**
    - **too open...**
- Archival Information Package (AIP)
  - **Digital Object to be stored**
  - **Metadata:** enough to ensure DO's preservation and access
    - model defined by PREMIS
- Dissemination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



Ingestion



Ingestion



Submission Contract

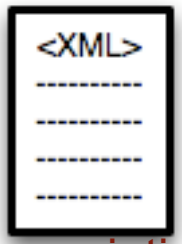
- SIP specification
- Ingestion workflow specification

SIP Structure (example)

one still image

SIP Structure (example)

one still image



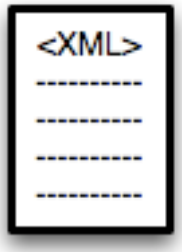
criation

properties:

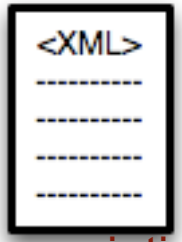
- date
- hardware
- ...

SIP Structure (example)

one still image



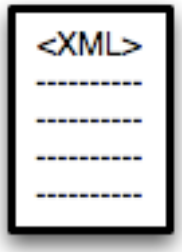
Technical Metadata:
- color
- dimensions
- ...



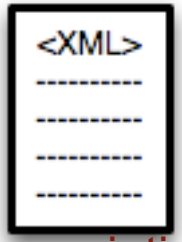
criation properties:
- date
- hardware
- ...

SIP Structure (example)

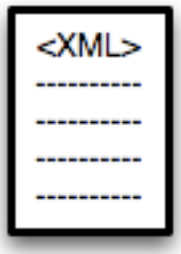
one still image



Technical Metadata:
- color
- dimensions
- ...



criation properties:
- date
- hardware
- ...



Descriptive Metadata:
- producer
- colection
- ...

SIP Structure (example)

one still image

<XML>

Manifest

<XML>

Technical Metadata:
- color
- dimensions
- ...

<XML>

---

criation properties:
- date
- hardware
- ...

<XML>

Descriptive Metadata:
- producer
- colection
- ...

SIP Structure (example)

one still image

<XML>

Manifest

<XML>

Technical Metadata:
- color
- dimensions
- ...

<XML>

---

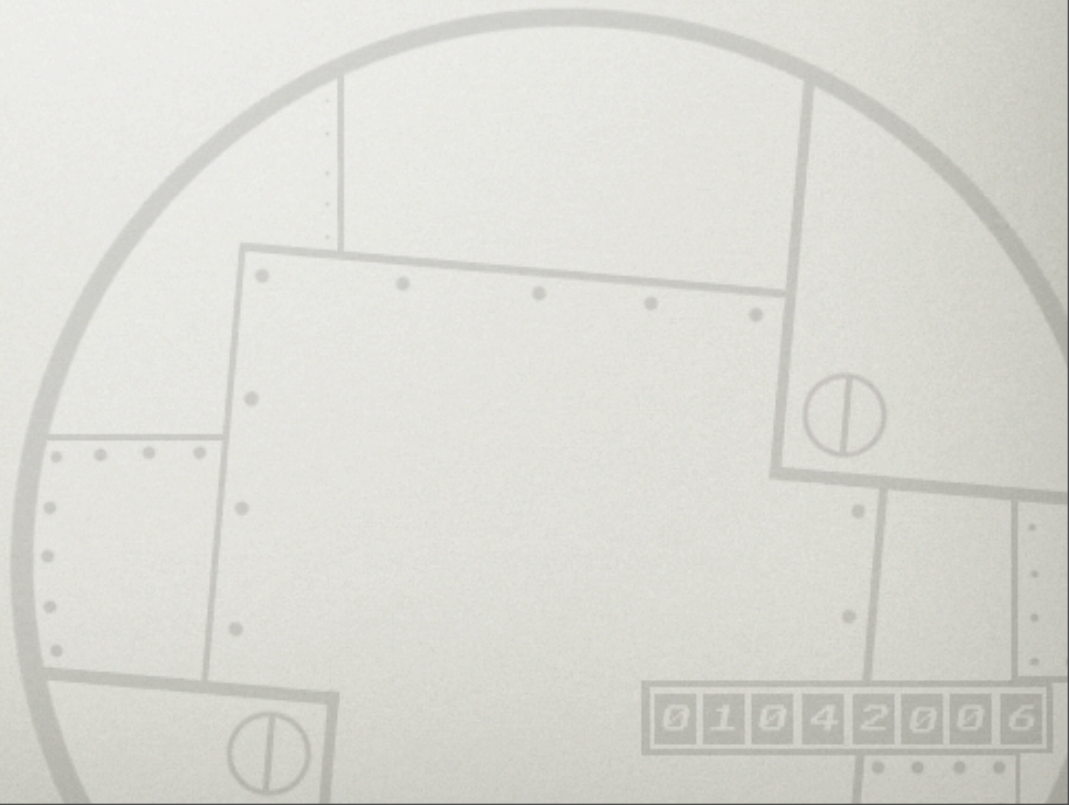
criation properties:
- date
- hardware
- ...

<XML>

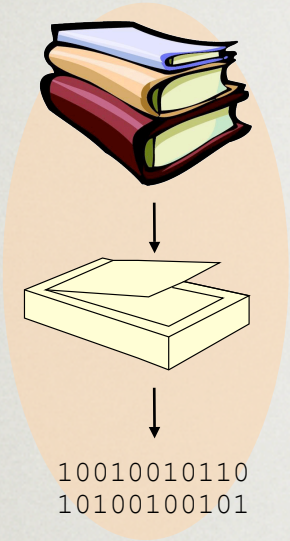
Descriptive Metadata:
- producer
- colection
- ...

Compressed File

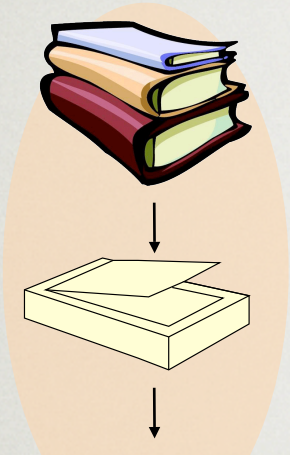
SIP Structure (+complex)

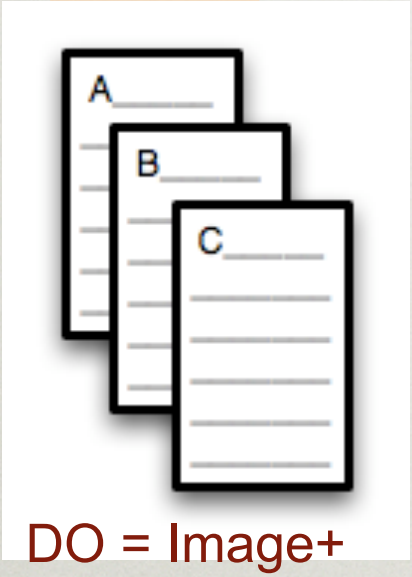


SIP Structure (+complex)



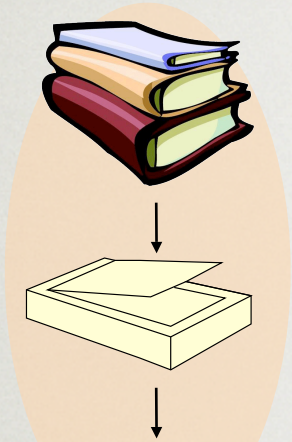
SIP Structure (+complex)

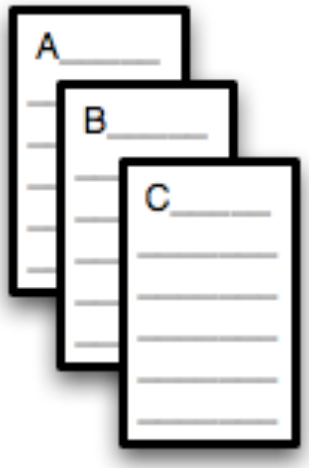




DO = Image+

SIP Structure (+complex)

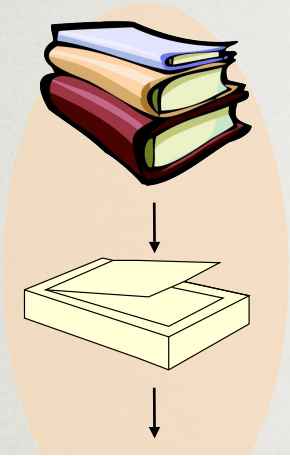


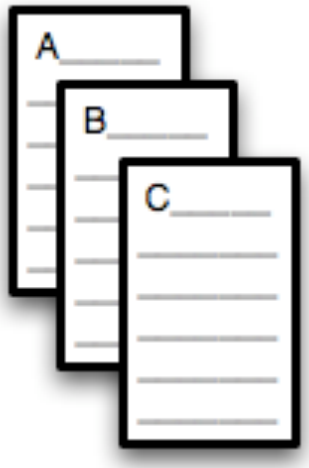


DO = Image+

Properties

SIP Structure (+complex)



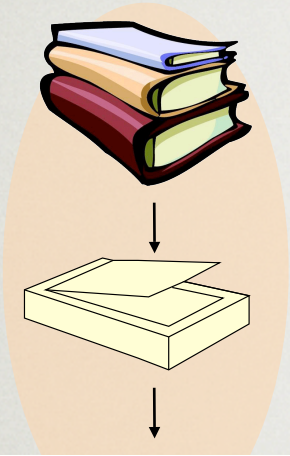


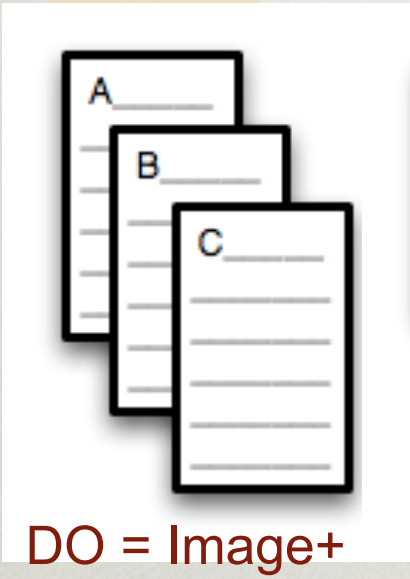
DO = Image+

Properties

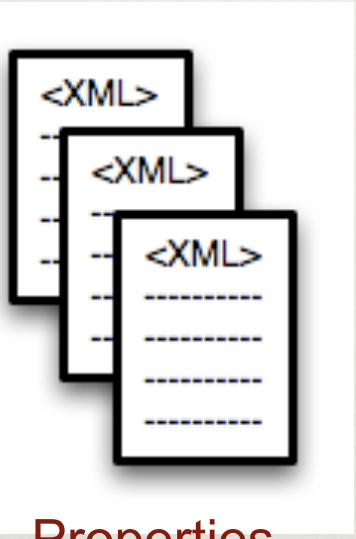
Technical Metadata

SIP Structure (+complex)

