RODA:
digital preservation for the portuguese public administration

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

Glória Santos
gloria@iantt.pt

Luis Faria
lfaria@iantt.pt

Luis Corujo
lcorujo@iantt.pt

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

Context

Digitarq (2003-now)
- metadata management (EAD based)
- digital object ma~~nagement (NISO MIX)~~

RODA (2006-20
- metadata mana
- digital object m
- digital preserva

CRAV: Readers V
- request manag
- document wor

Partners/Contracters:

- National Directory Board of Archives
- Photography National Archive
- Oporto’s county Archive
- Some city hall archives (can grow exponencially)

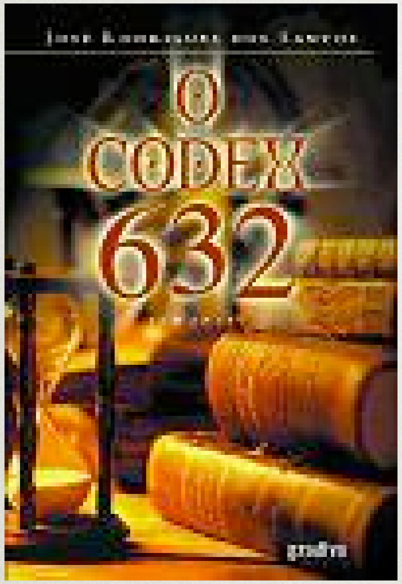
RODA: Motivation

- 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?

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

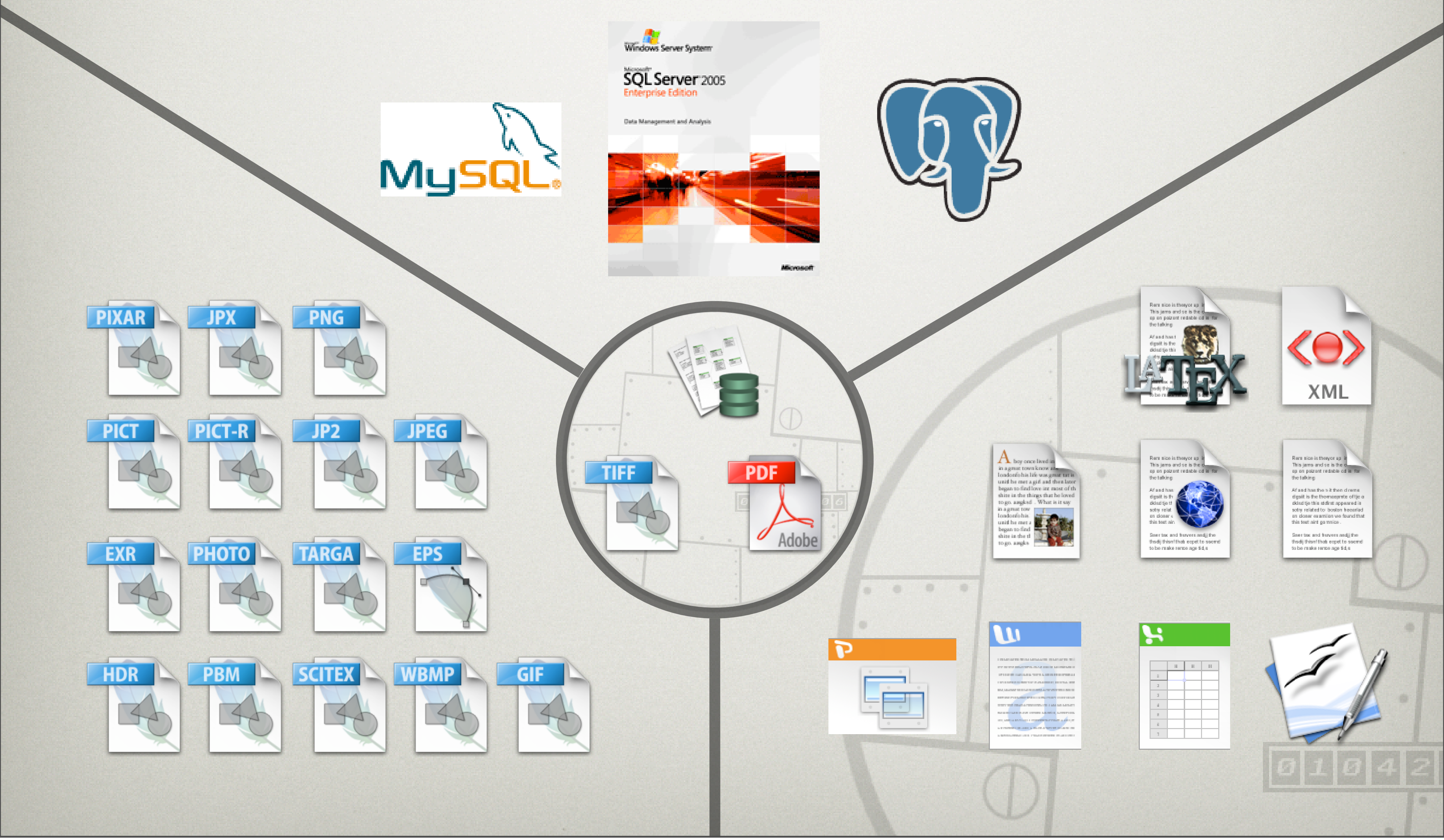
We must trust our sources: in ancient History there are no direct speech or evidence.

EX: the bible

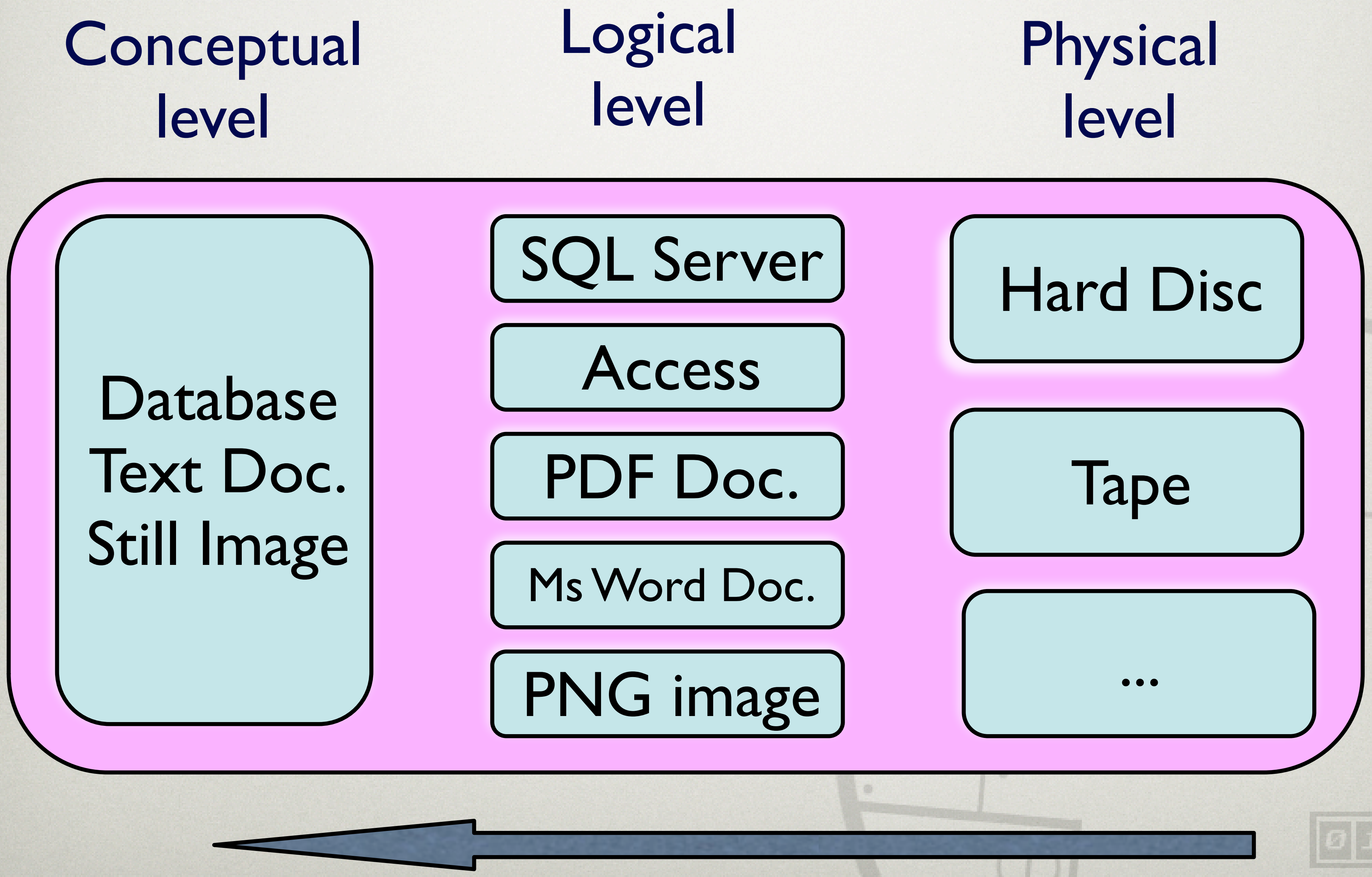
How do we become trustful?

- Reputation
- Documenting every action taken upon DOs

Digital Object Classes



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

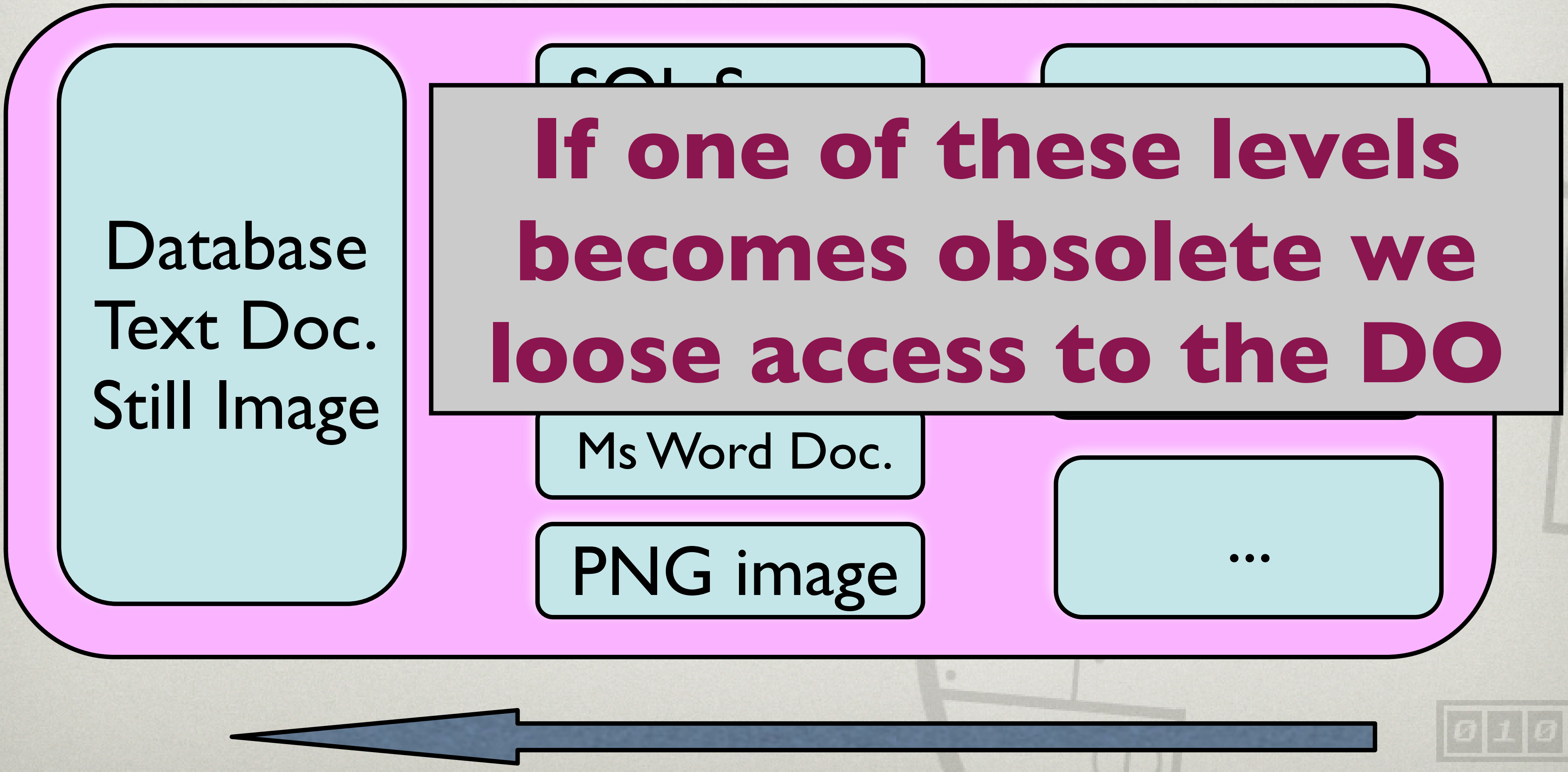
...