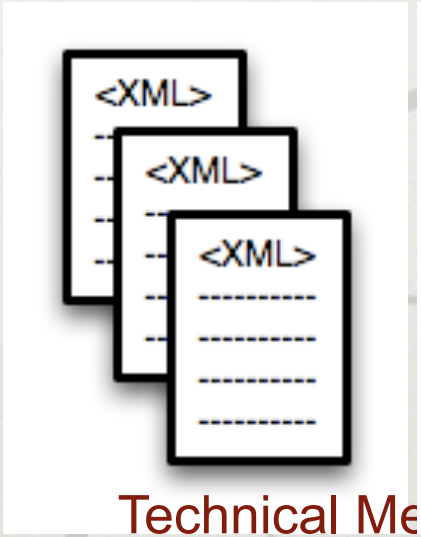




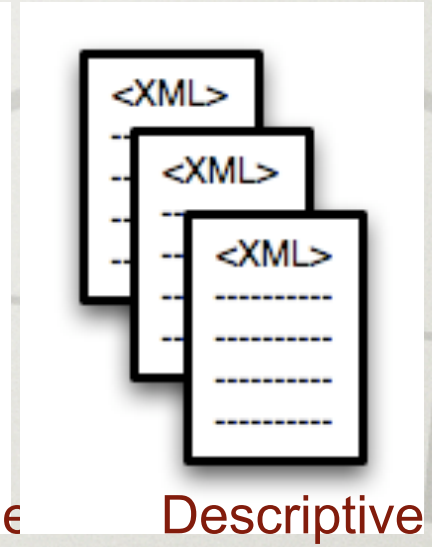
DO = Image+



Properties

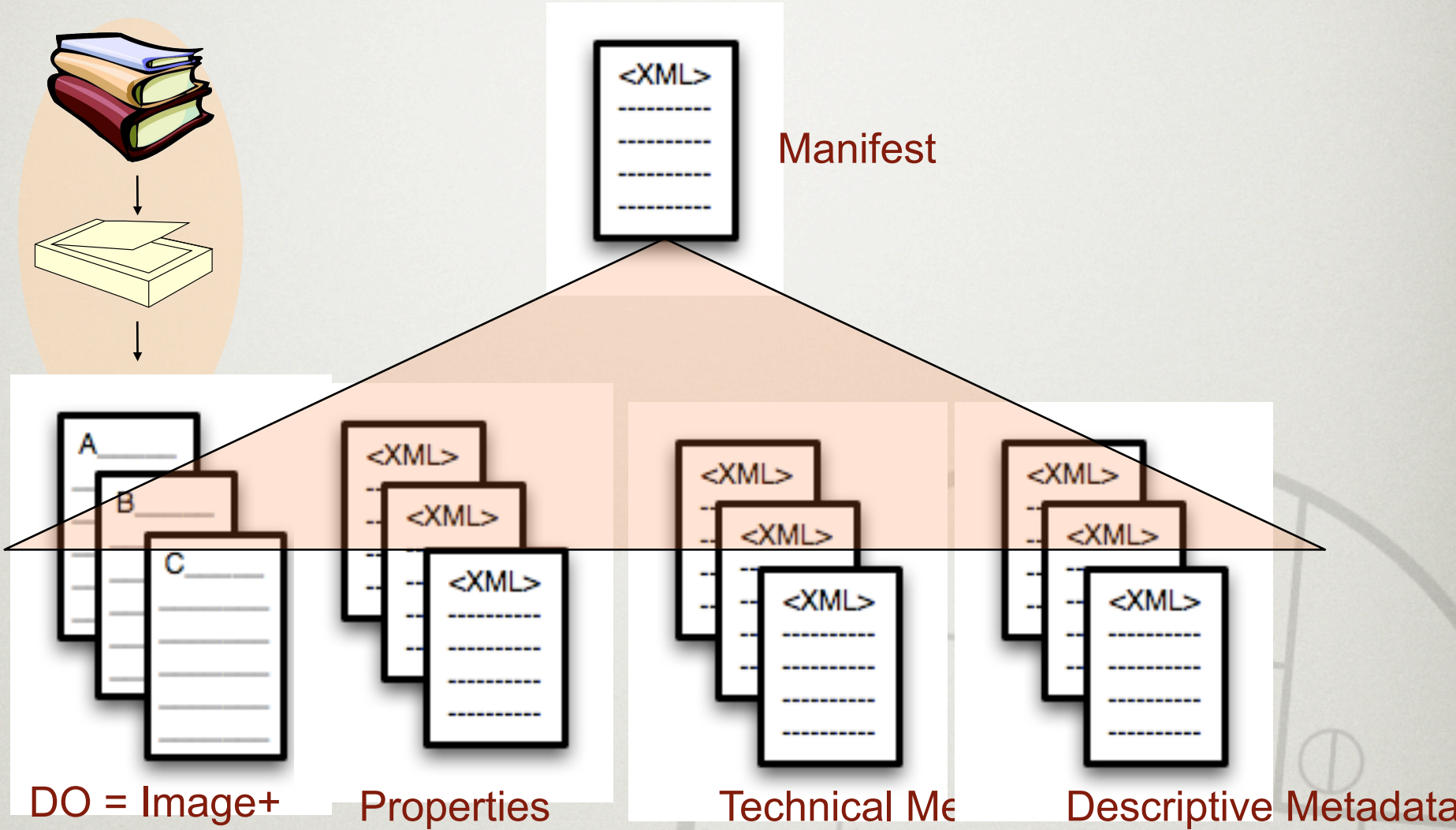


Technical Me

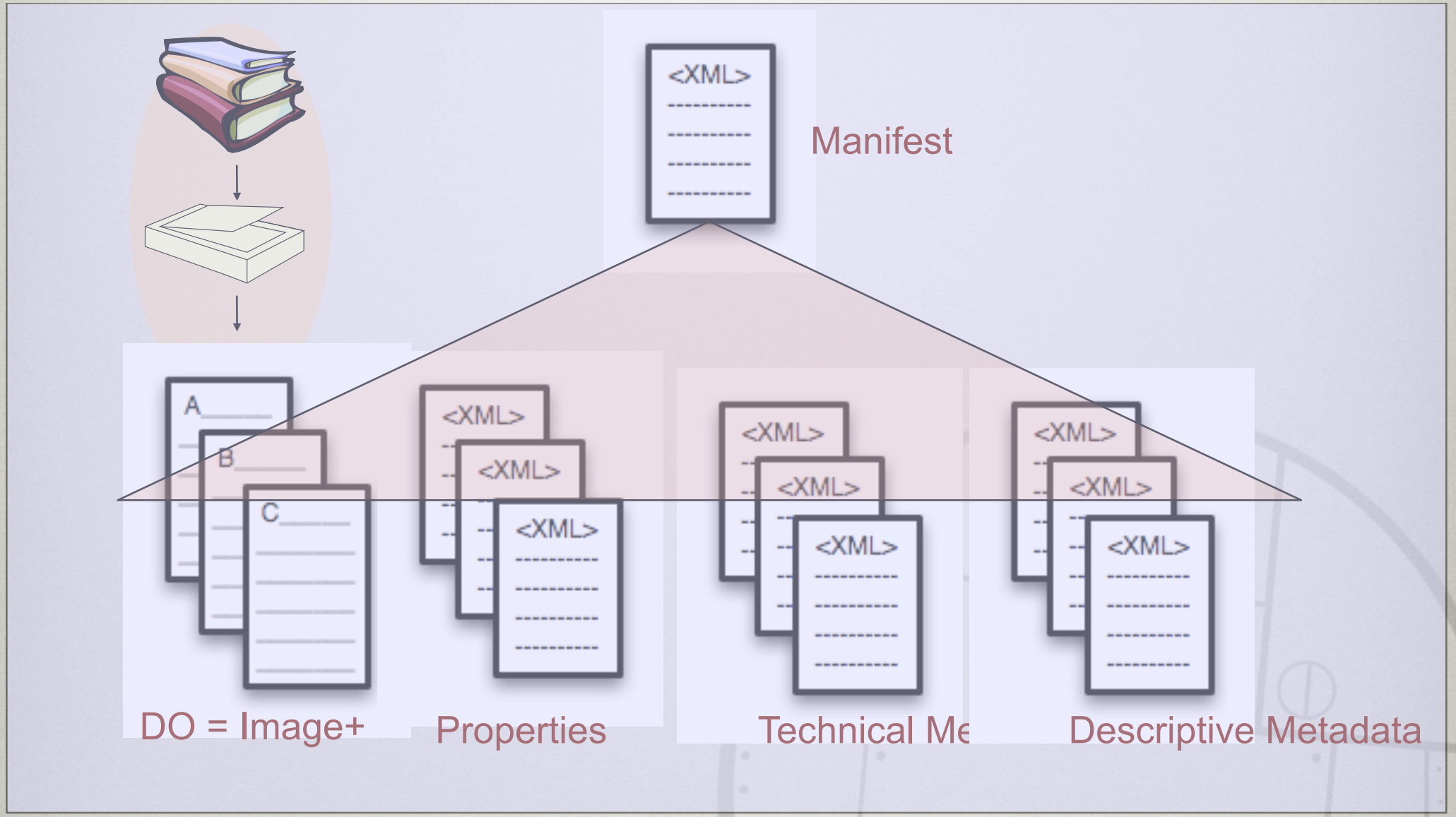


Descriptive Metadata

SIP Structure (+complex)



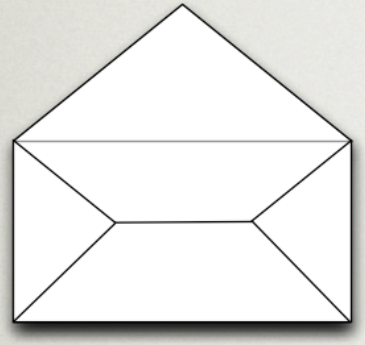
SIP Structure (+complex)



Compressed File

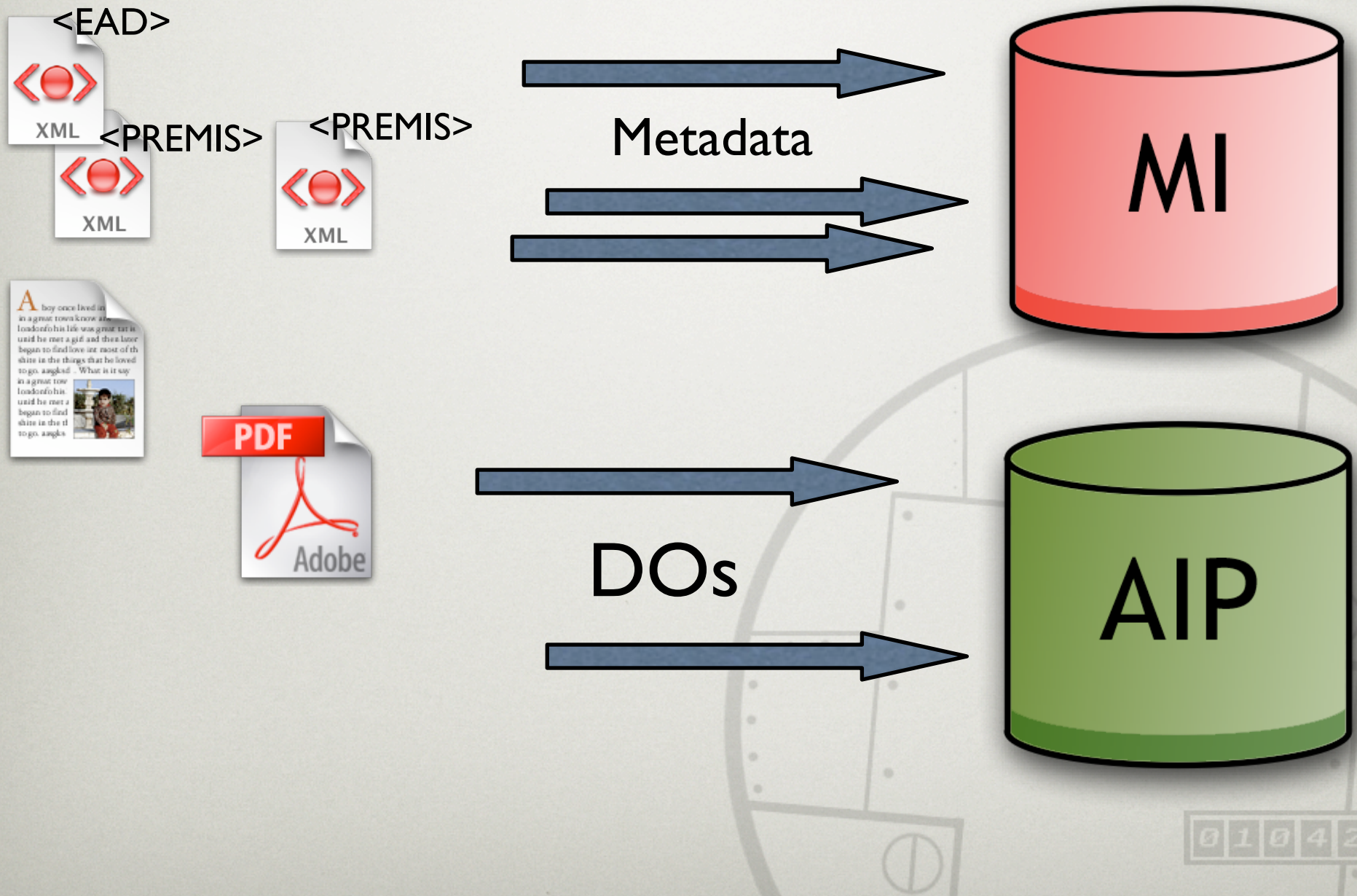
Ingestion Workflow

SIP → AIP

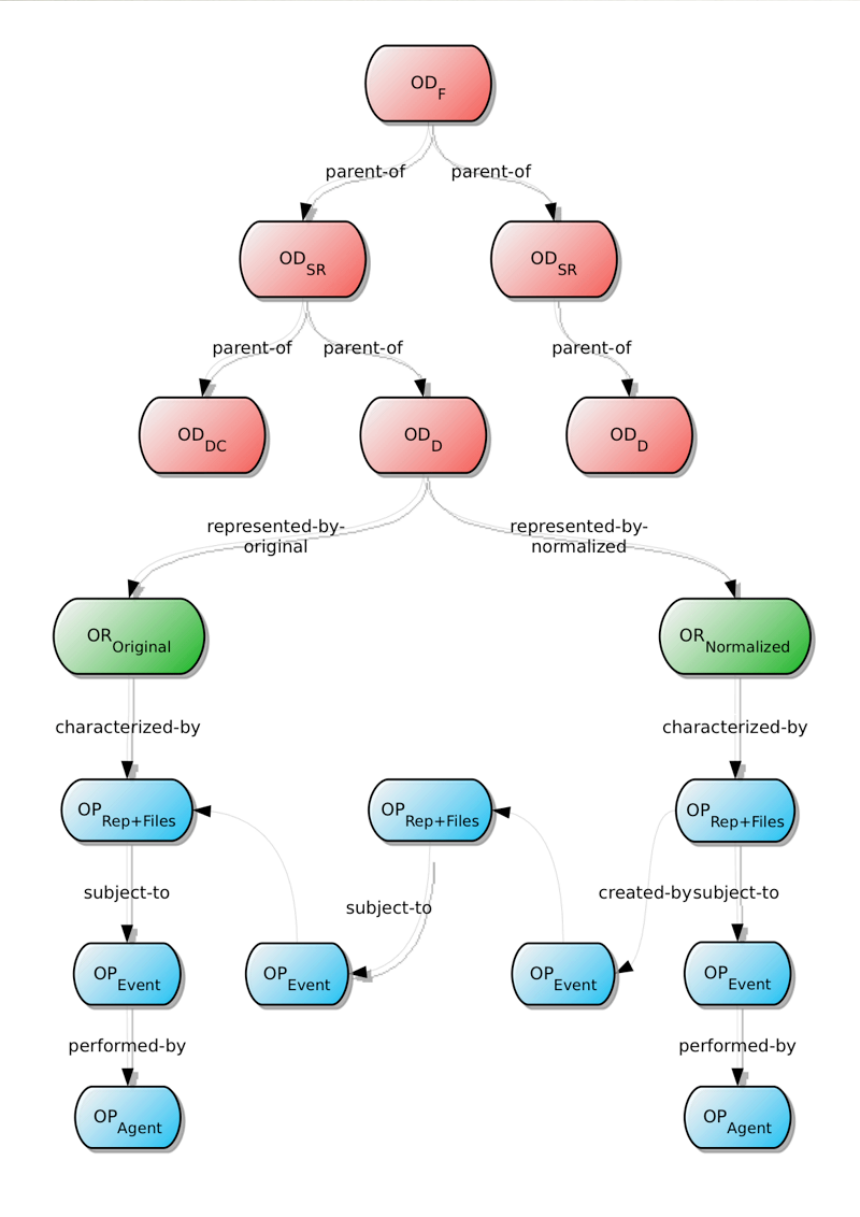


- integrity check
- virus check
- generation of preservation metada (PREMIS)
- conversion to a normalized format
- generation of technical metadata
- generation of preservation metadata (PREMIS)

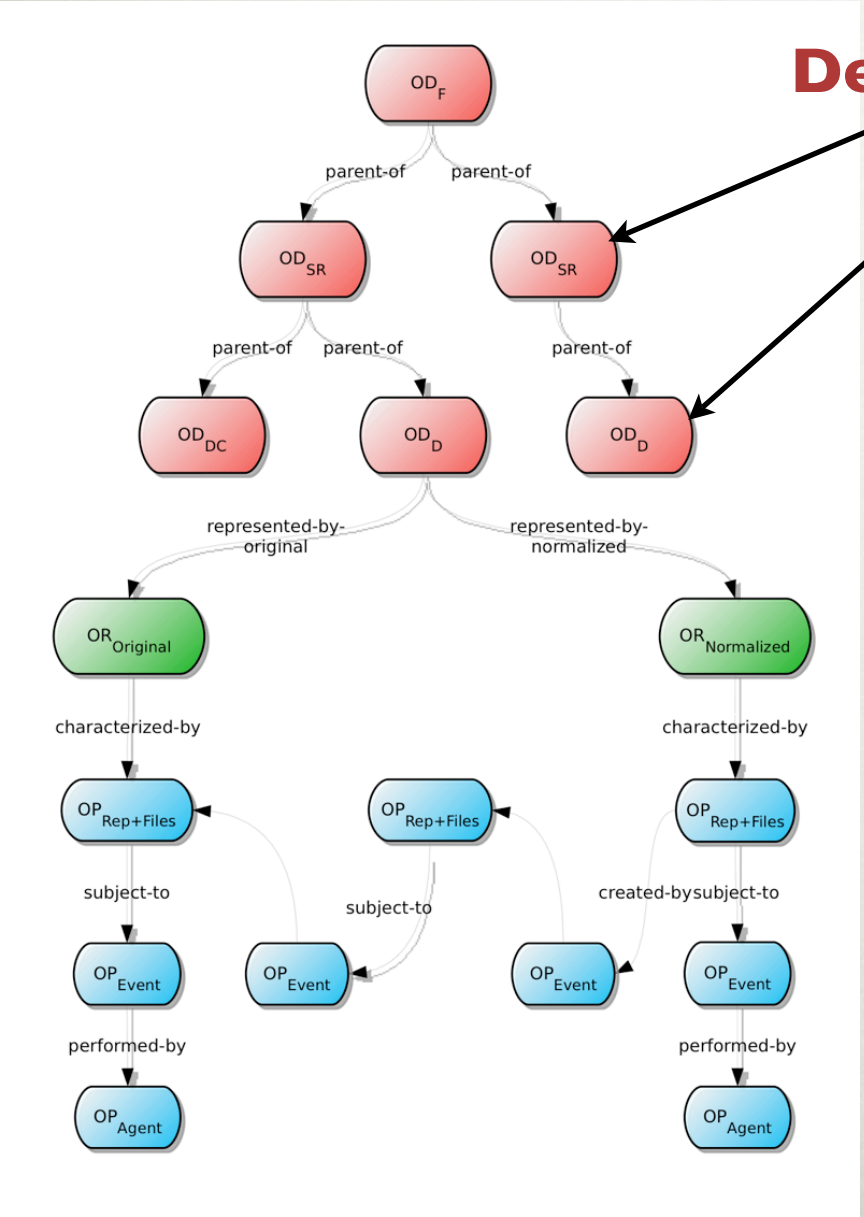
AIP Storage



Roda Data Model

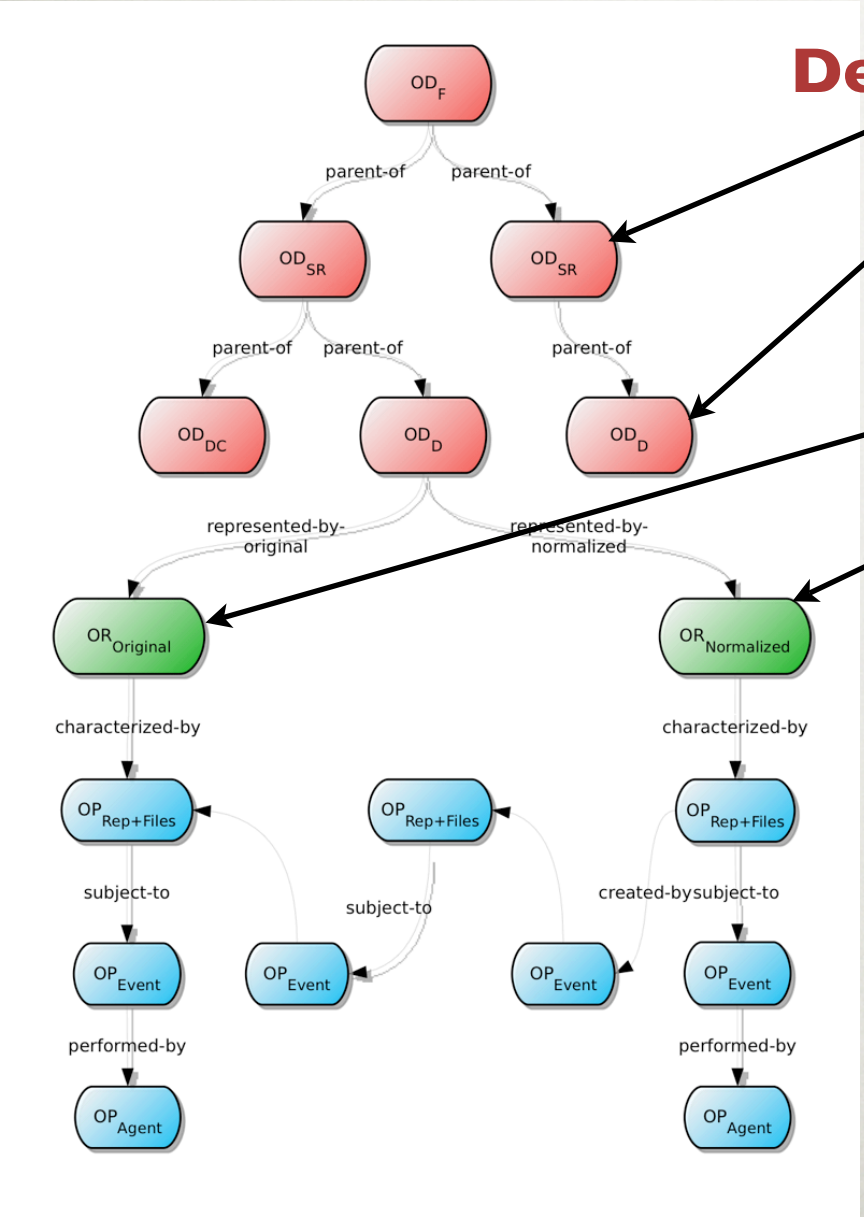


Roda Data Model



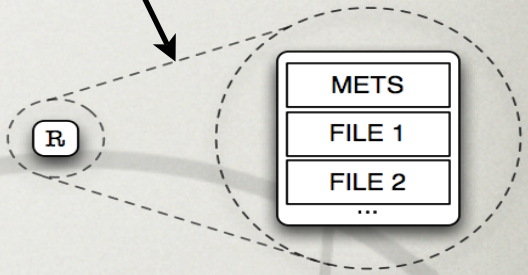
Description Objects

Roda Data Model

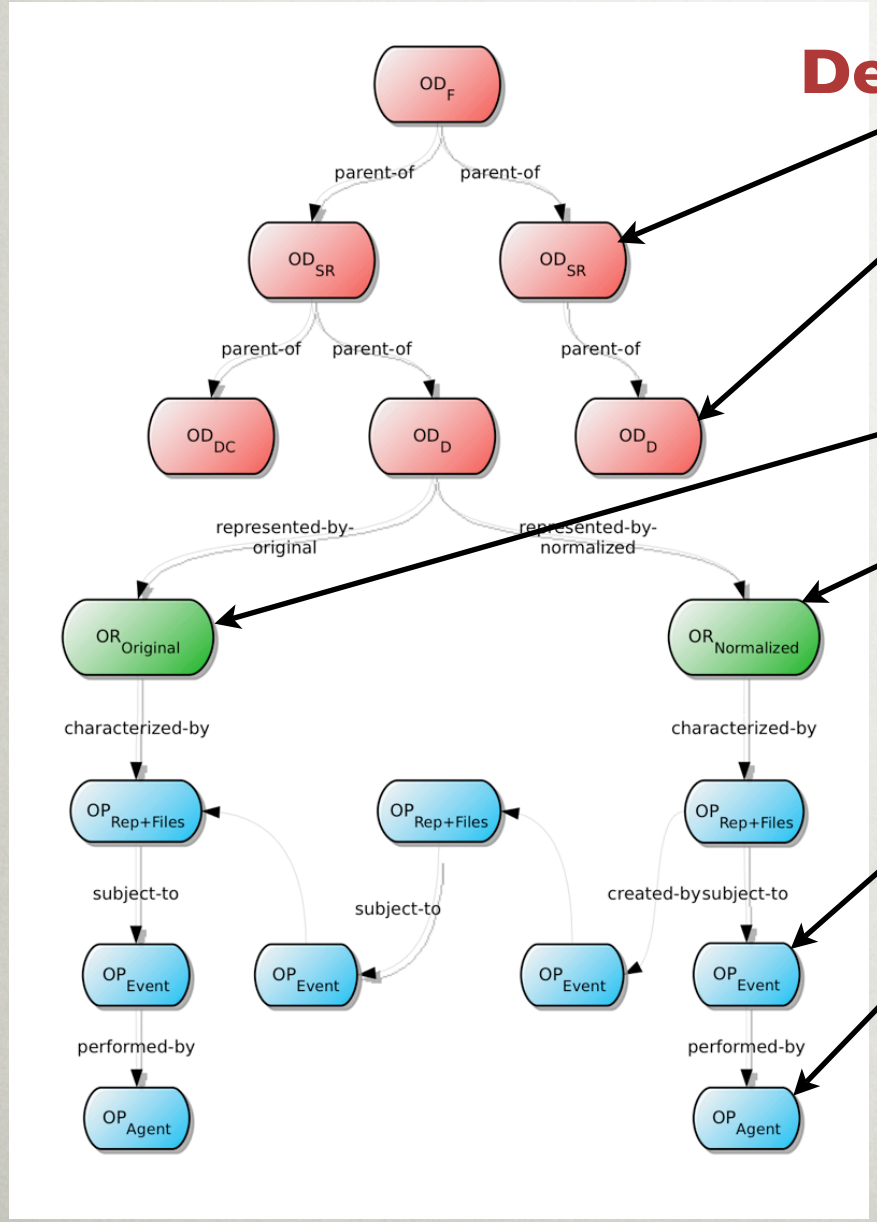


Description Objects

Representation Objects

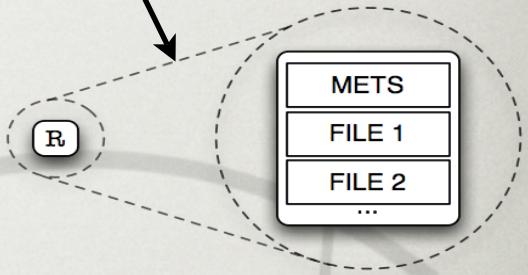


Roda Data Model

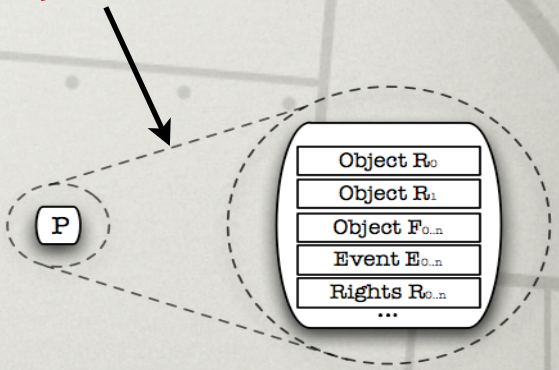


Description Objects

Representation Objects

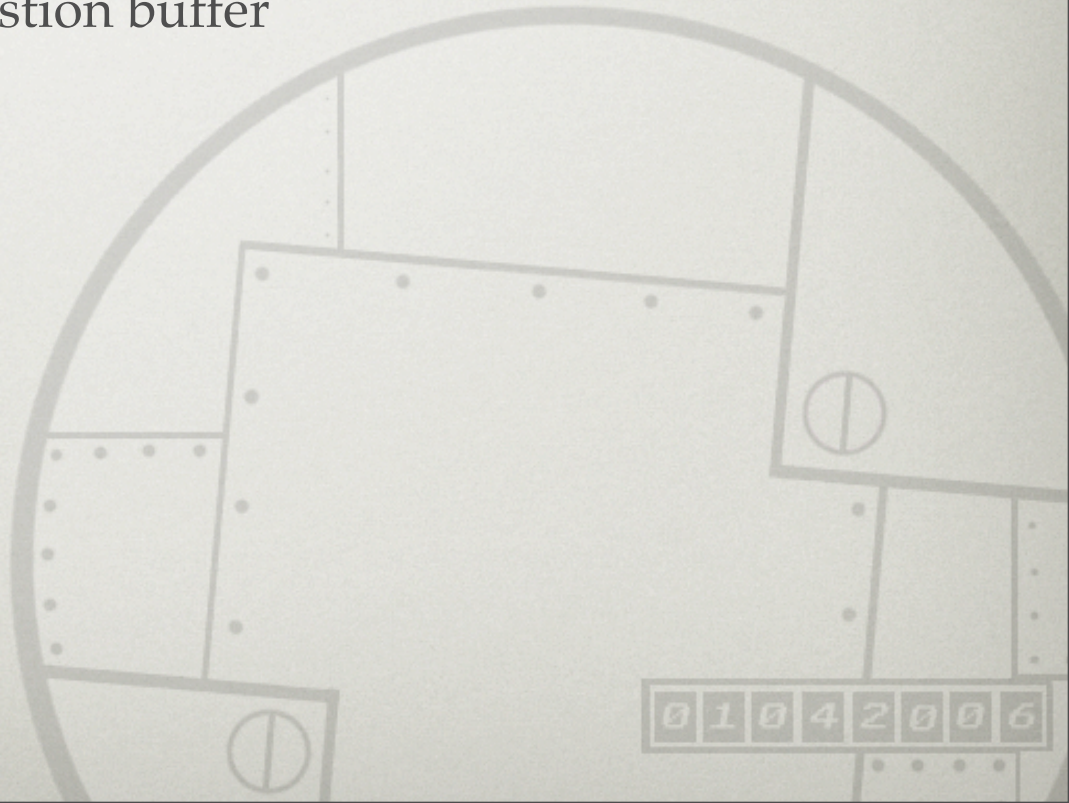


Preservation Objects



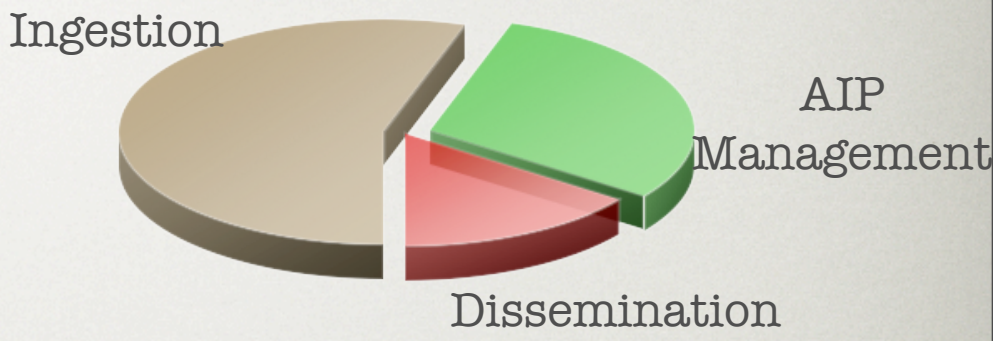
Implementation Requisites

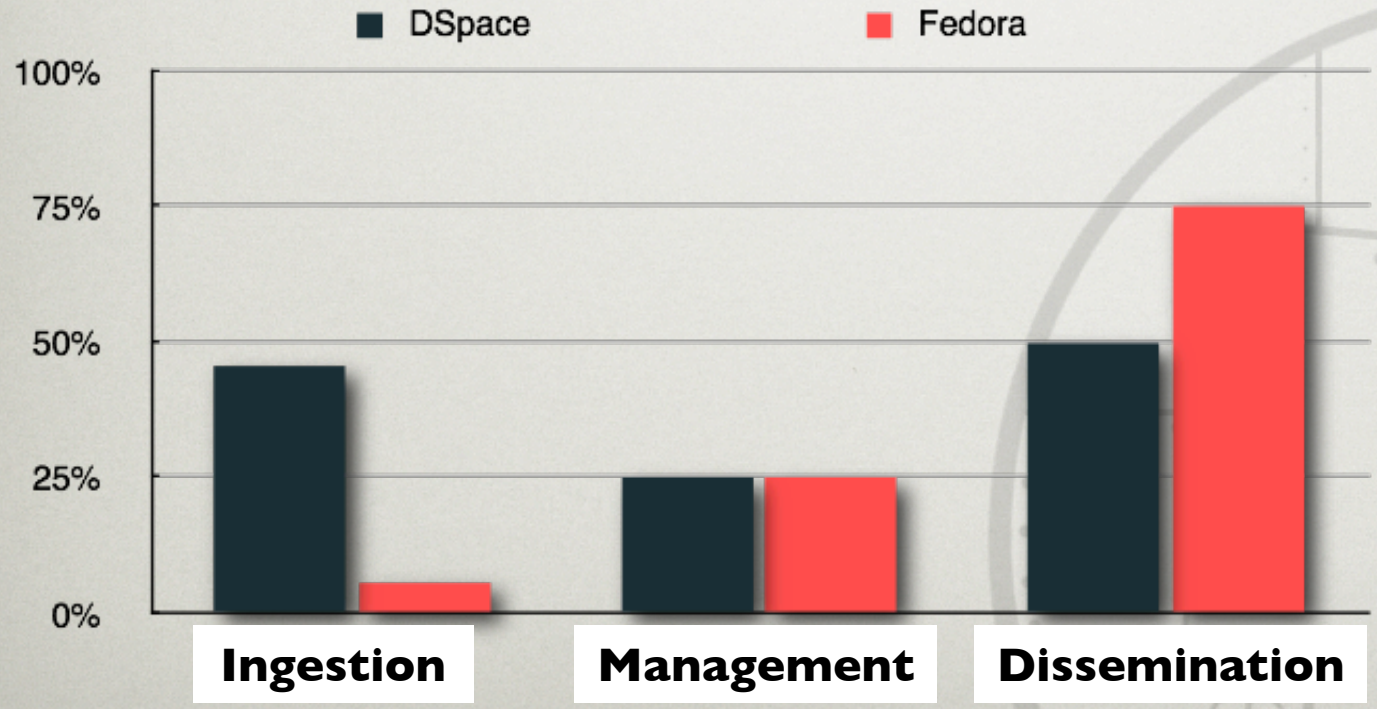
- Graphical Interface for Ingestion process
- Producer registry
- SIP production tool
- Ingestion feedback
- Partial Ingestion
- “Quarantine” zone: cache, ingestion buffer
- SIP validation
- Error reporting
- Persistent identifiers
- PREMIS event generation
- DIP digital signature
- ...