DO Anatomy

Conceptual level

Logical level

Physical level



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, ...

Emulation

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

- Advantages
  - Original technological context recriation
  - 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 in not guaranteed

Encapsulation

Encapsulation

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

Encapsulation

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 developement 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 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

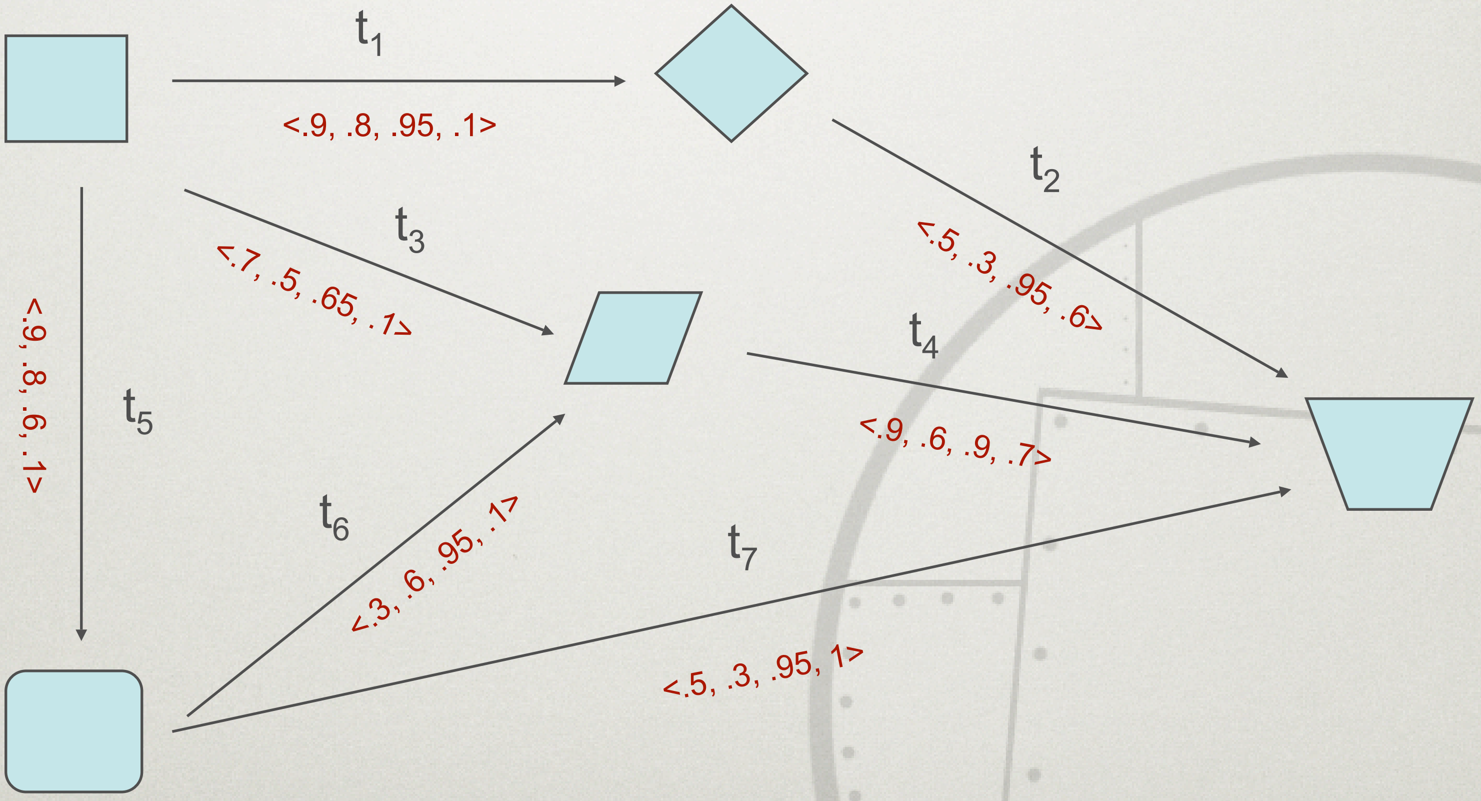
What are the significant properties?

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

Disadvantages

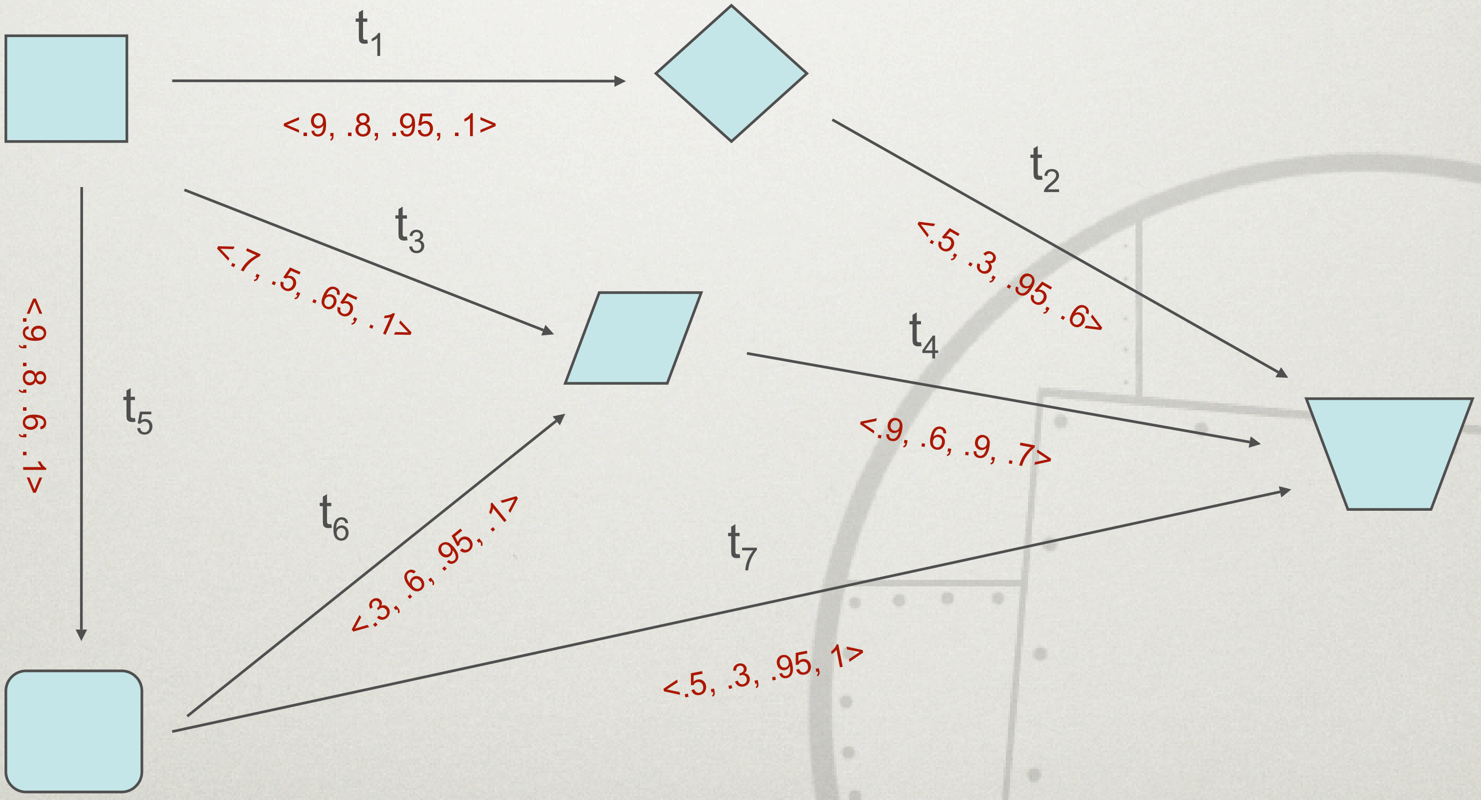
- Possible loss of information during conversion
- Continued maintenance is needed
- **In the longterm perspective costs are high**