Development framework

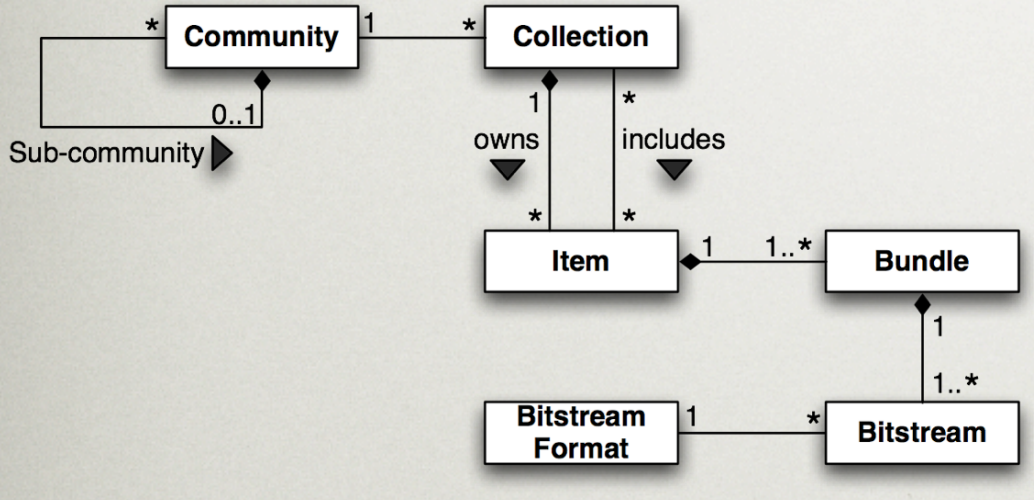
Requisites based comparaison

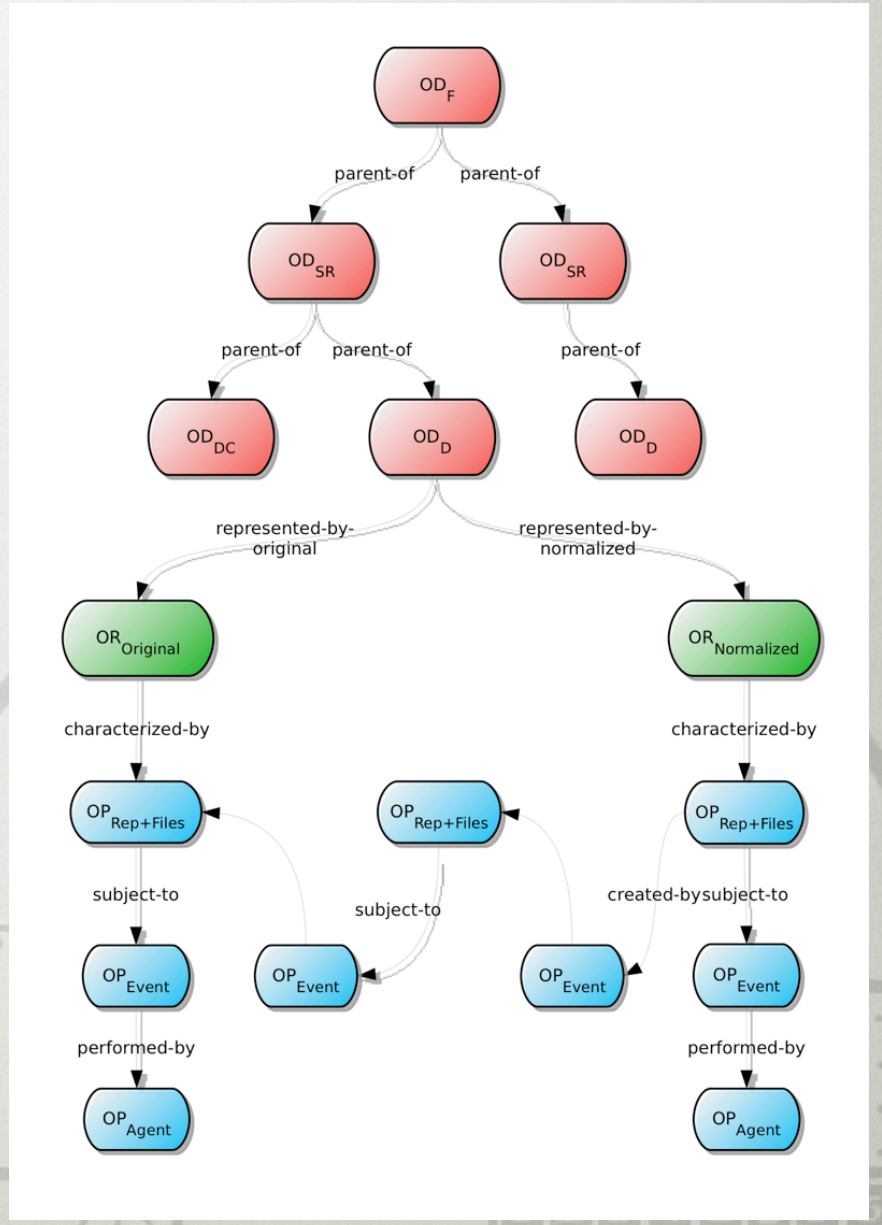




Matching data models

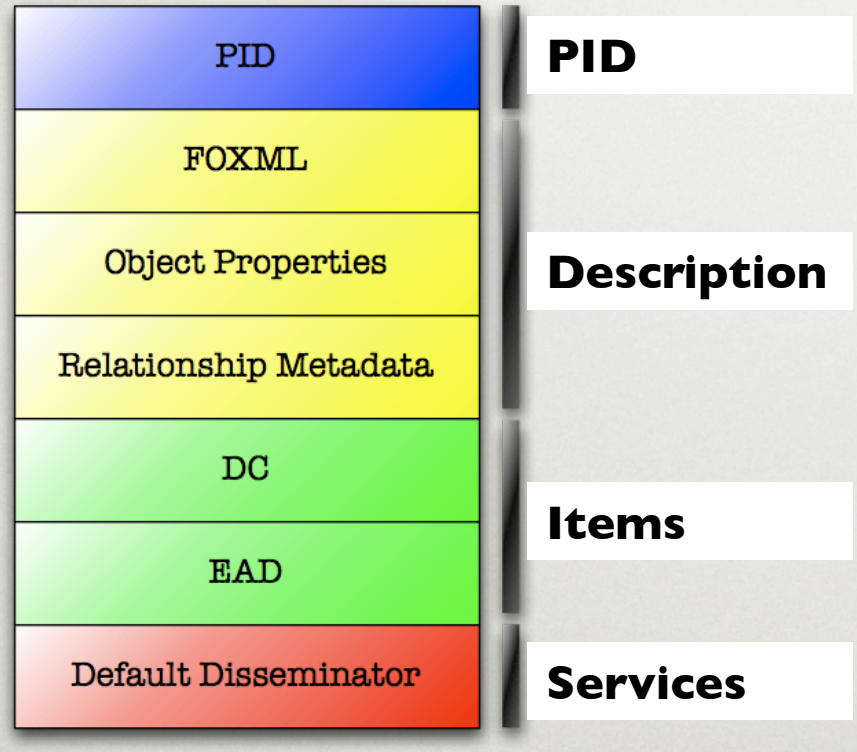
DSpace

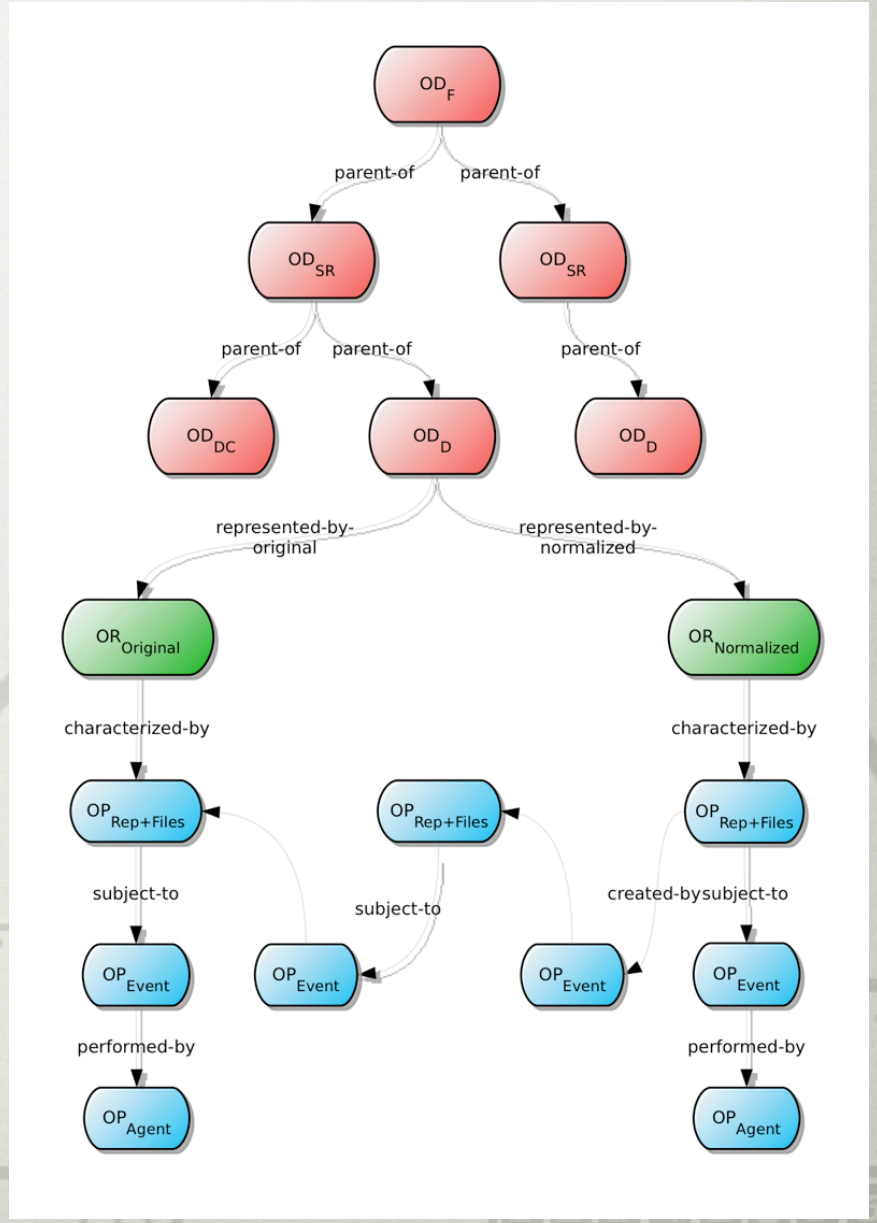




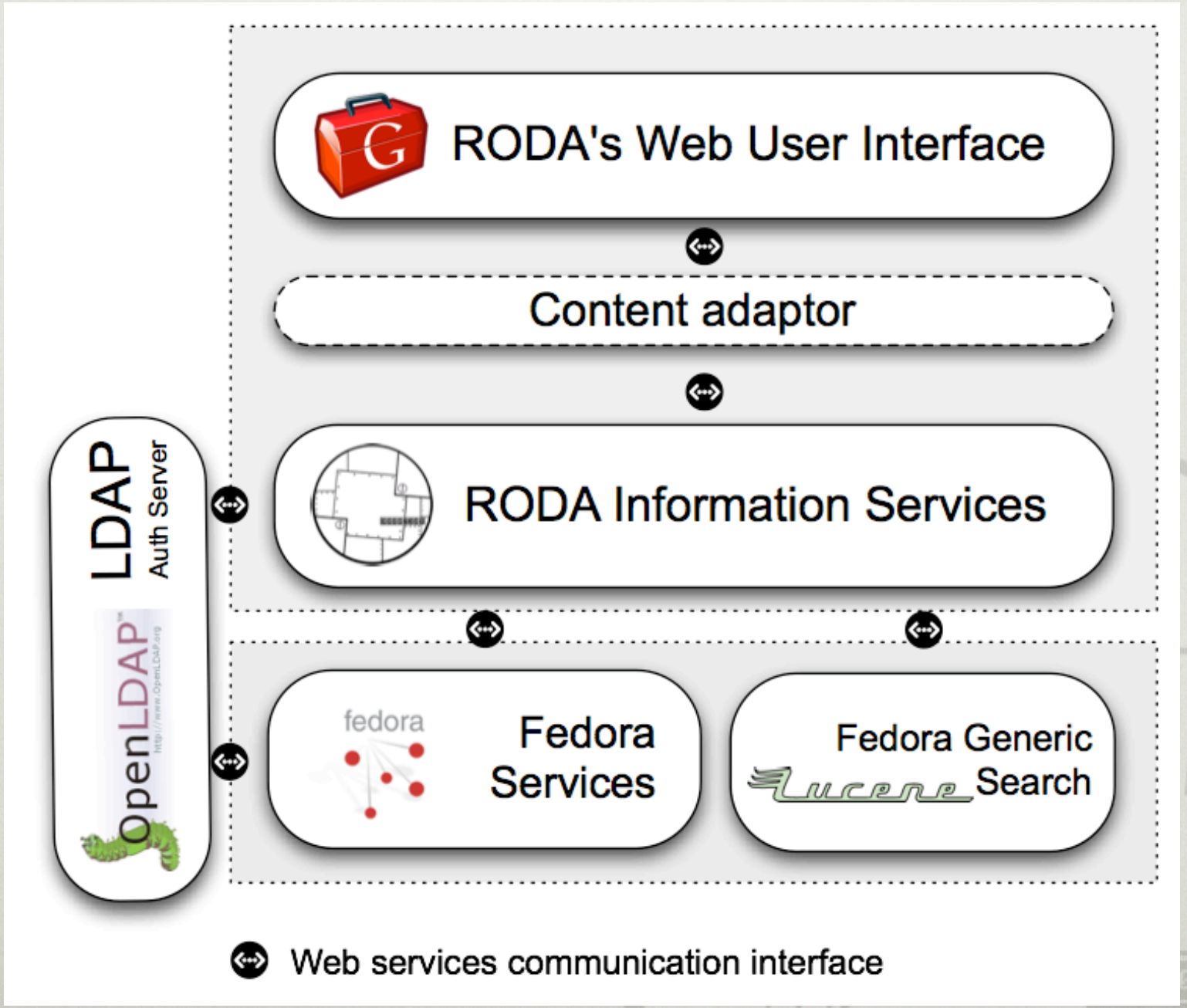
Matching data models

Fedora

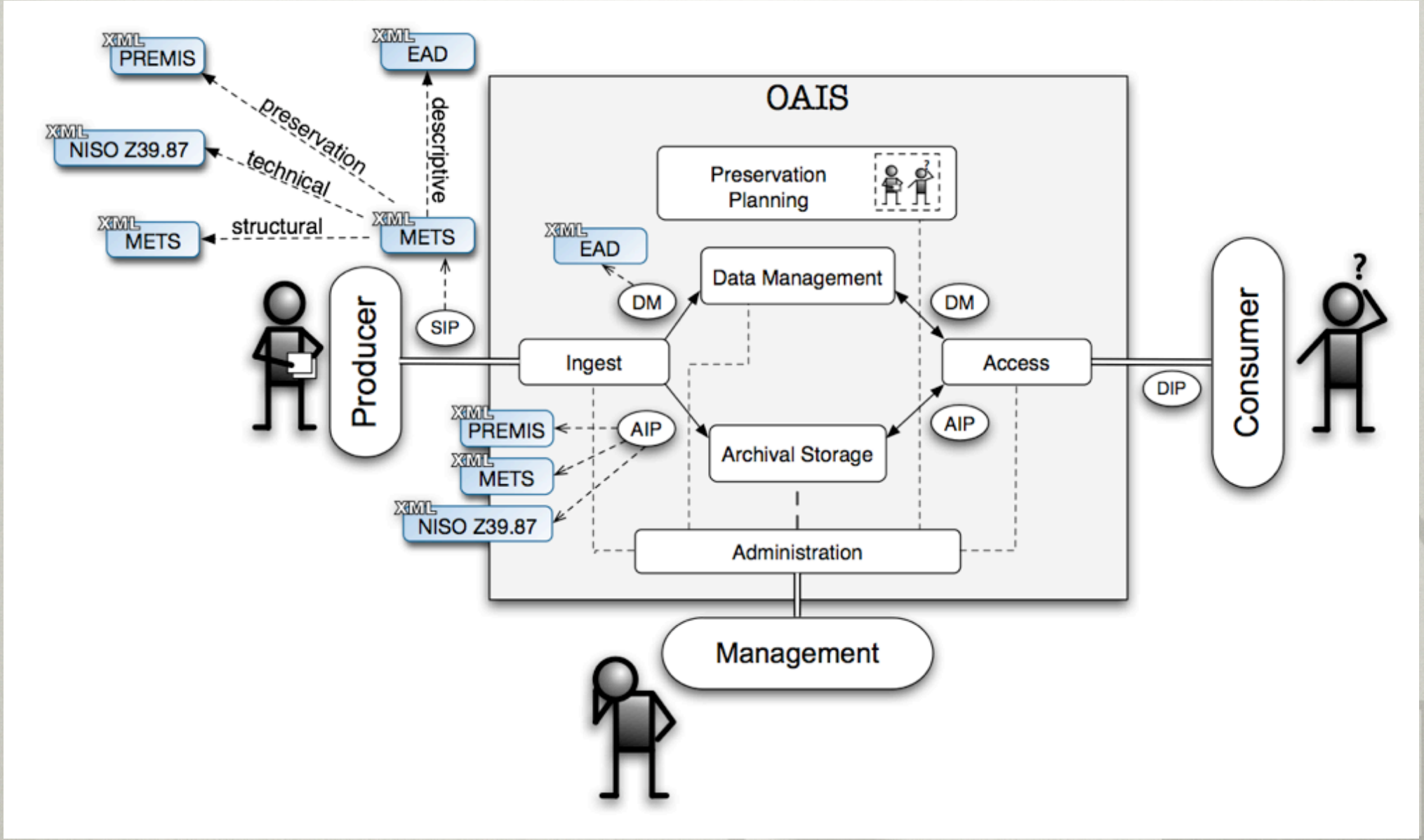




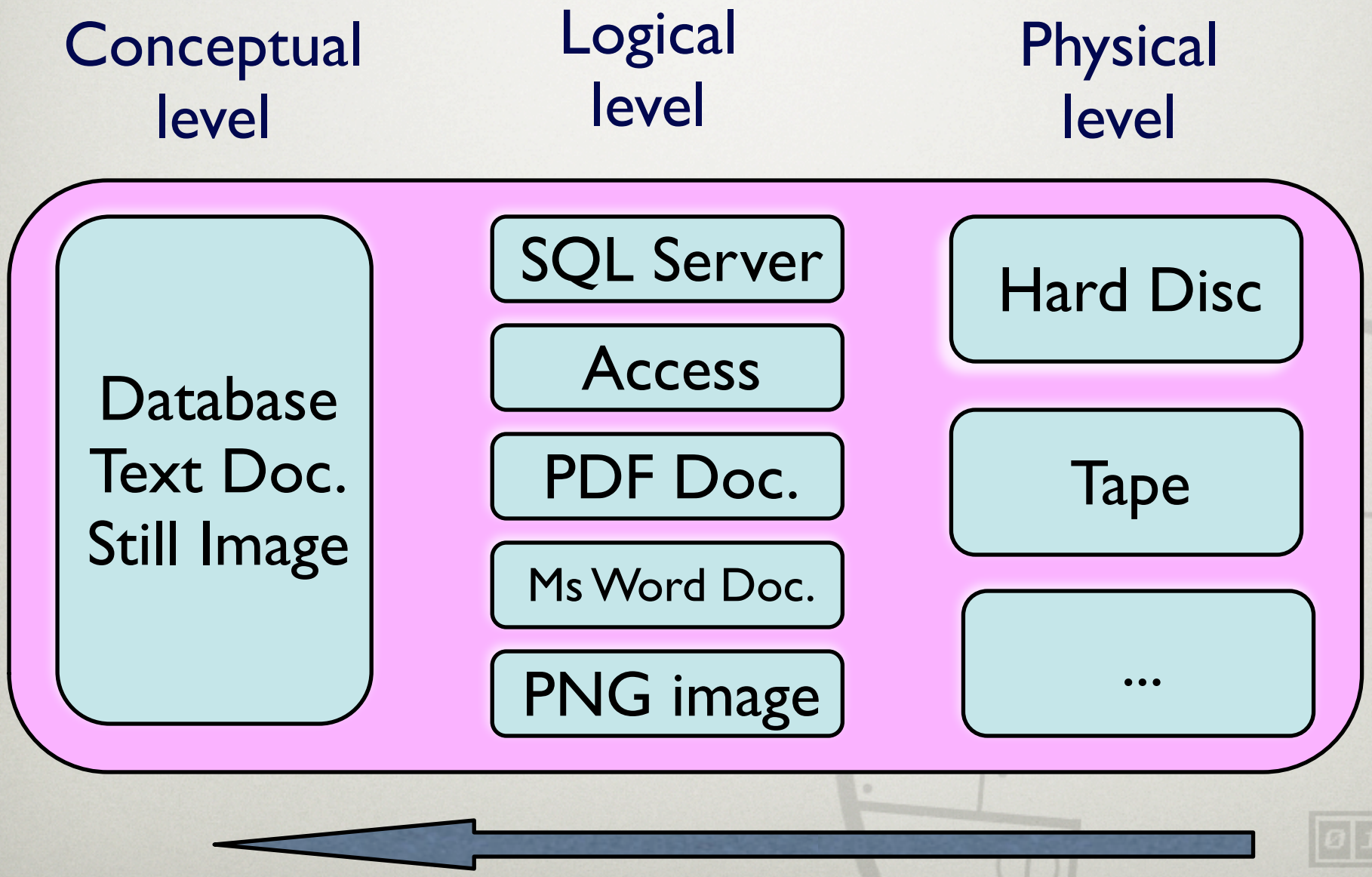
Architecture



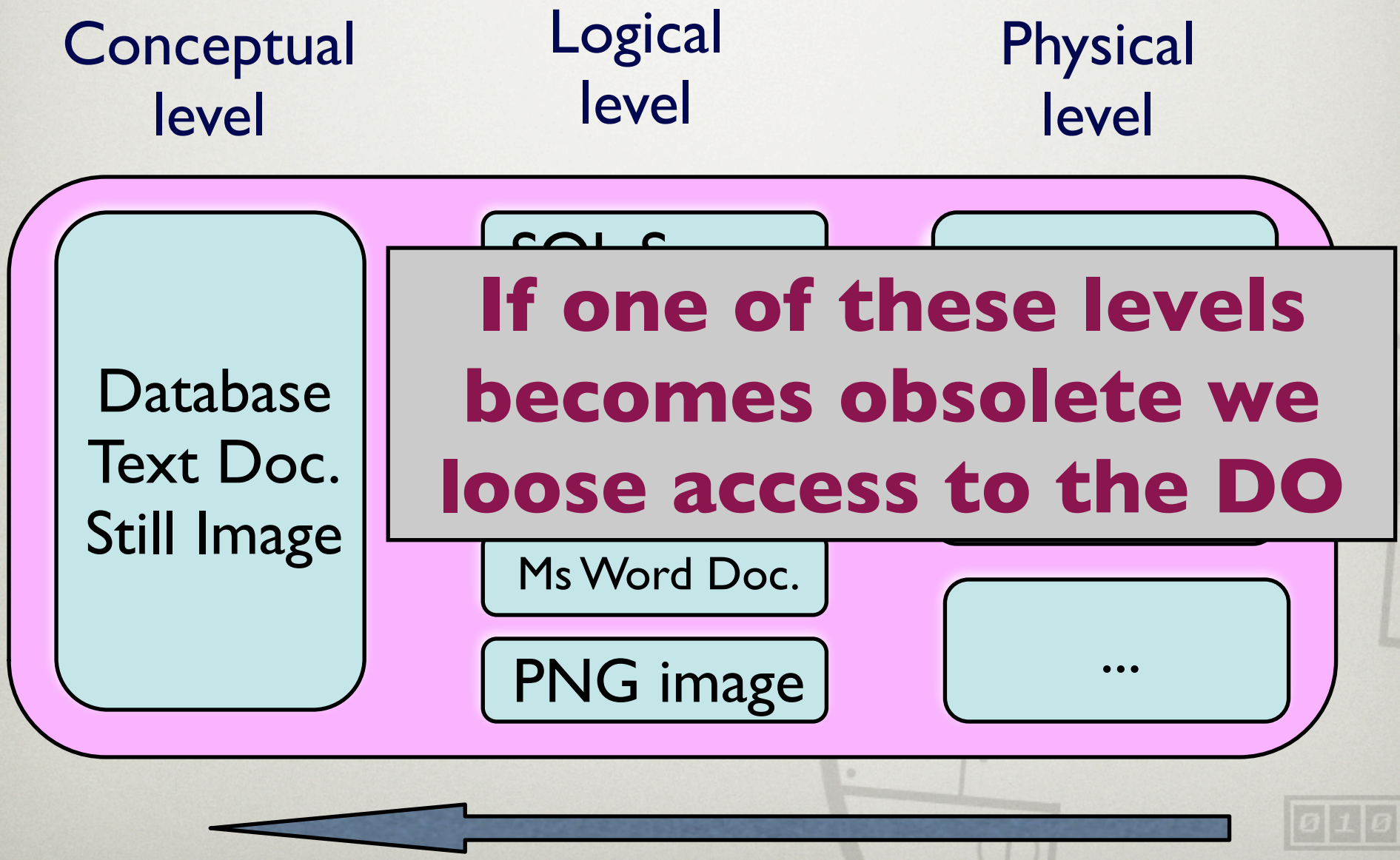
RODA Schemas



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
...

DO Anatomy


Conceptual level
Logical level
Physical level
Database
Text Doc.
Still Image
Ms Word Doc.
PNG image
...
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**

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 then preserving the original bit stream).

Advantag

- DO are
- No need to preserve the original hw/sw platform
- Most used strategy and the only that has worked so far

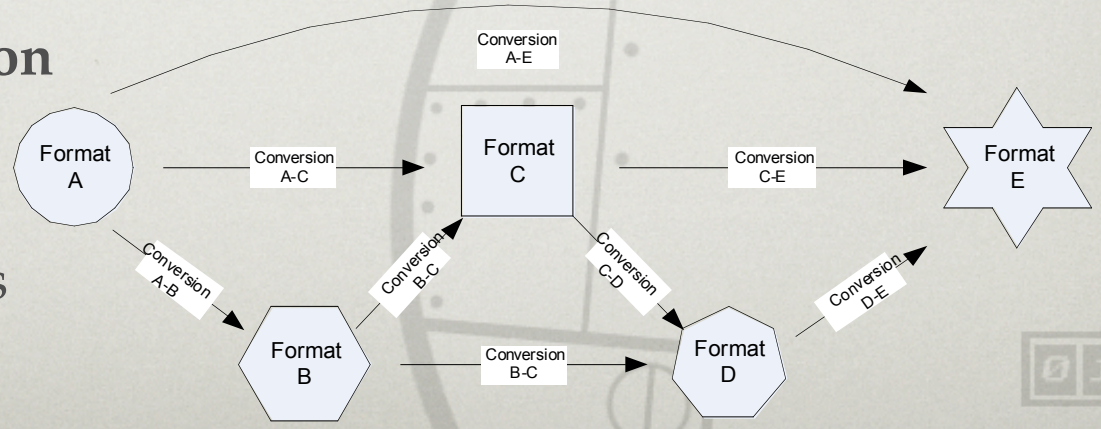
What are the significant properties?

Disadvantages

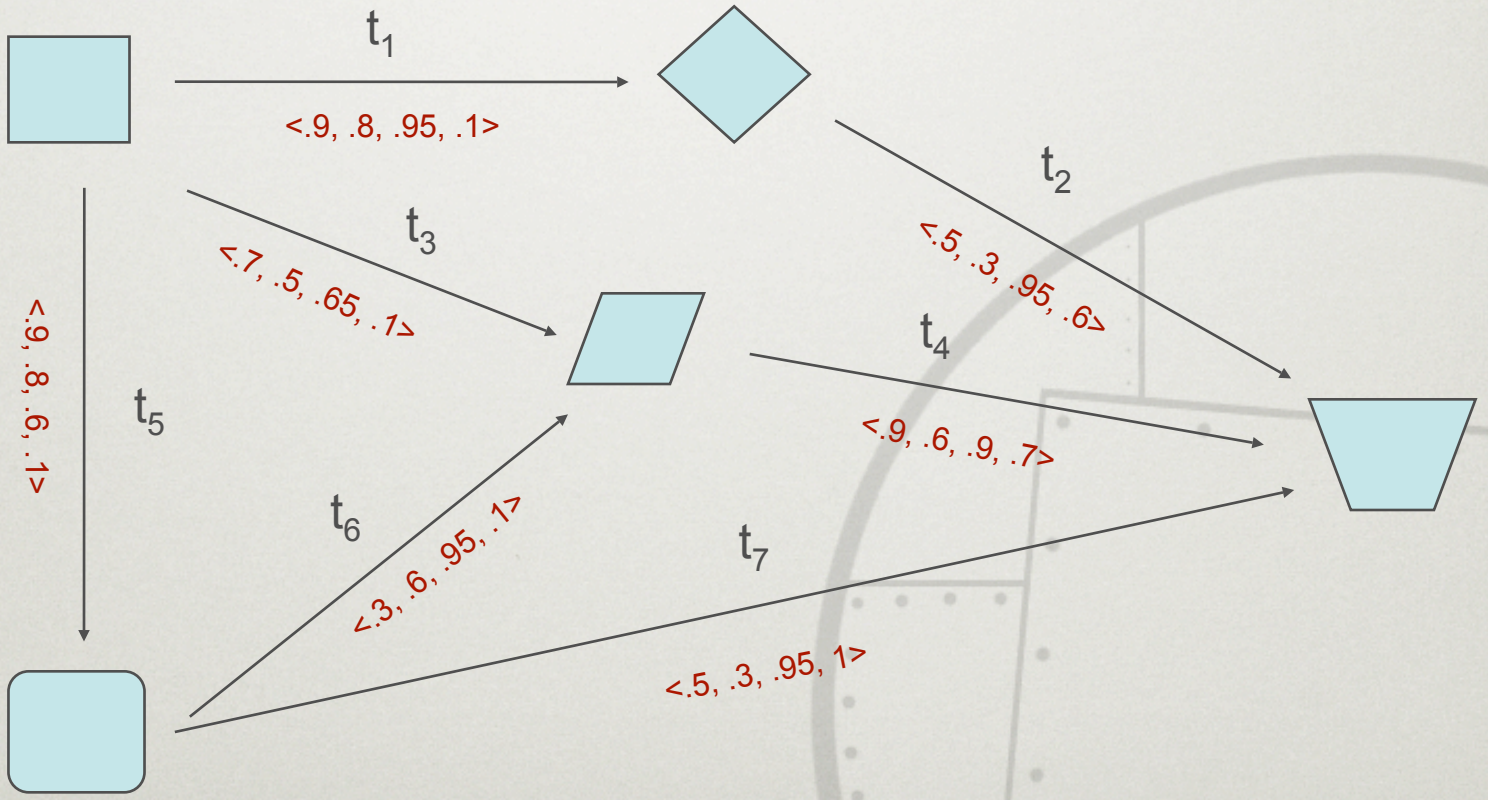
- Possible loss of information during conversion
- Continued maintenance is needed
- **In the longterm perspective costs are high**

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
    - Normalizatio, migration on request
  - Generalized **cost reduction**
- Disavantages
  - **Bandwidth** requirements
  - **Slow**