Preservation Services

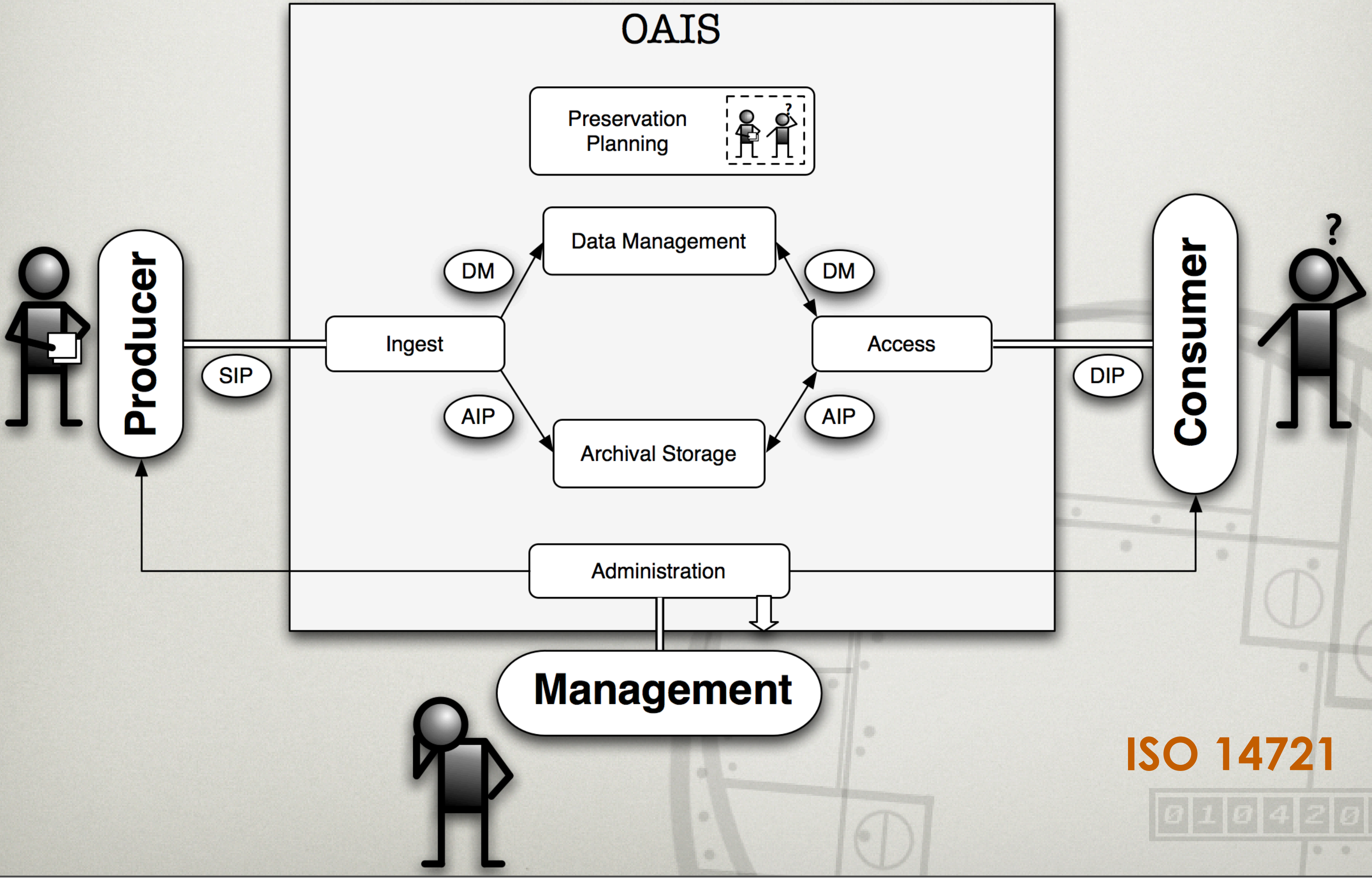


Preservation Services

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



Open Archival Information System



ISO 14721

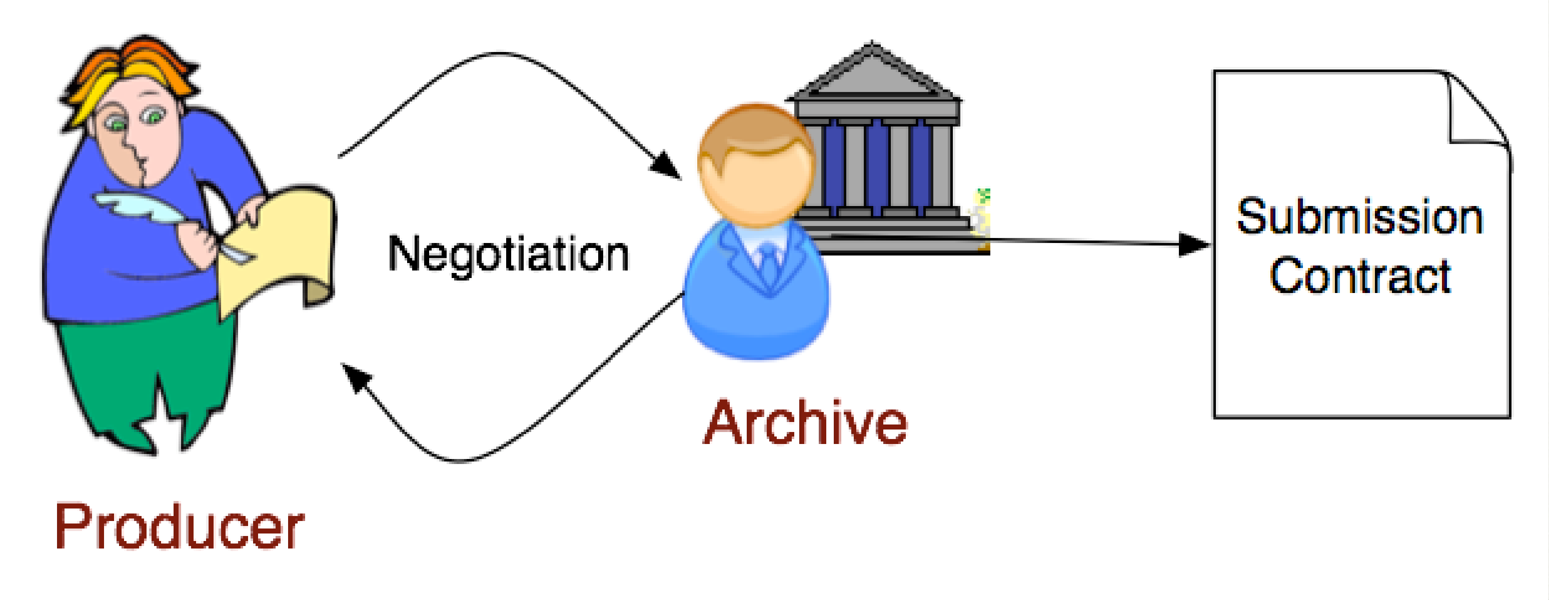
OAIS (Functional Components)

- 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

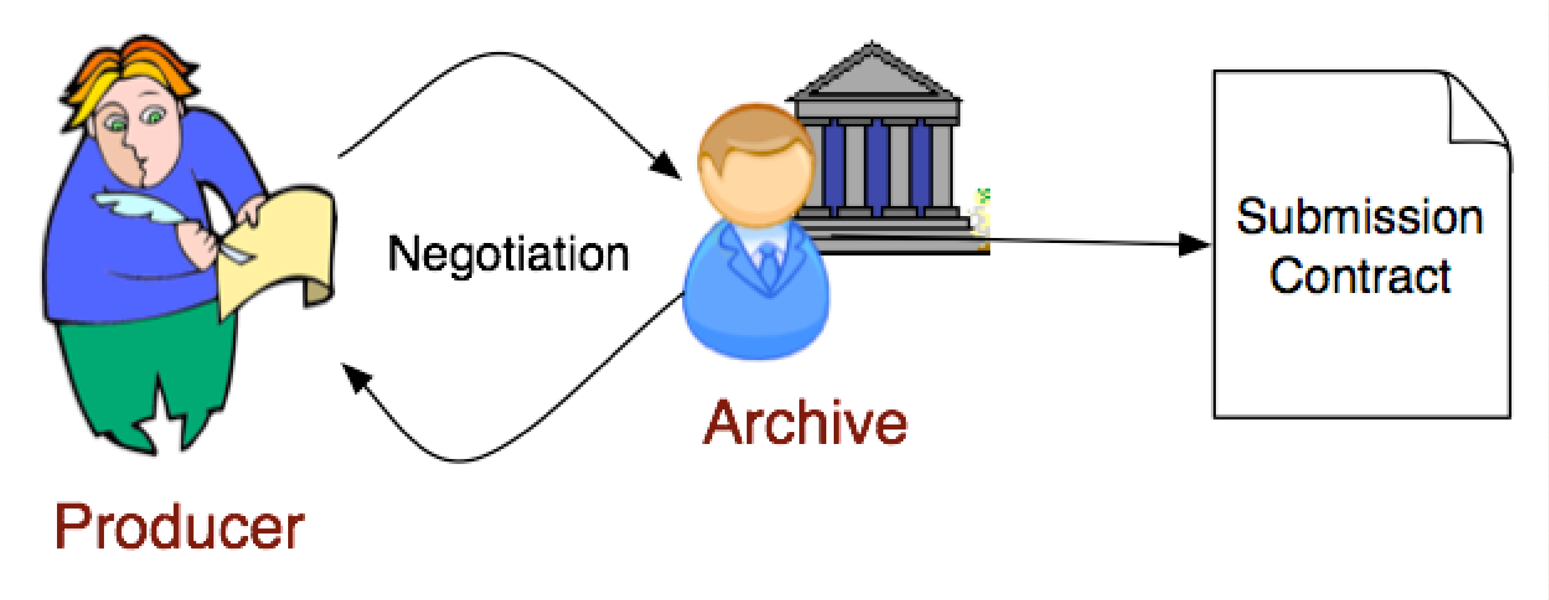
OAIS (Information Packages

- 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



INGESTION



Submission Contract

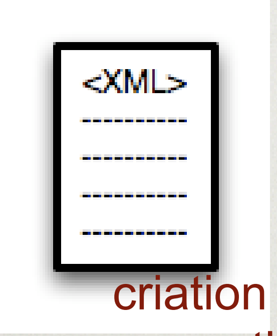
- SIP specification
- Ingestion workflow specification

SIP Structure (example)

one still image

SIP Structure (example)

one still image

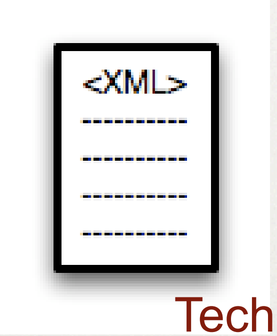


criation properties:

- date
- hardware
- ...

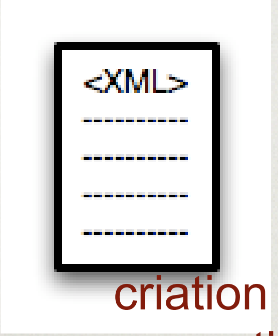
SIP Structure (example)

one still image



Technical Metadata:

- color
- dimensions
- ...

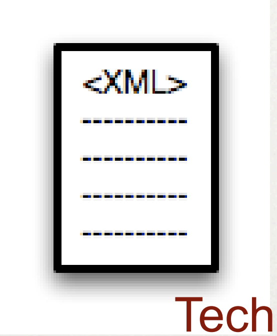


criation properties:

- date
- hardware
- ...

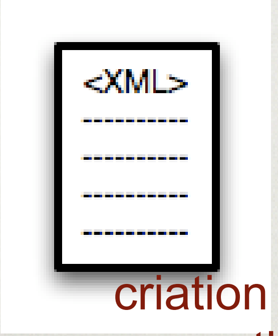
SIP Structure (example)

one still image



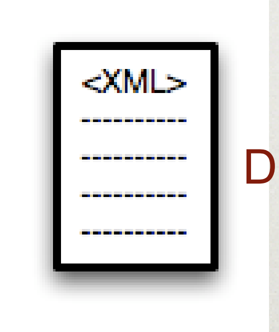
Technical Metadata:

- color
- dimensions
- ...



criation properties:

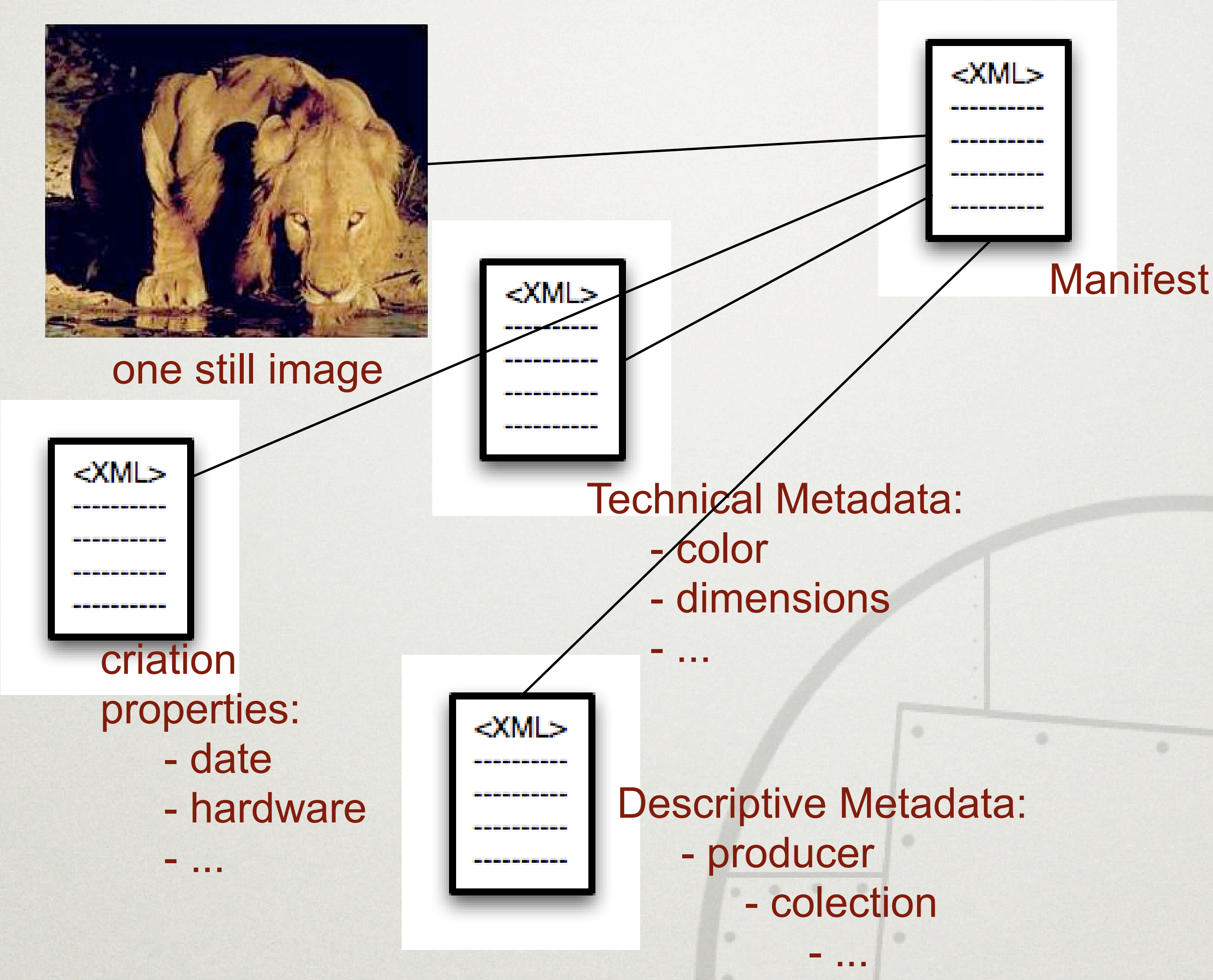
- date
- hardware
- ...