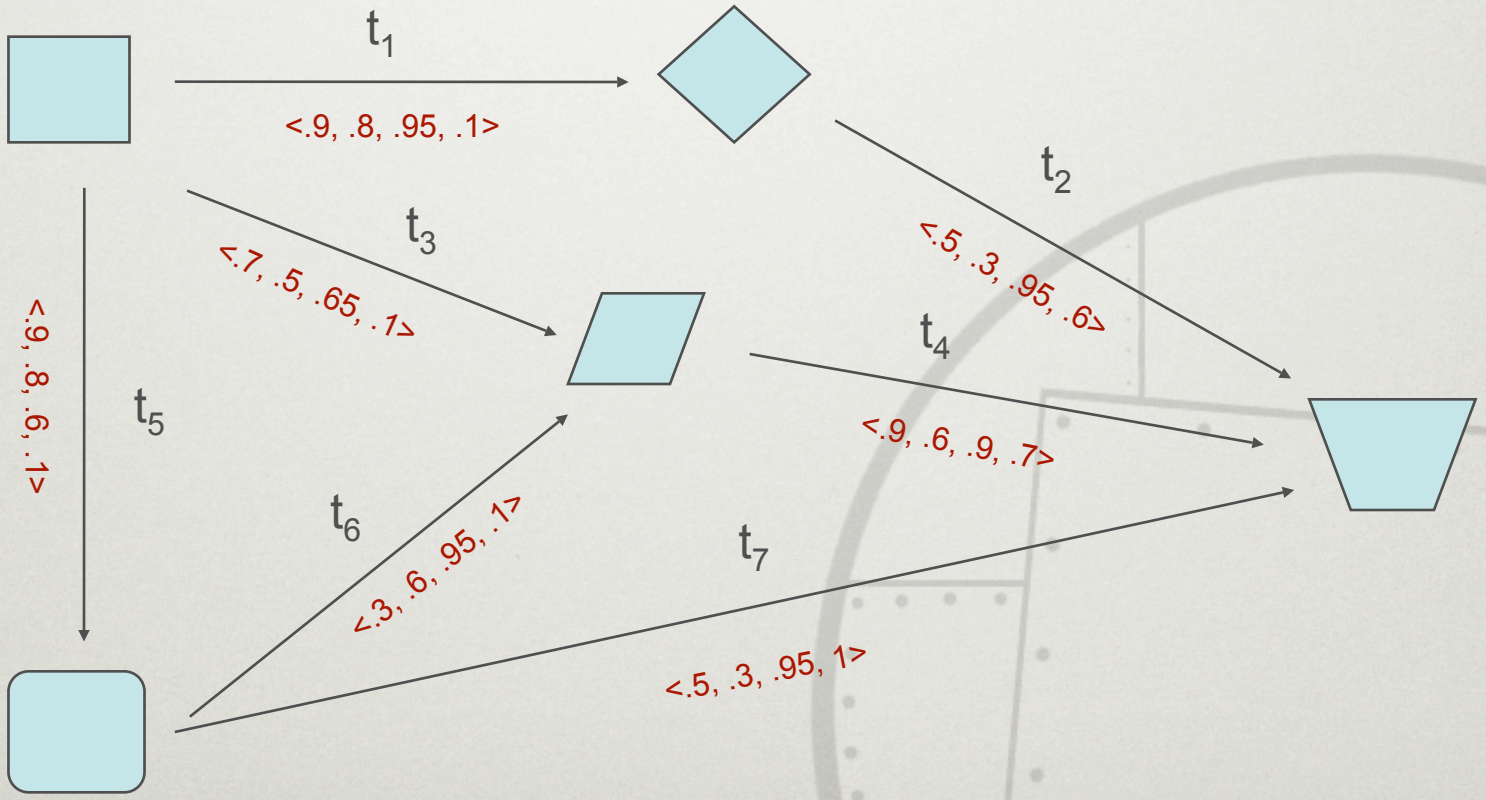


Distributed Preservation Services

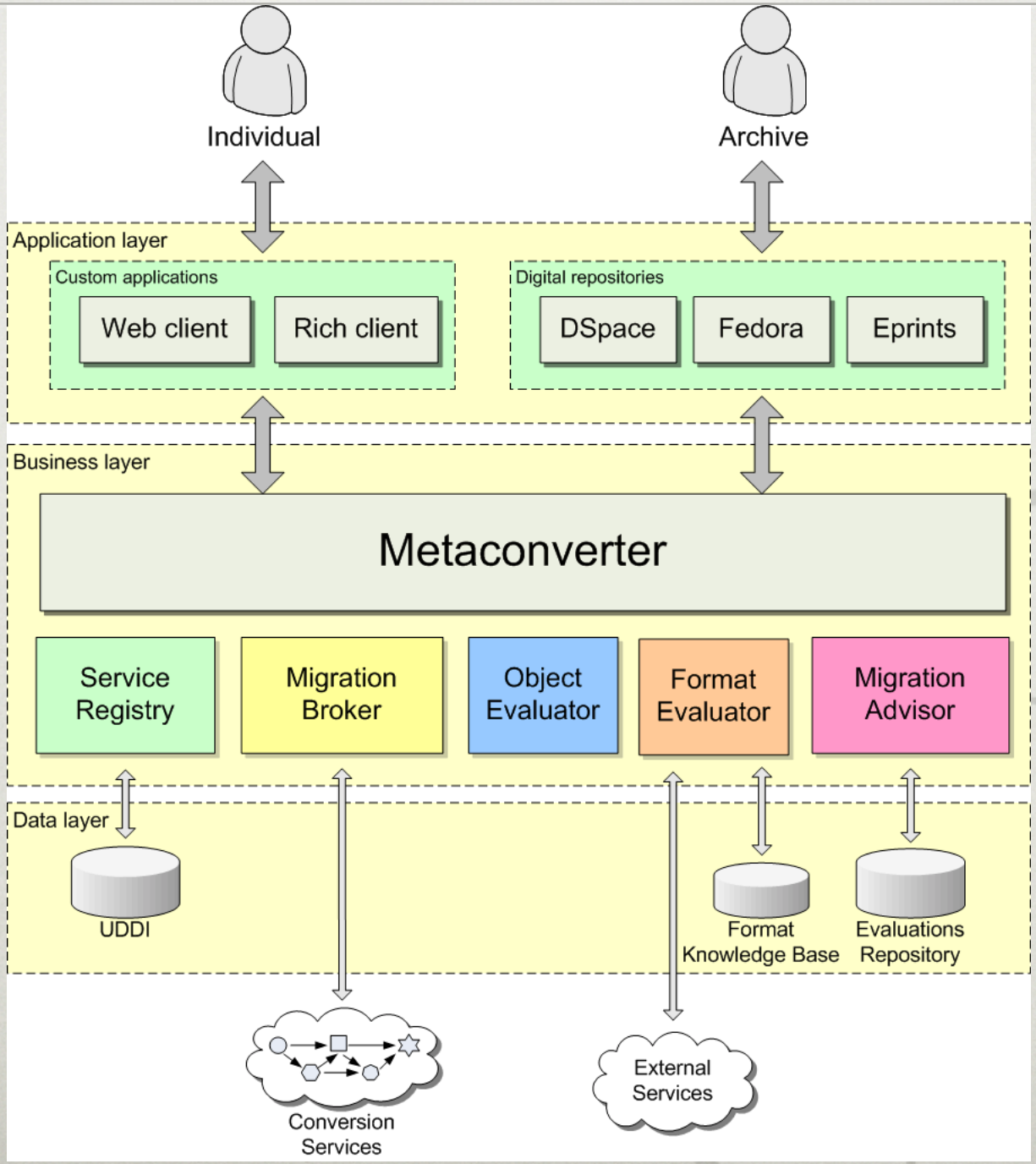


Distributed Preservation Services

CRiB project: http://crib.dsi.uminho.pt



CRiB: architecture



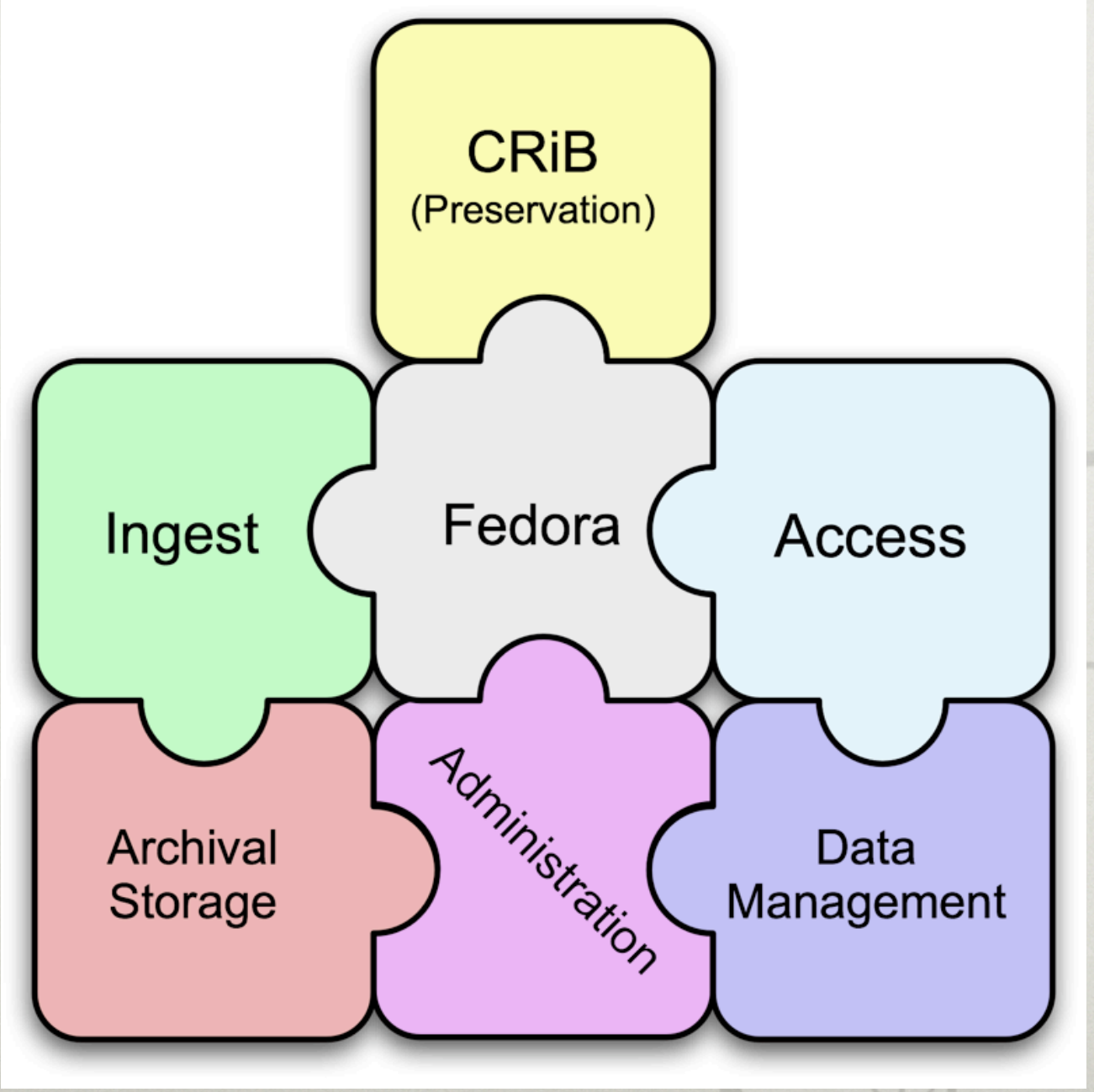
Migration Broker

- 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

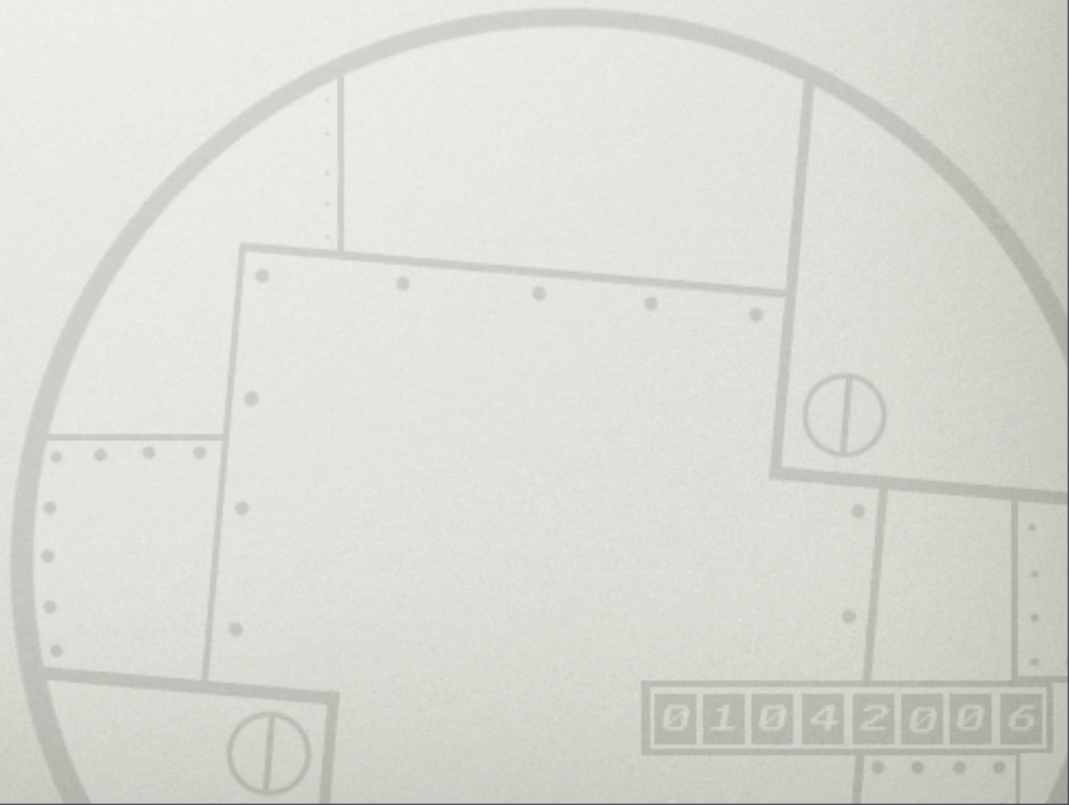
Object Evaluator

- 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

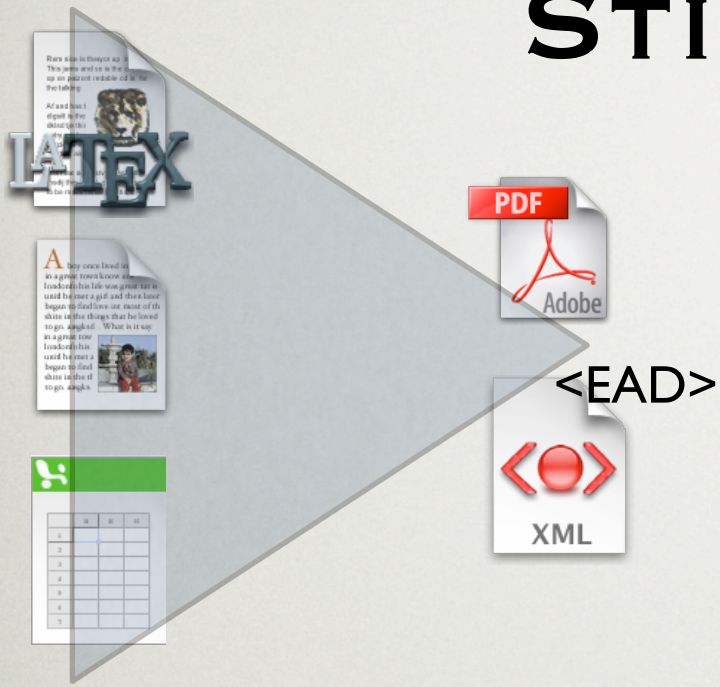
Text Documents and Still Images



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



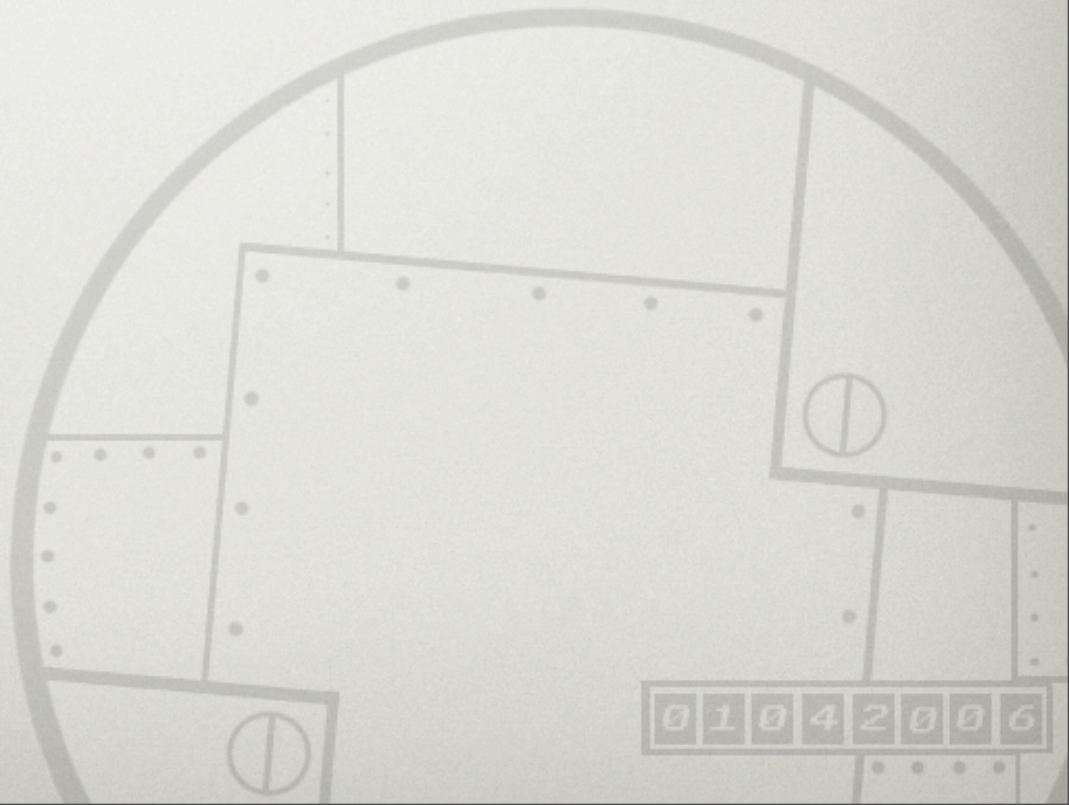
- EAD elements capture most of the significant properties: provenance, producer history, context, ...
- Content is kept in a normalized format: PDF and uncompressed TIFF.

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?
- ...

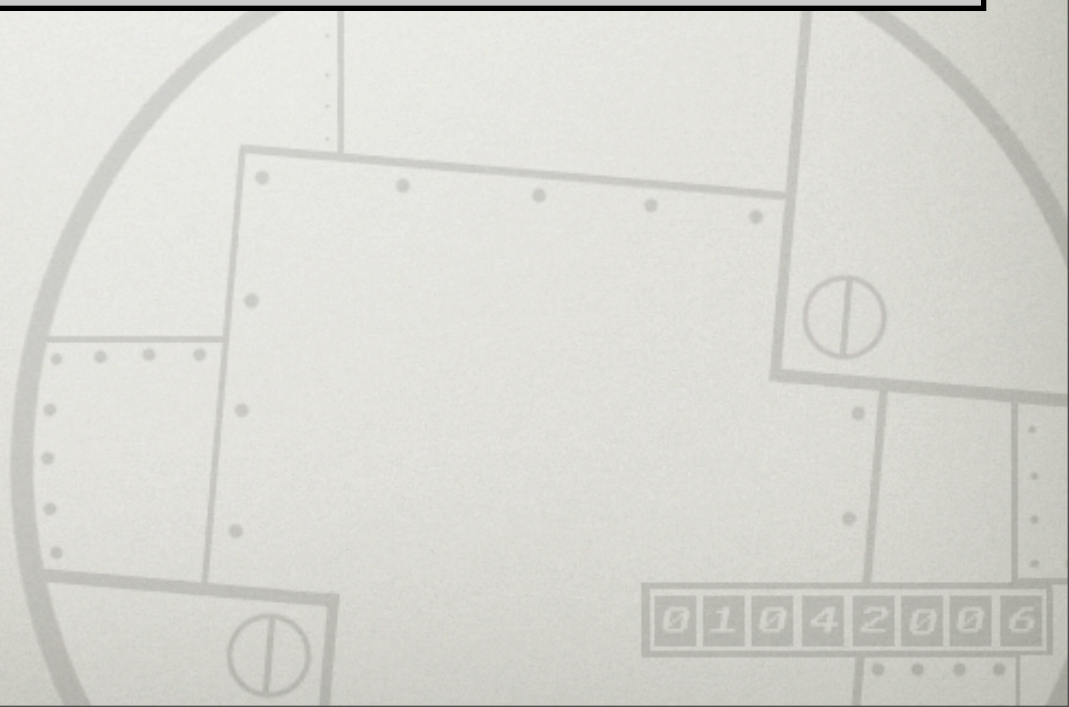


Databases

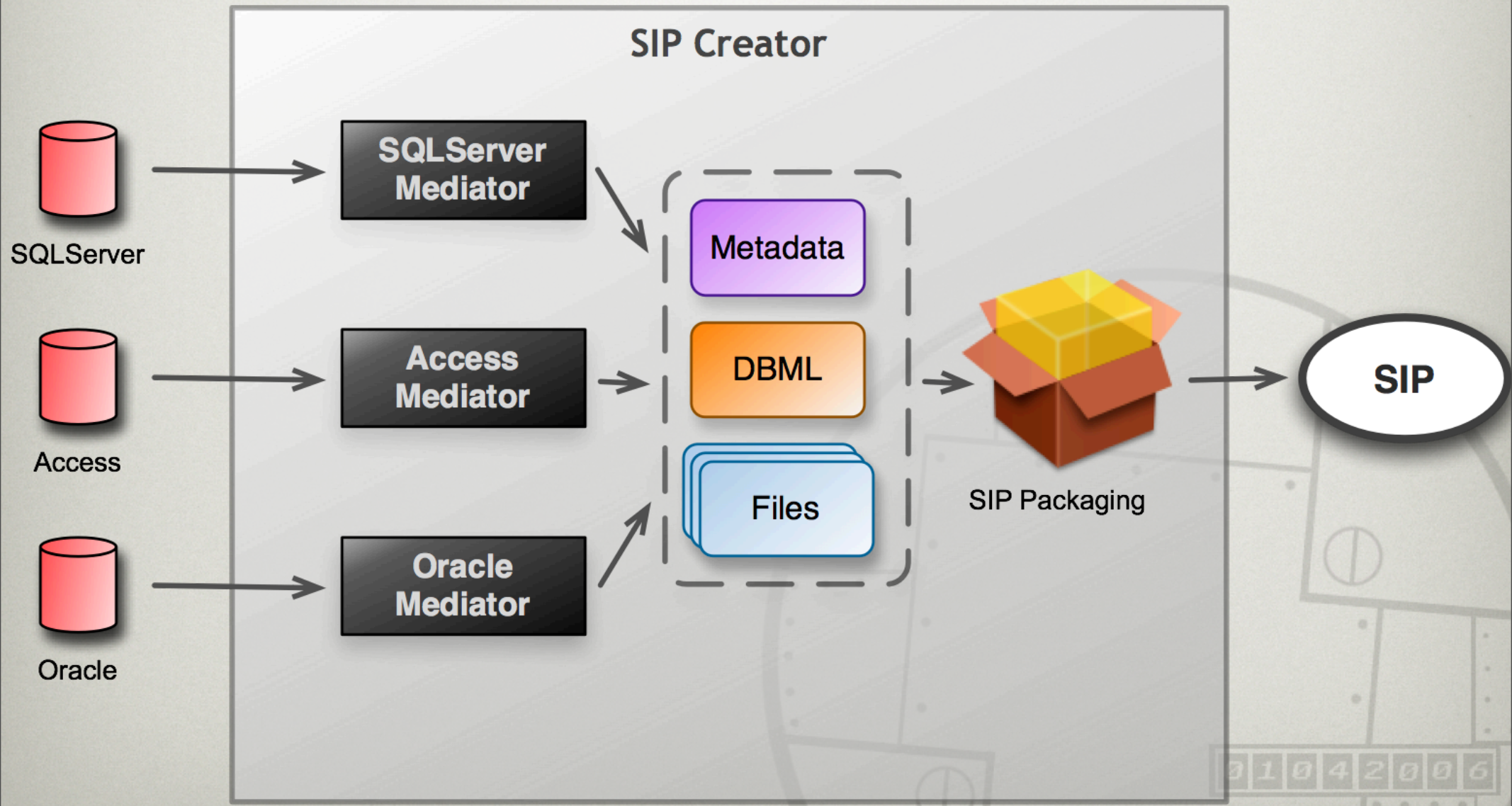
- Data?
- Structure?
- Views?
- Reports?
- Stored Procedures?
- ...

First prototype:

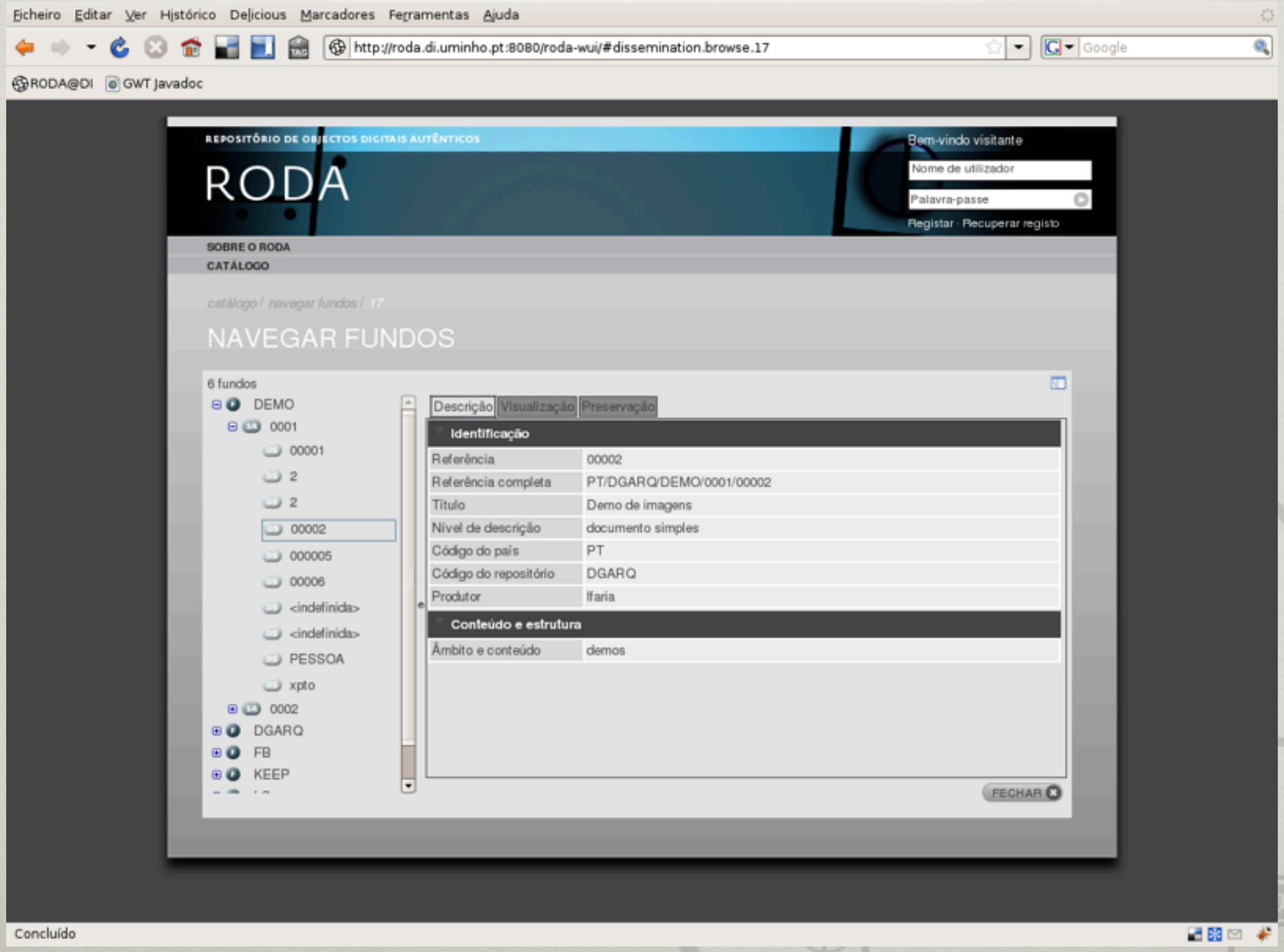
- Data
- Structure



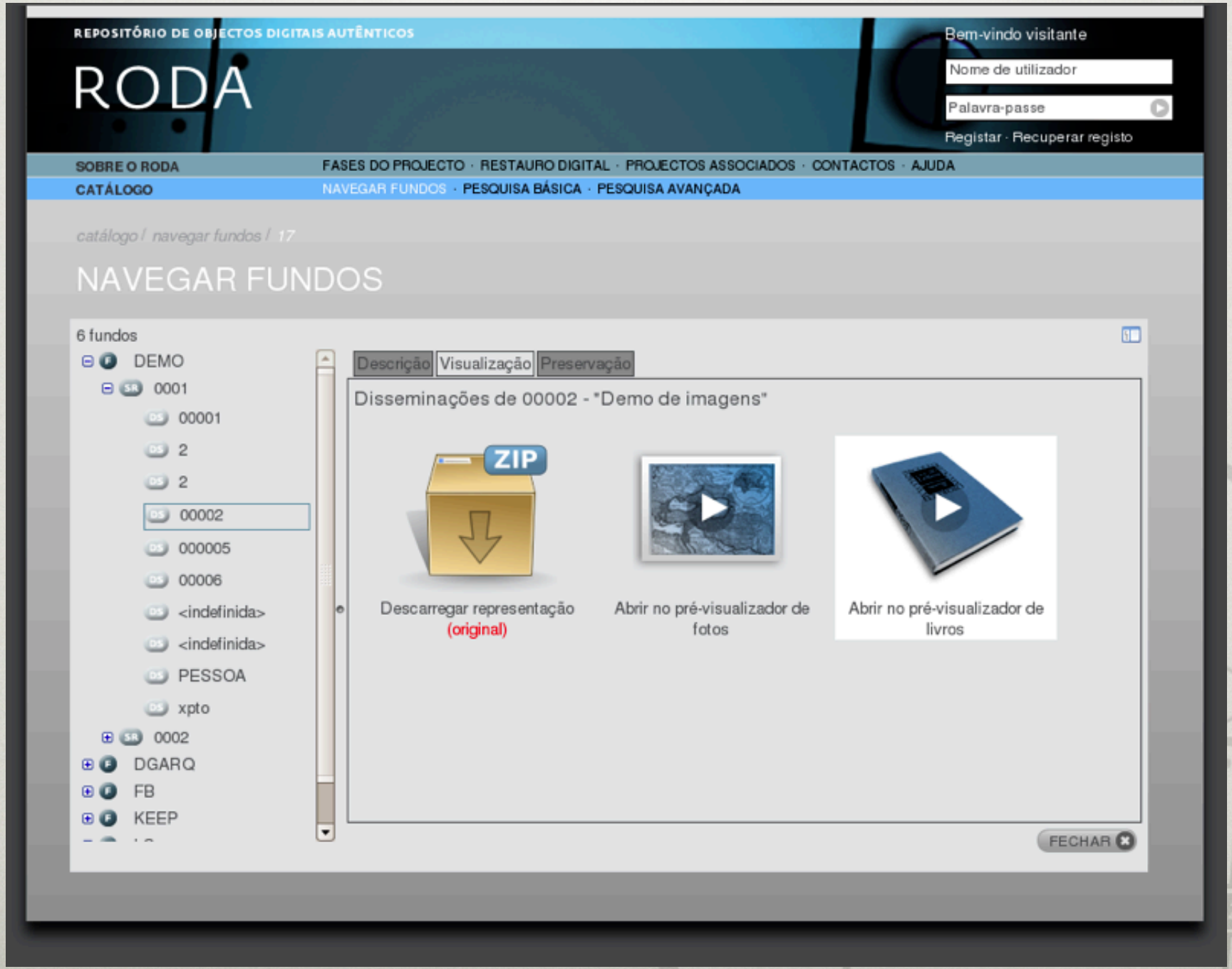
SIP Builder


SIP Creator
SQLServer
Access
Oracle
SQLServer Mediator
Access Mediator
Oracle Mediator
Metadata
DBML
Files
SIP Packaging
SIP