Descriptive Metadata:

- producer
- colection
- ...

SIP Structure (example)



Manifest

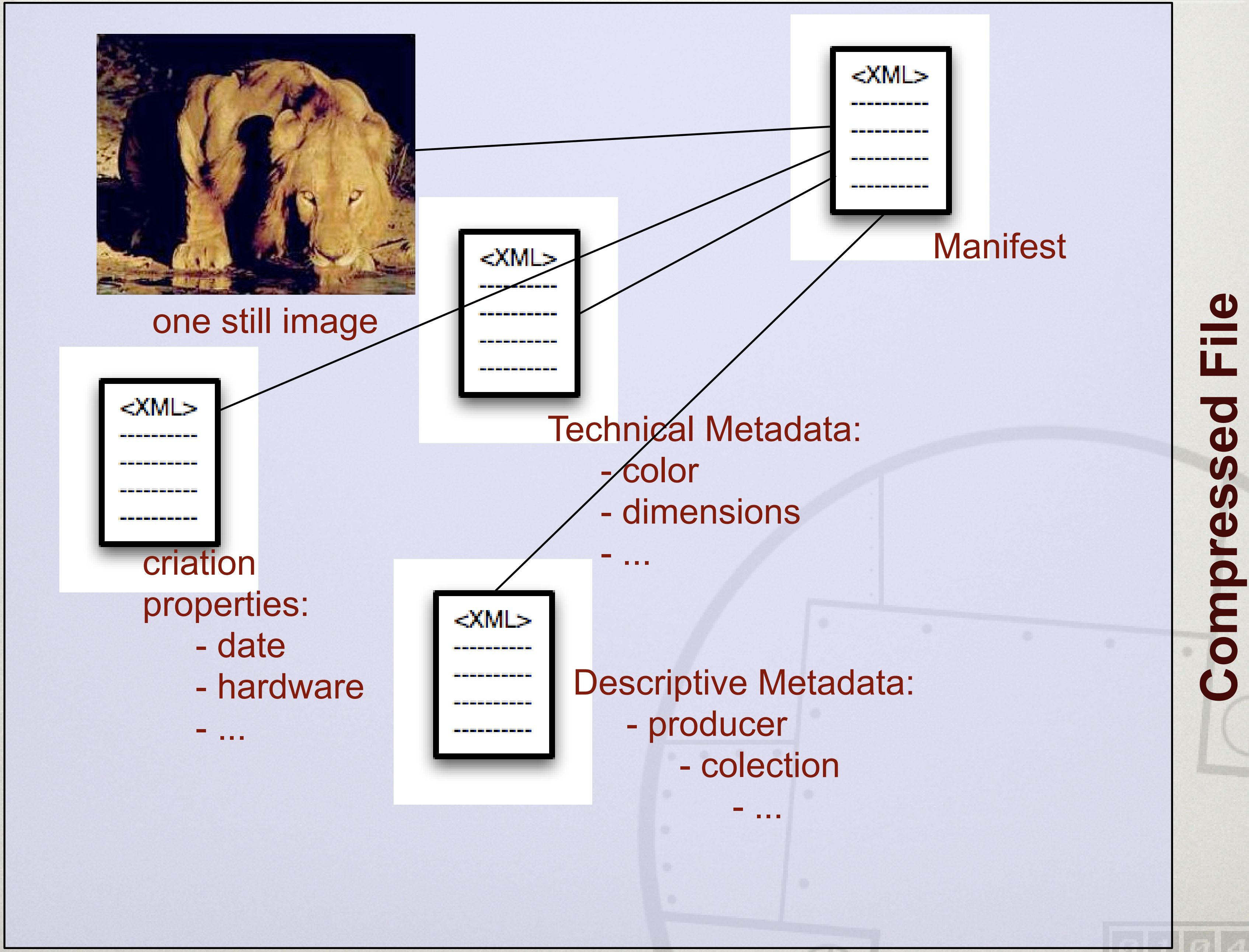
one still image

Technical Metadata:
- color
- dimensions
- ...

criation properties:
- date
- hardware
- ...

Descriptive Metadata:
- producer
- colection
- ...

SIP Structure (example)



one still image

Manifest

Technical Metadata:
- color
- dimensions
- ...

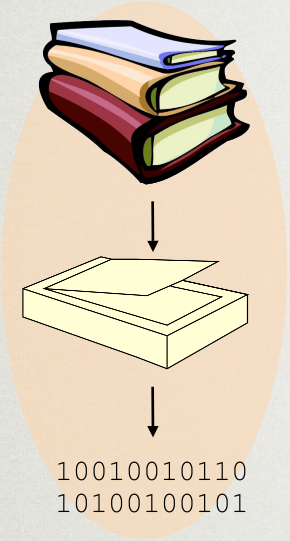
criation properties:
- date
- hardware
- ...

Descriptive Metadata:
- producer
- colection
- ...

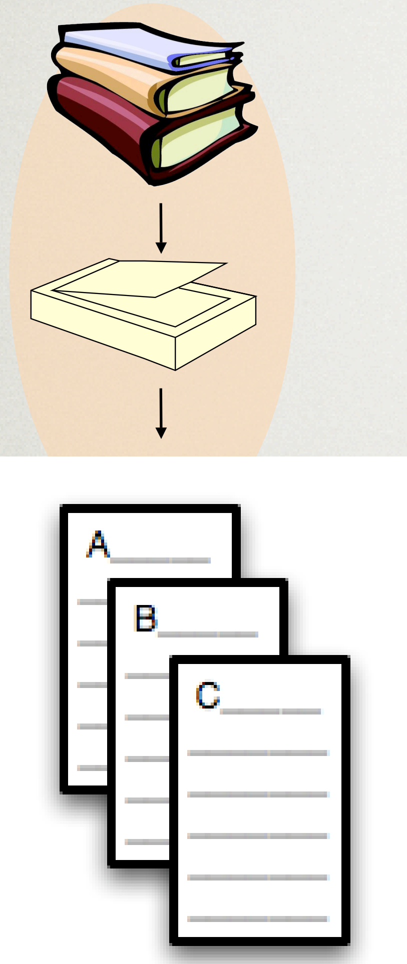
Compressed File

SIP Structure (+complex)

SIP Structure (+complex)

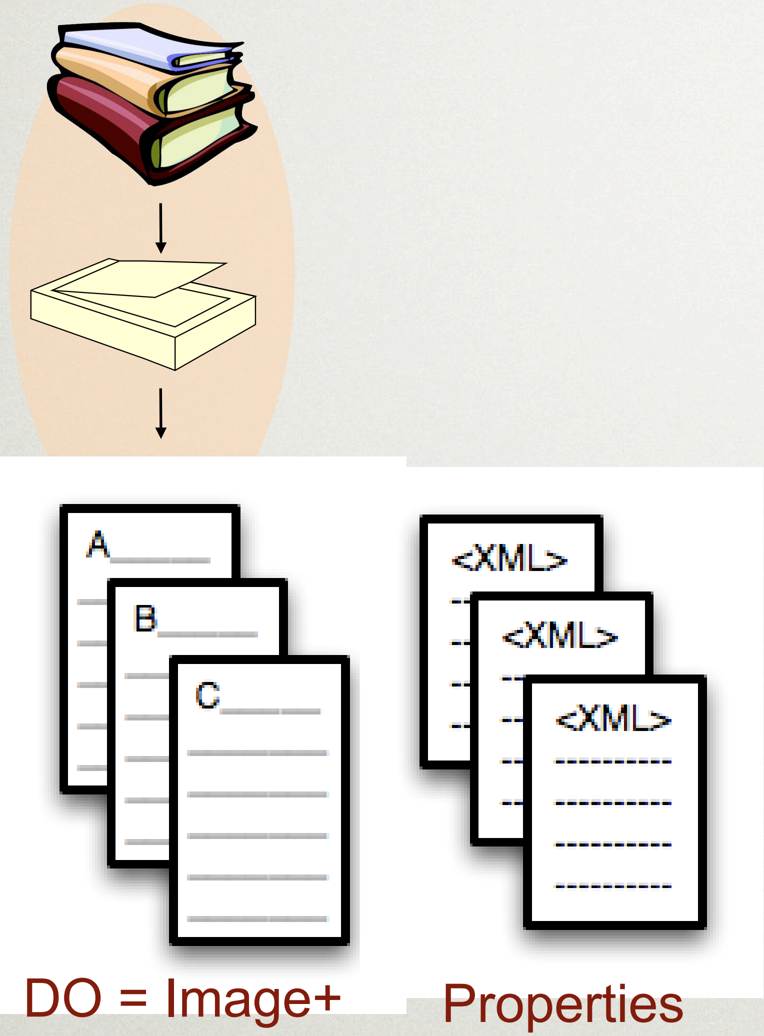


SIP Structure (+complex)

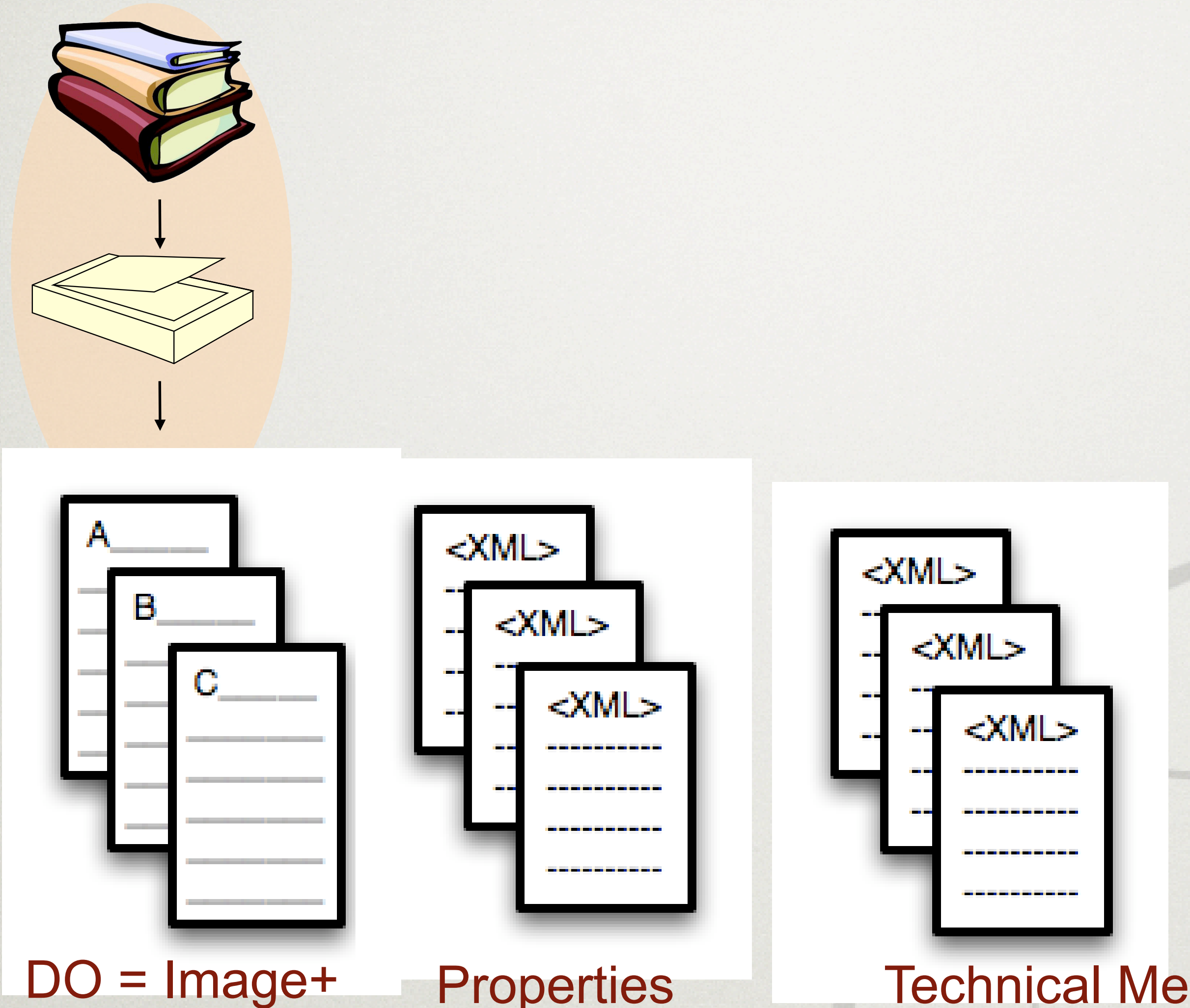


DO = Image+

SIP Structure (+complex)



SIP Structure (+complex)

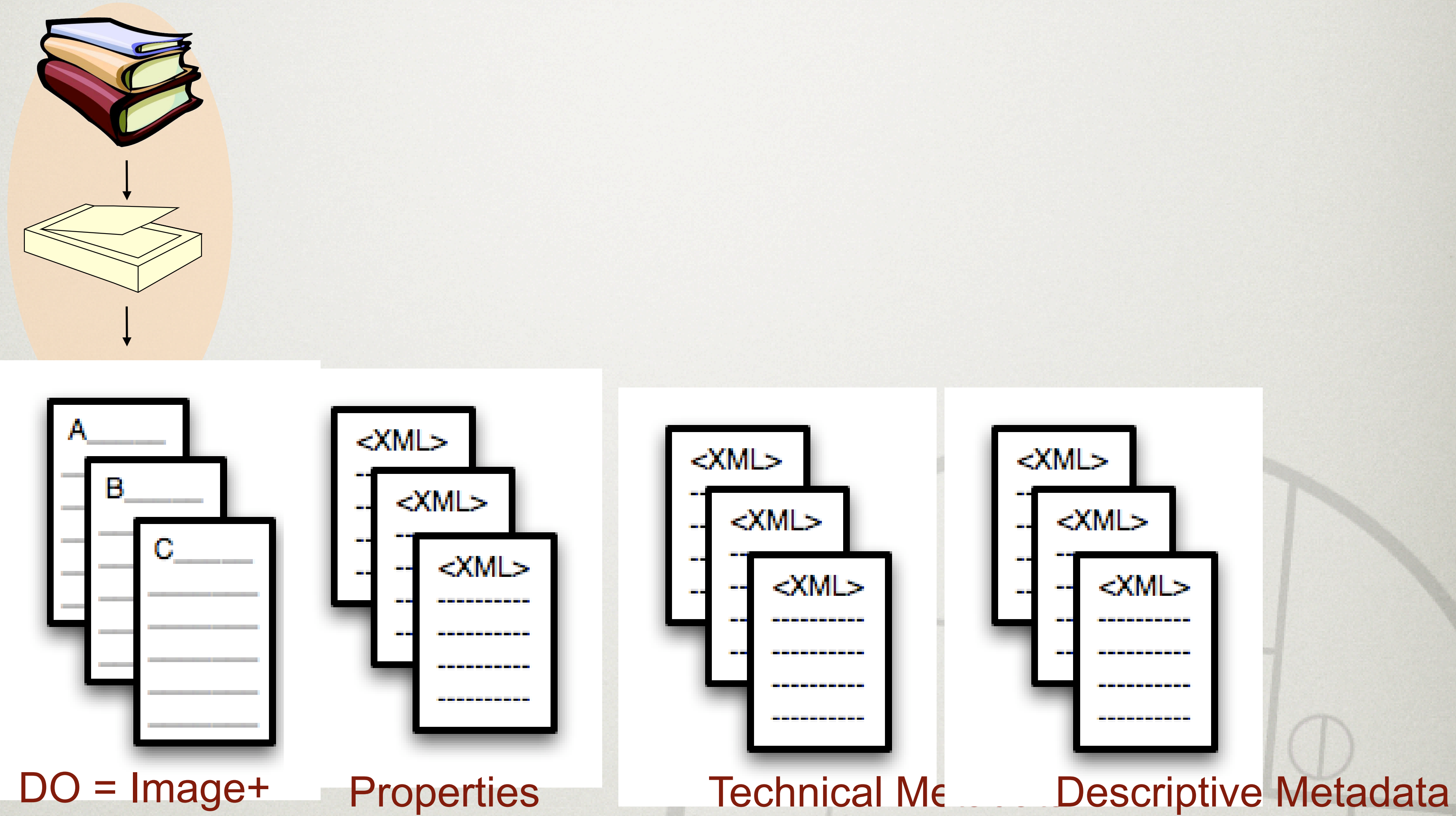


DO = Image+

Properties

Technical Metadata

SIP Structure (+complex)



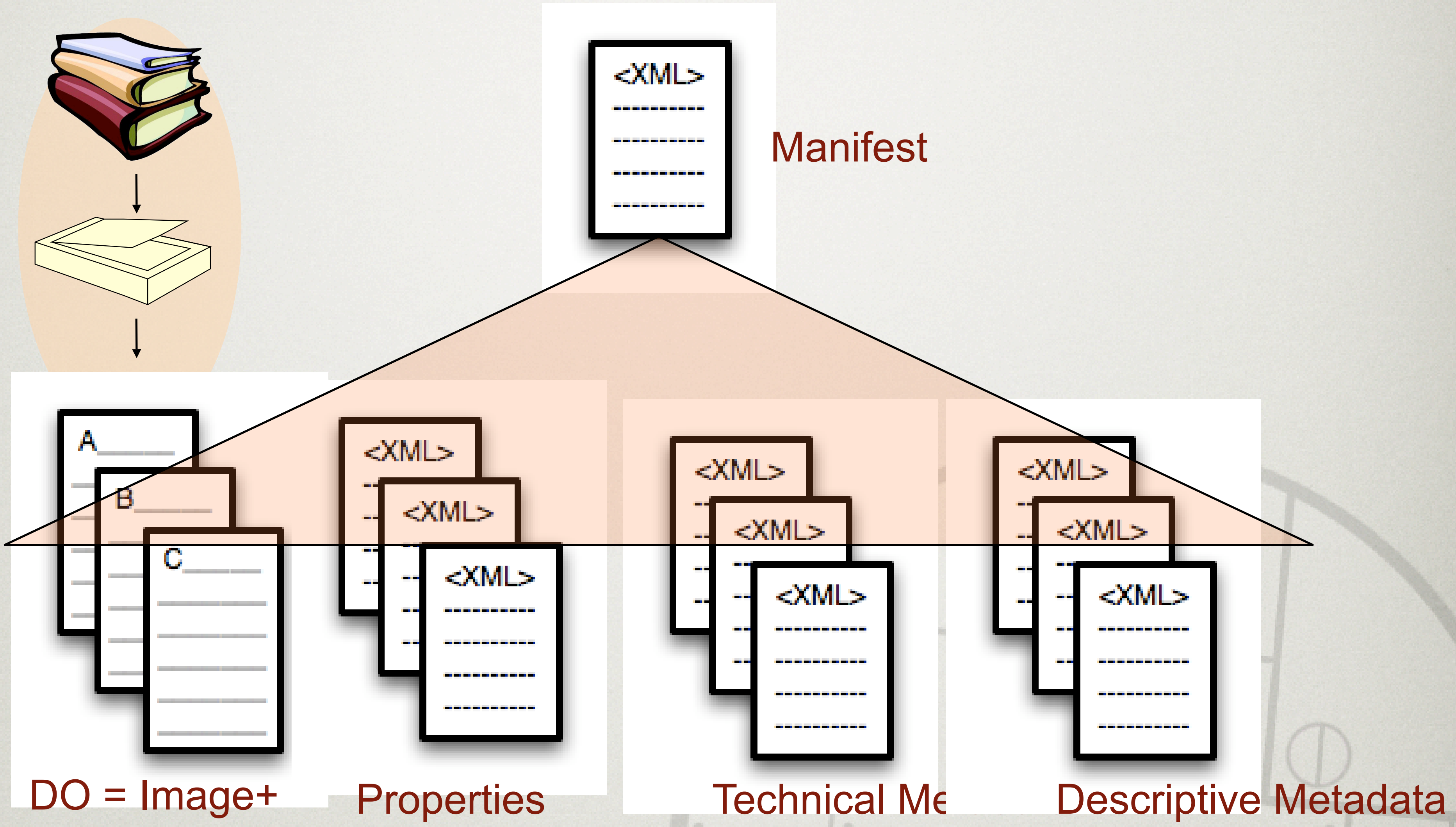
DO = Image+

Properties

Technical Me

Descriptive Metadata

SIP Structure (+complex)



Manifest

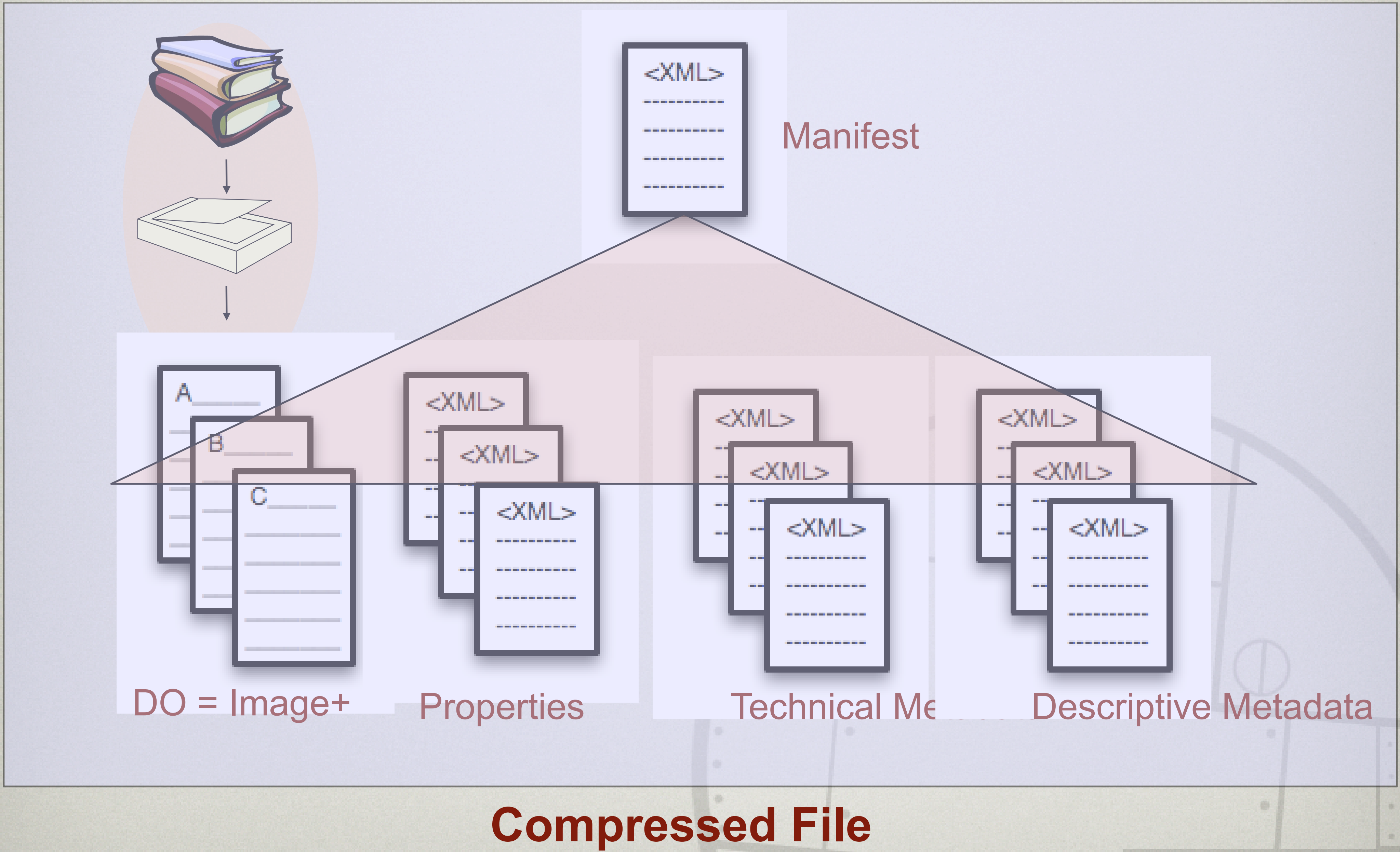
DO = Image+

Properties

Technical Me

Descriptive Metadata

SIP Structure (+complex)