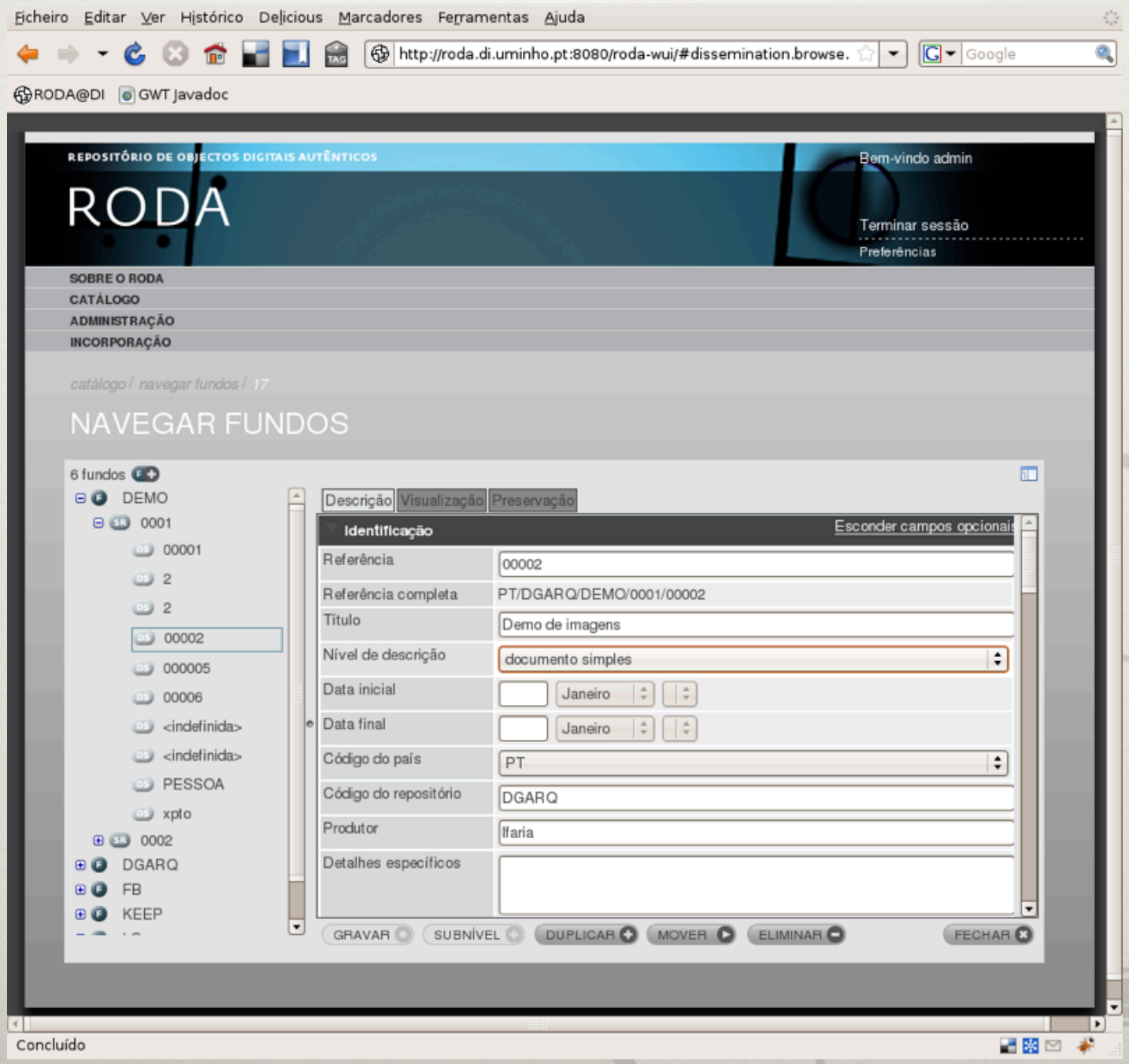
Browser


Ficheiro Editar Ver Histórico Delicious Marcadores Ferramentas Ajuda
http://roda.di.uminho.pt:8080/roda-wui/#dissemination.browse.17
Google
RODA@DI
GWT Javadoc
REPOSITÓRIO DE OBJECTOS DIGITAIS AUTÊNTICOS
RODA
Bem-vindo visitante
Nome de utilizador
Palavra-passe
Registar · Recuperar registo
SOBRE O RODA
CATÁLOGO
catálogo / navegar fundos / 17
NAVEGAR FUNDOS
6 fundos
DEMO
0001
00001
2
2
00002
000005
00006
<indefinida>
<indefinida>
PESSOA
xpto
0002
DGARQ
FB
KEEP
Descrição
Visualização
Preservação
Identificação
Referência
00002
Referência completa
PT/DGARQ/DEMO/0001/00002
Título
Demo de imagens
Nível de descrição
documento simples
Código do país
PT
Código do repositório
DGARQ
Produtor
Ifaria
Conteúdo e estrutura
Âmbito e conteúdo
demos
FECHAR
Concluído

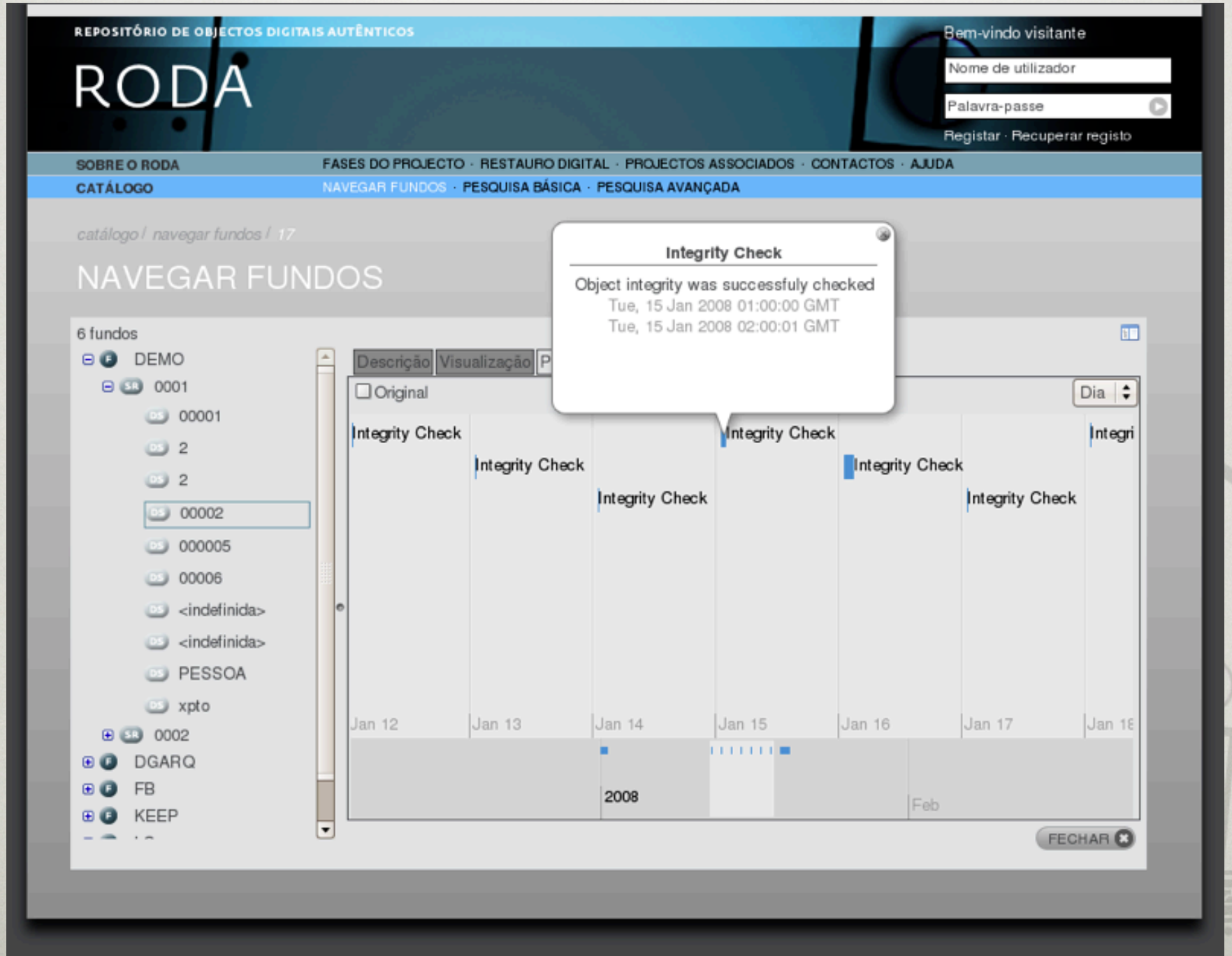
Disseminators



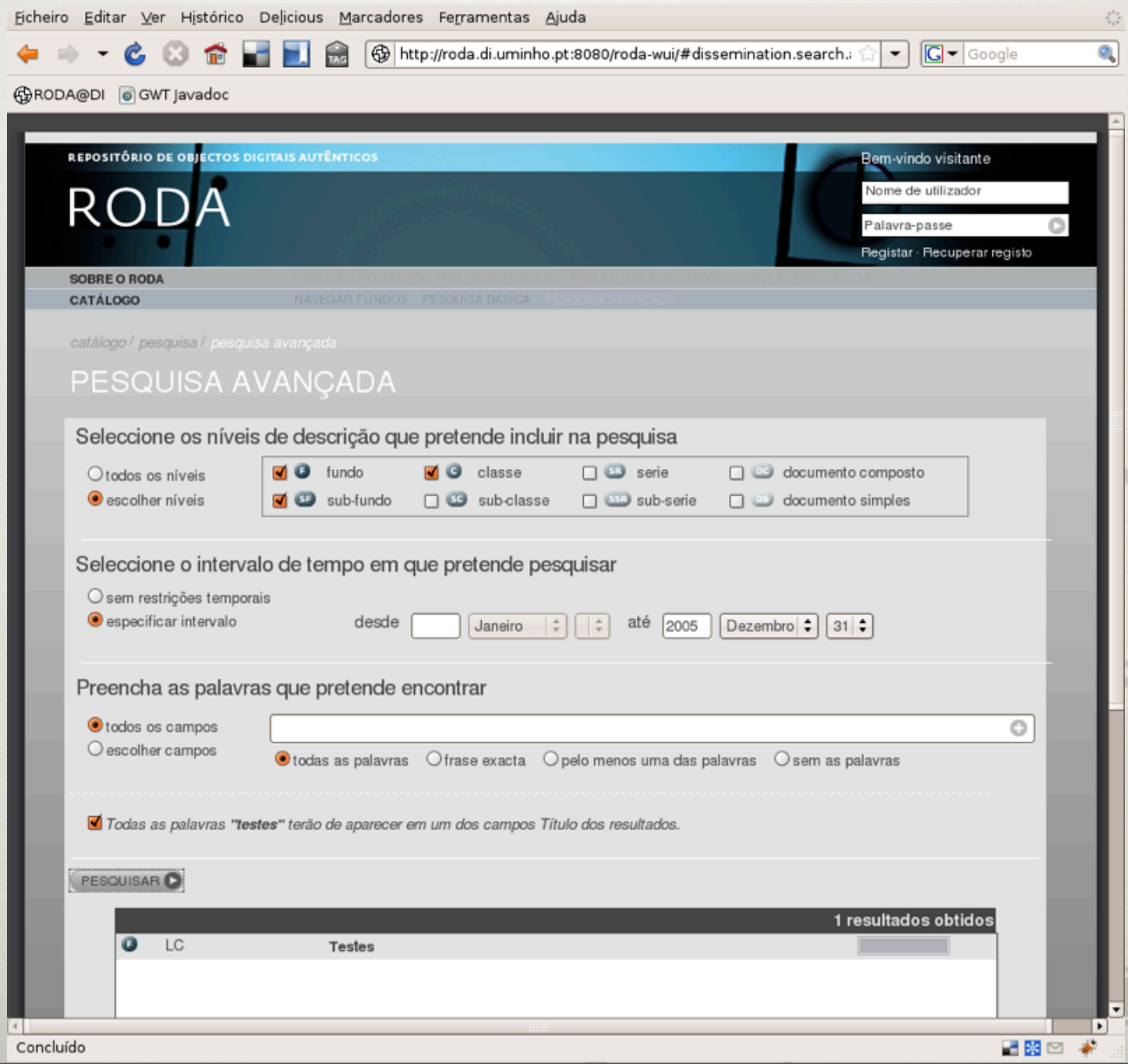
Metadata Editor



Preservation Metadata Viewer

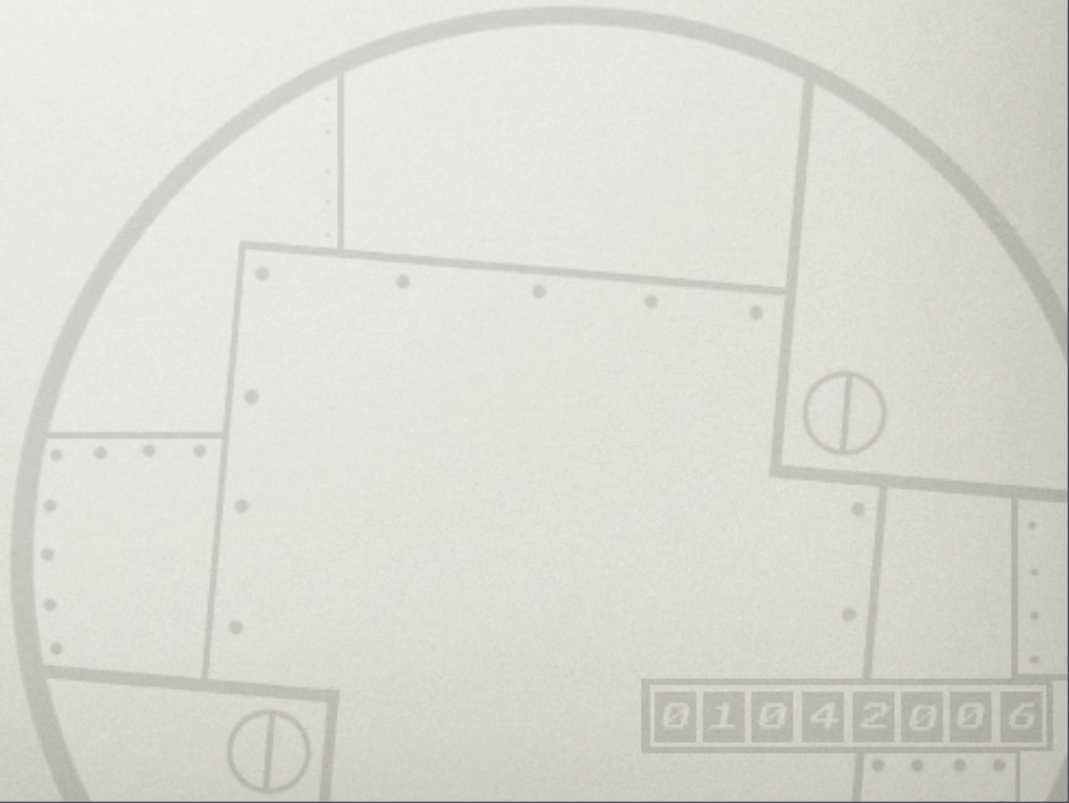


Advanced Search



Final thoughts

"Data Preservation is a people problem"
Michael Lesk



Final thoughts

"Data Preservation is a people problem"
Michael Lesk

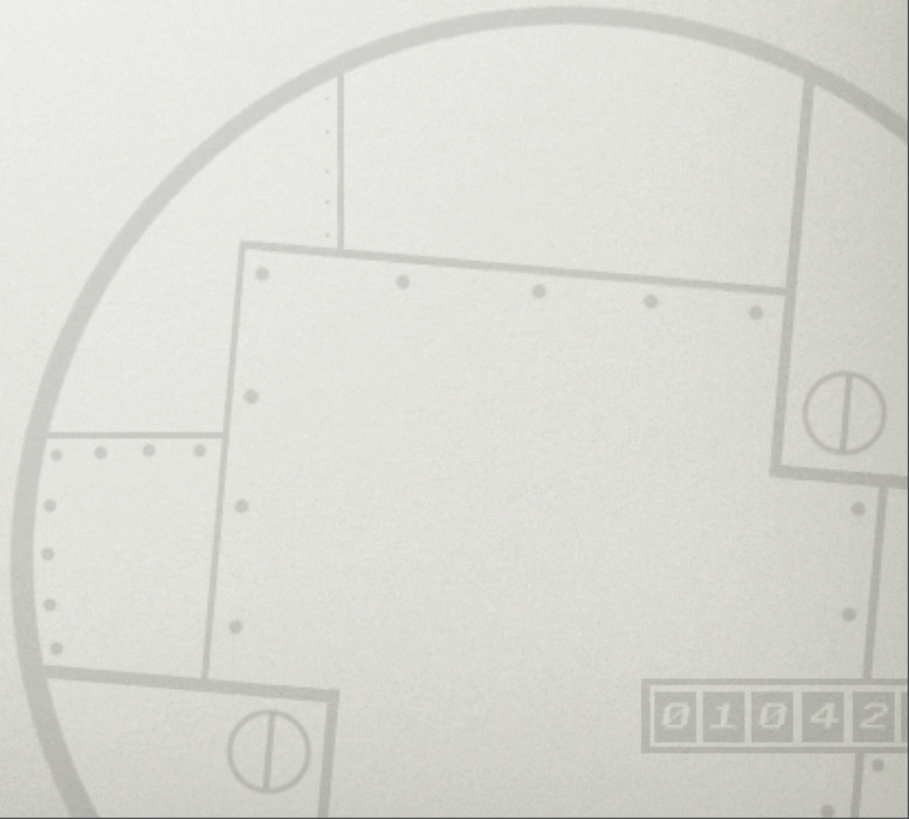
- People need to be trained to save data in a proper way.
- What to preserve? Data, Structure, Semantics...
- Preservation is for future users but only today users vote on budget
- We need to make data collecting people have preservation concerns
- Preservation is fault tolerance. All systems are imperfect

RODA Homepage

http://roda.iantt.pt/

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



Let's Preserve
Tomorrow's History...
Questions?