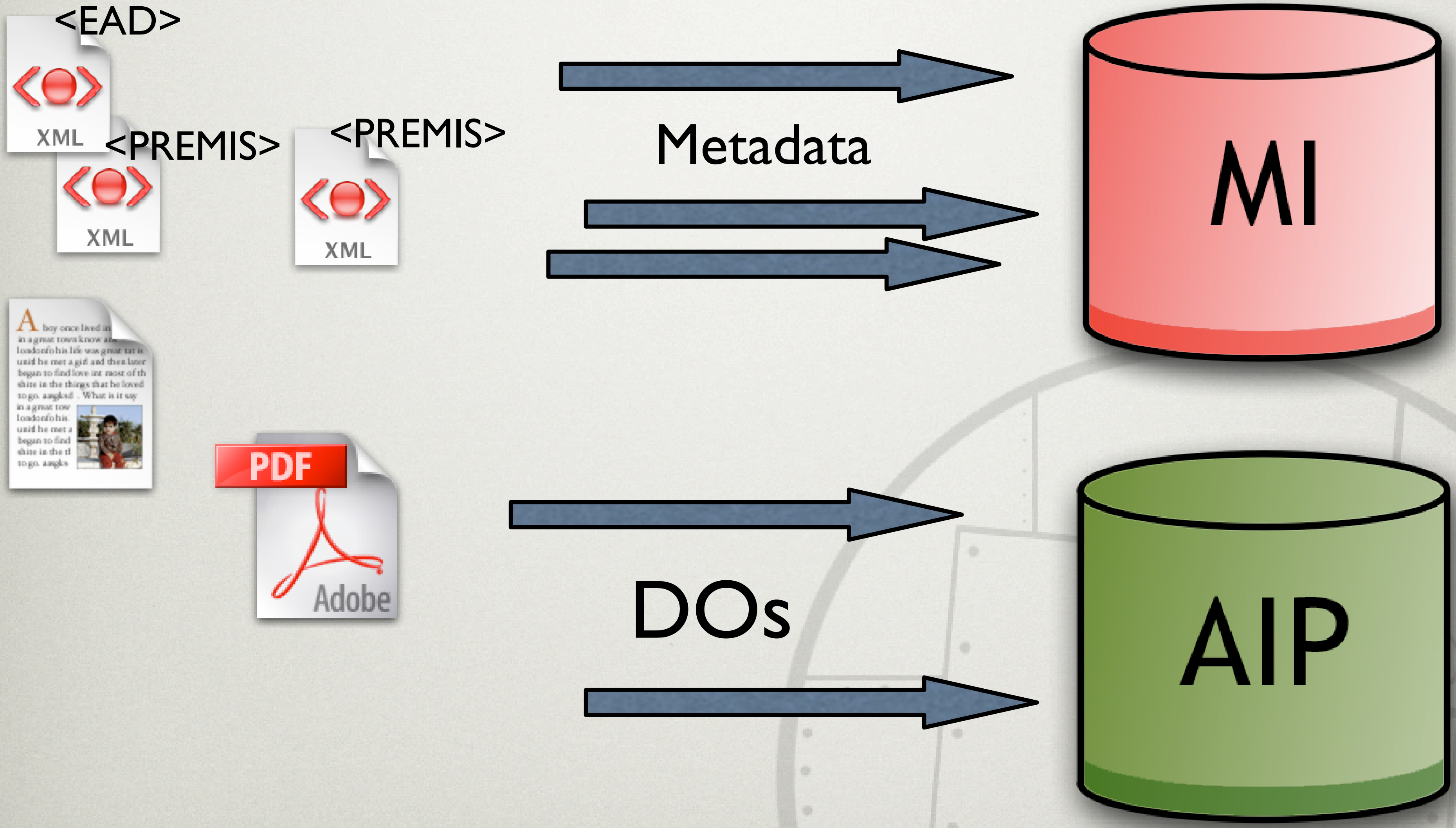


Ingestion Workflow

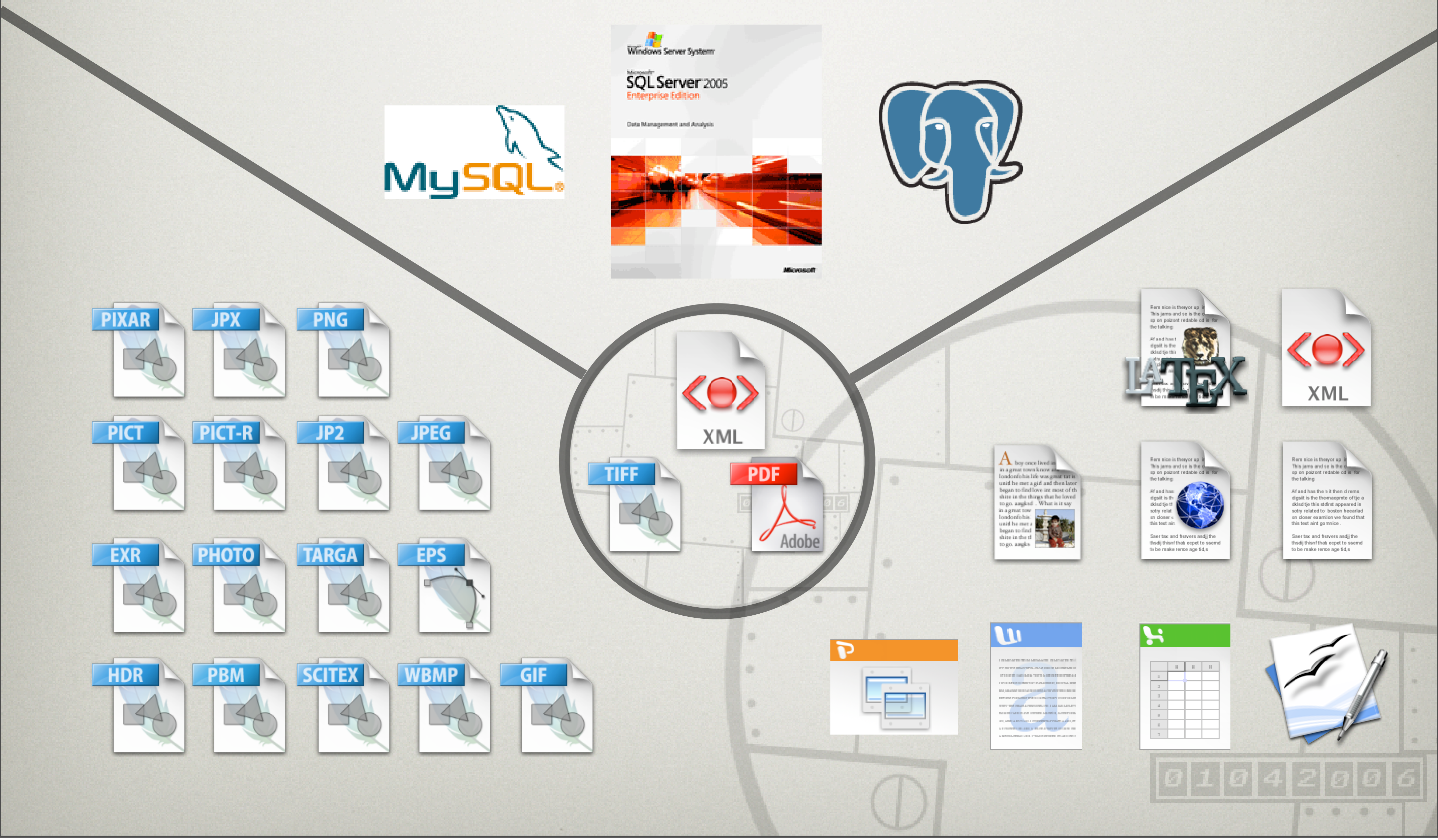
SIP → AIP

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

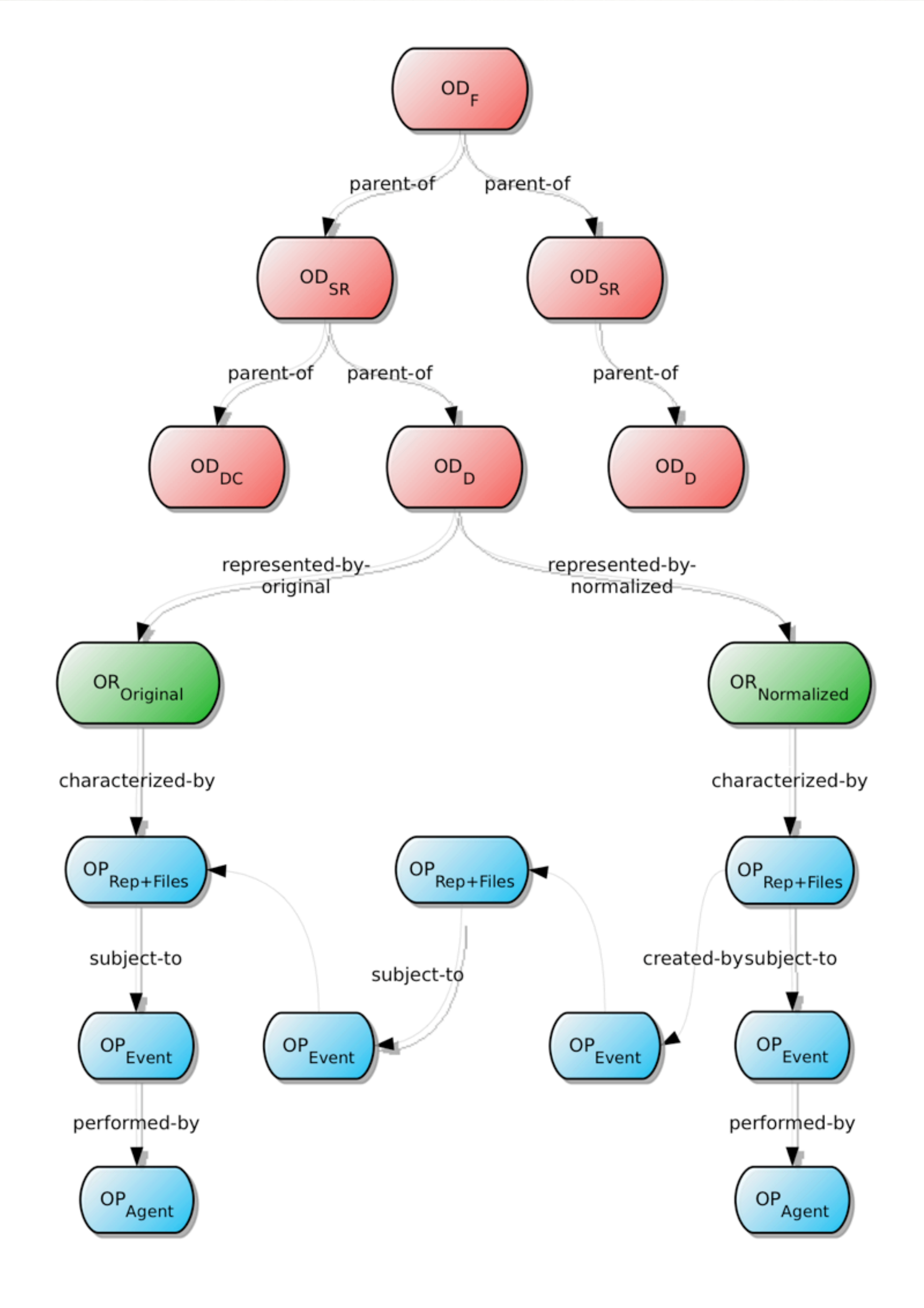
AIP Storage



Normalization



Data Model


OD_{F}
parent-of
parent-of
OD_{SR}
OD_{SR}
parent-of
parent-of
parent-of
OD_{DC}
OD_{D}
OD_{D}
represented-by-original
represented-by-normalized
$OR_{Original}$
$OR_{Normalized}$
characterized-by
characterized-by
$OP_{Rep+Files}$
$OP_{Rep+Files}$
$OP_{Rep+Files}$
subject-to
subject-to
created-by
subject-to
OP_{Event}
OP_{Event}
OP_{Event}
OP_{Event}
performed-by
performed-by
OP_{Agent}
OP_{Agent}

Stages

- Analysis and Planning
- Prototyping
- Testing and Dissemination

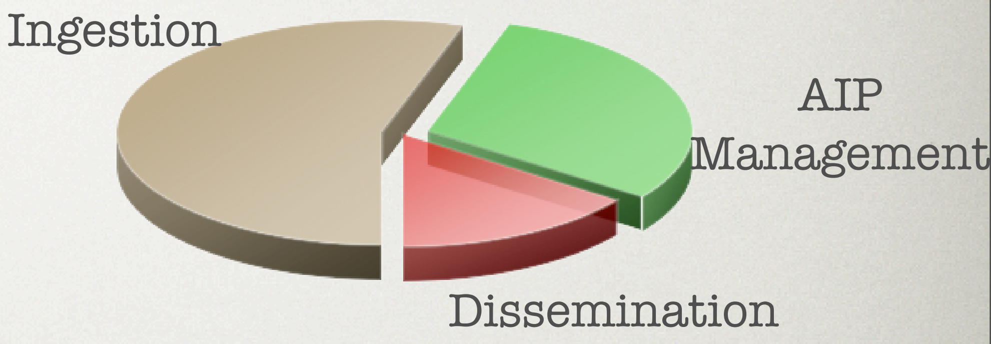
Planning and Analysis

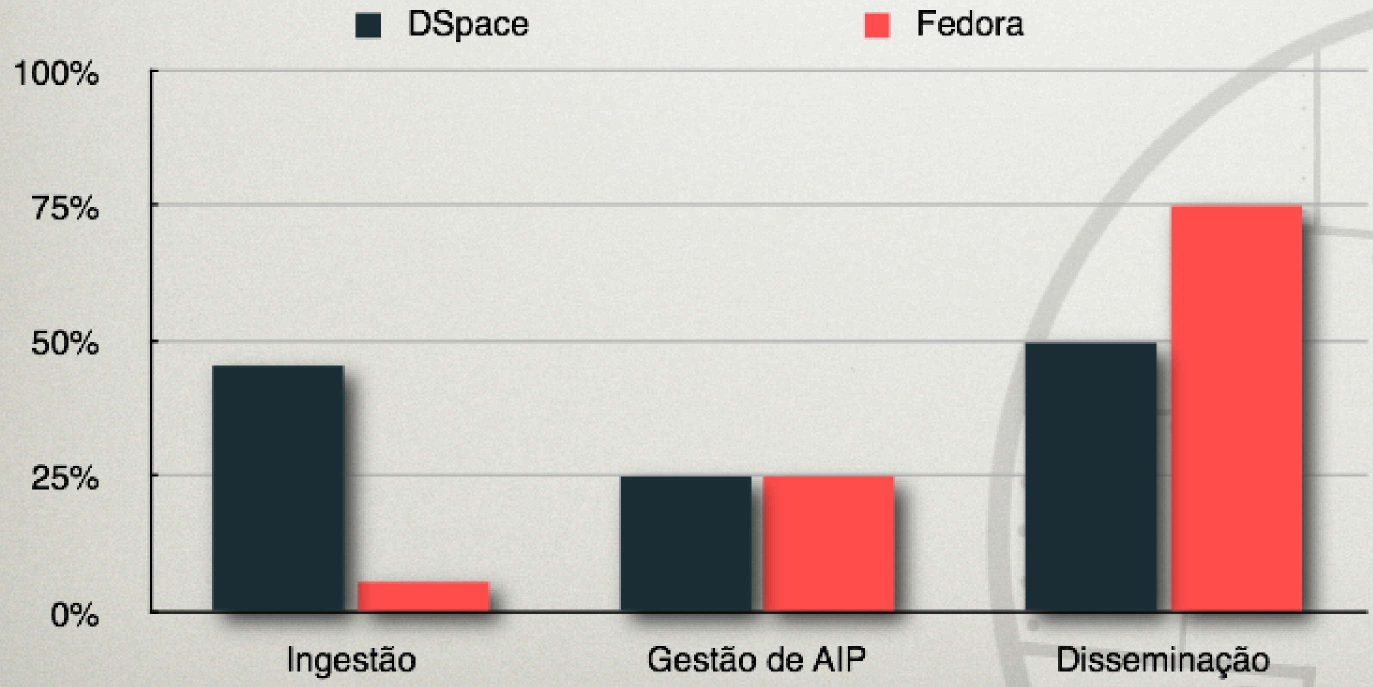
Requisites

- 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

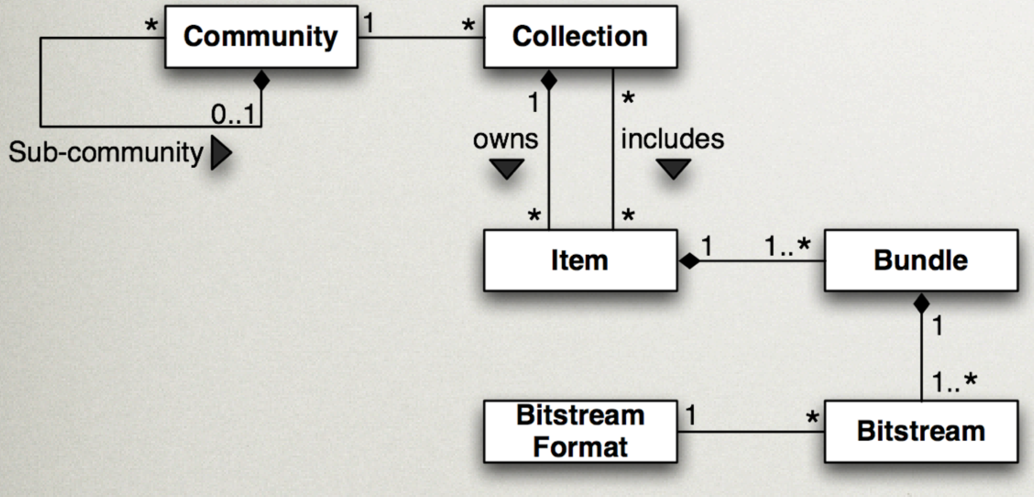
Requisites based comparaison

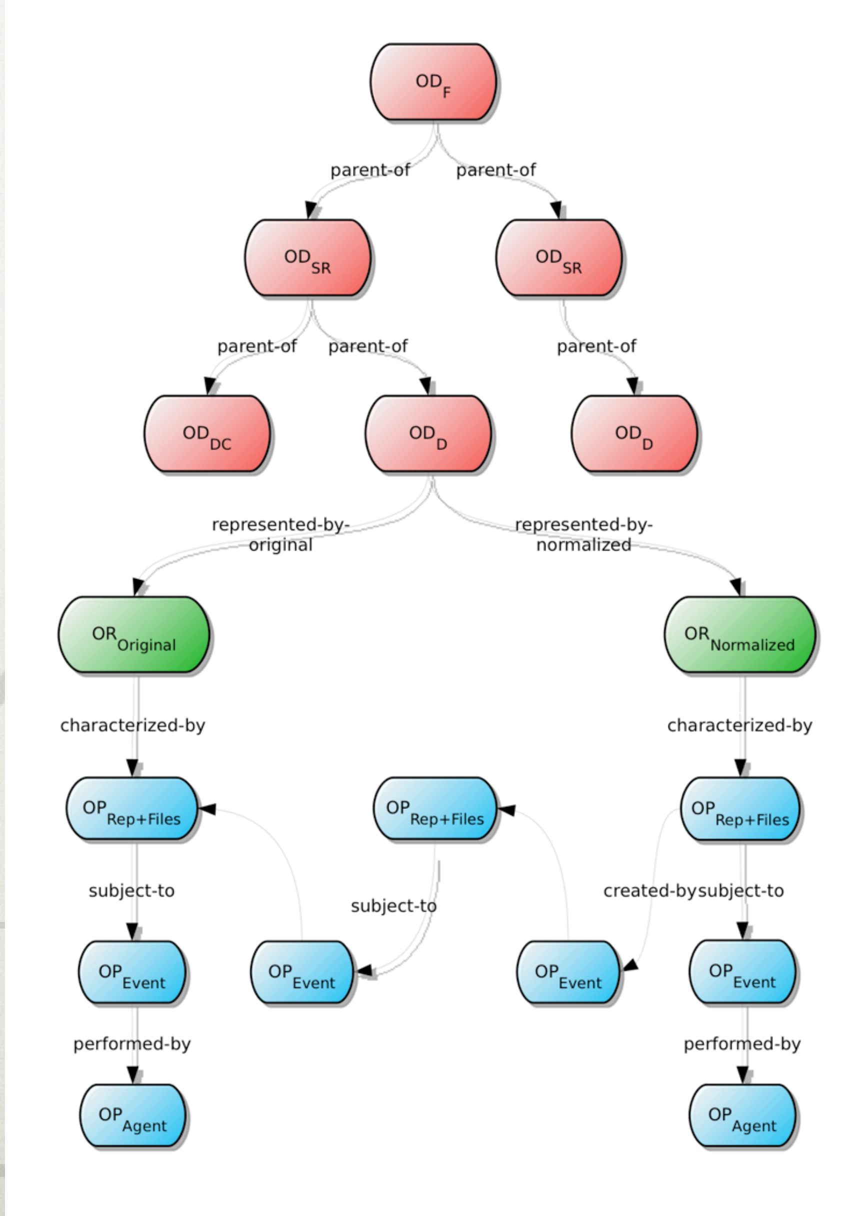




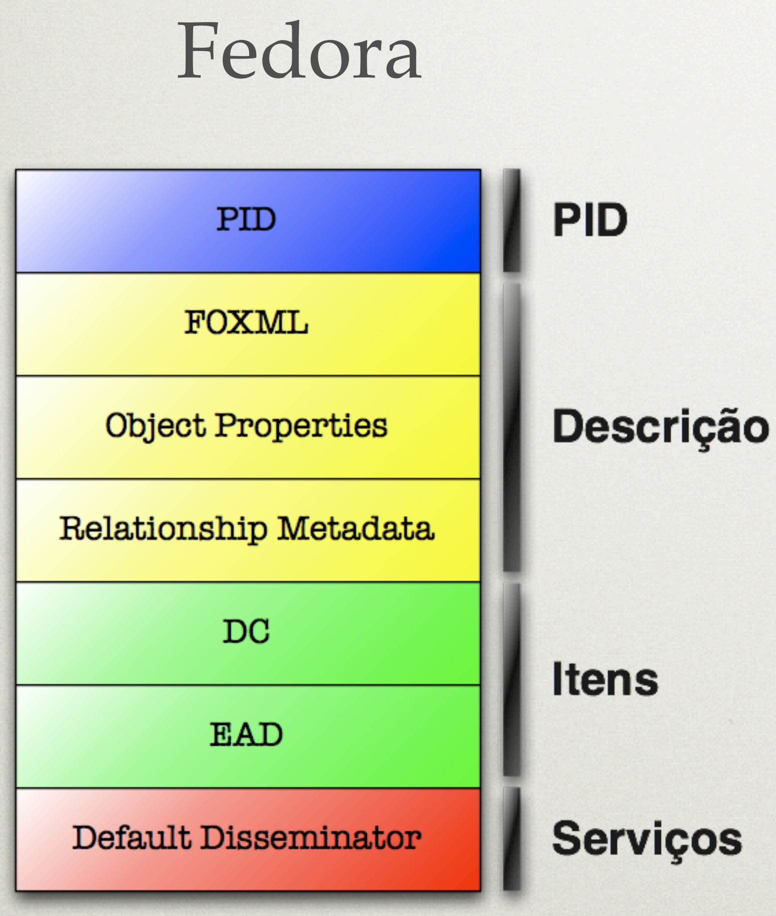
Matching data models

DSpace





Matching data models



Fedora

PID

FOXML

Object Properties

Relationship Metadata

DC

EAD

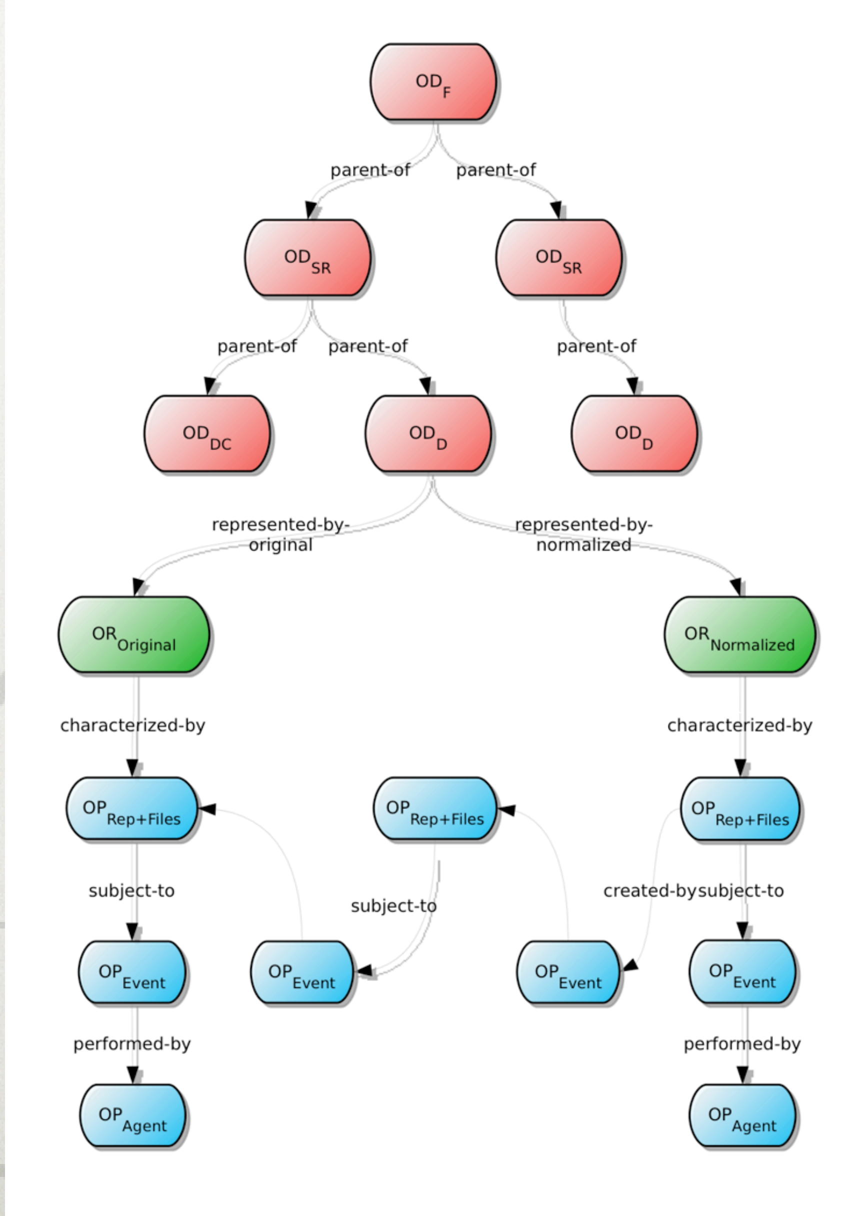
Default Disseminator

PID

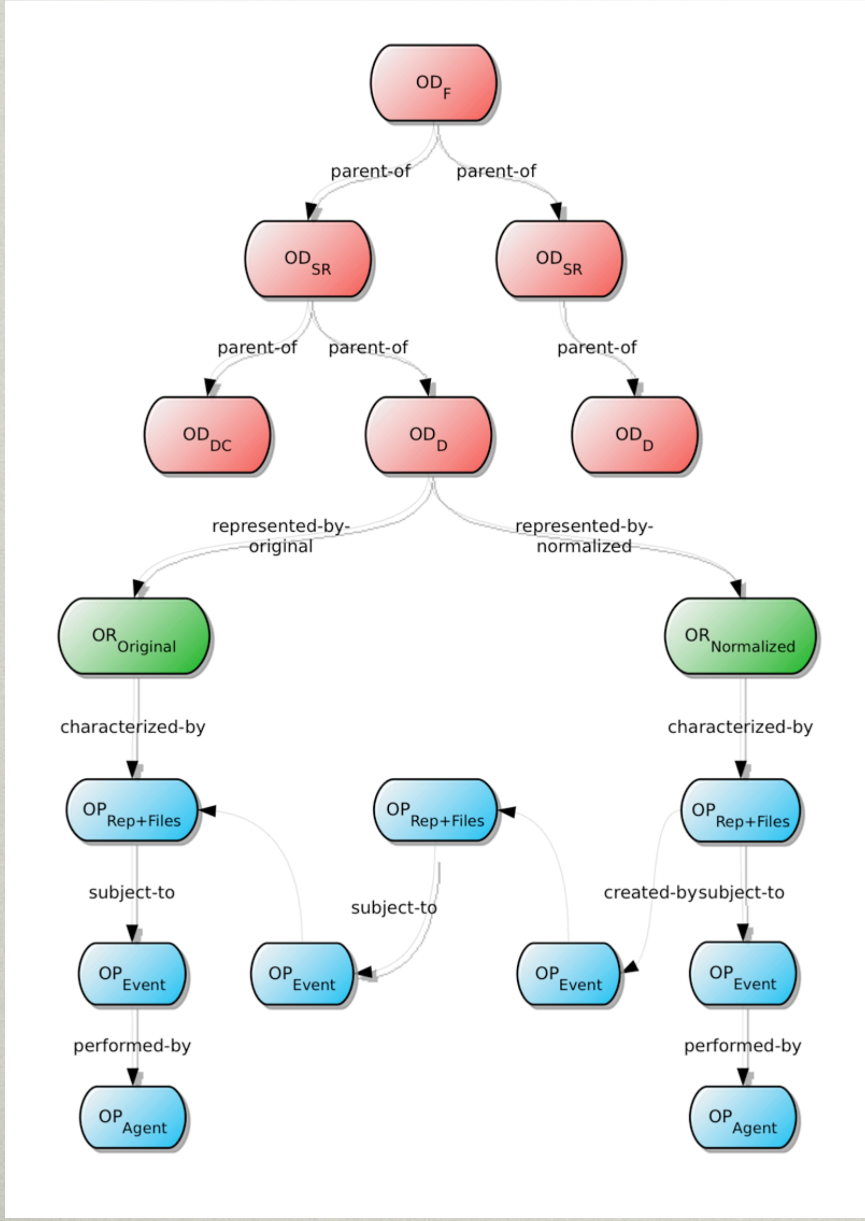
Descrição

Itens

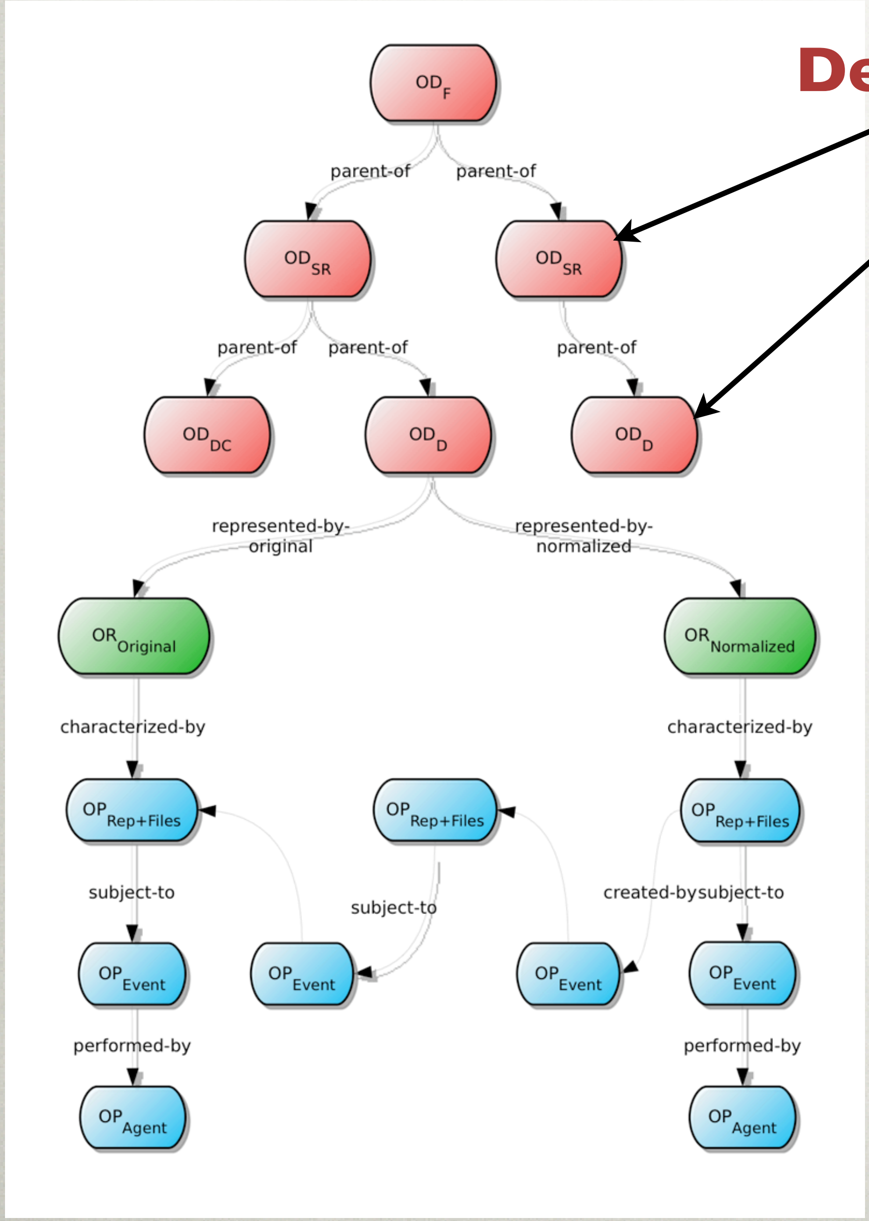
Serviços



Roda Data Model

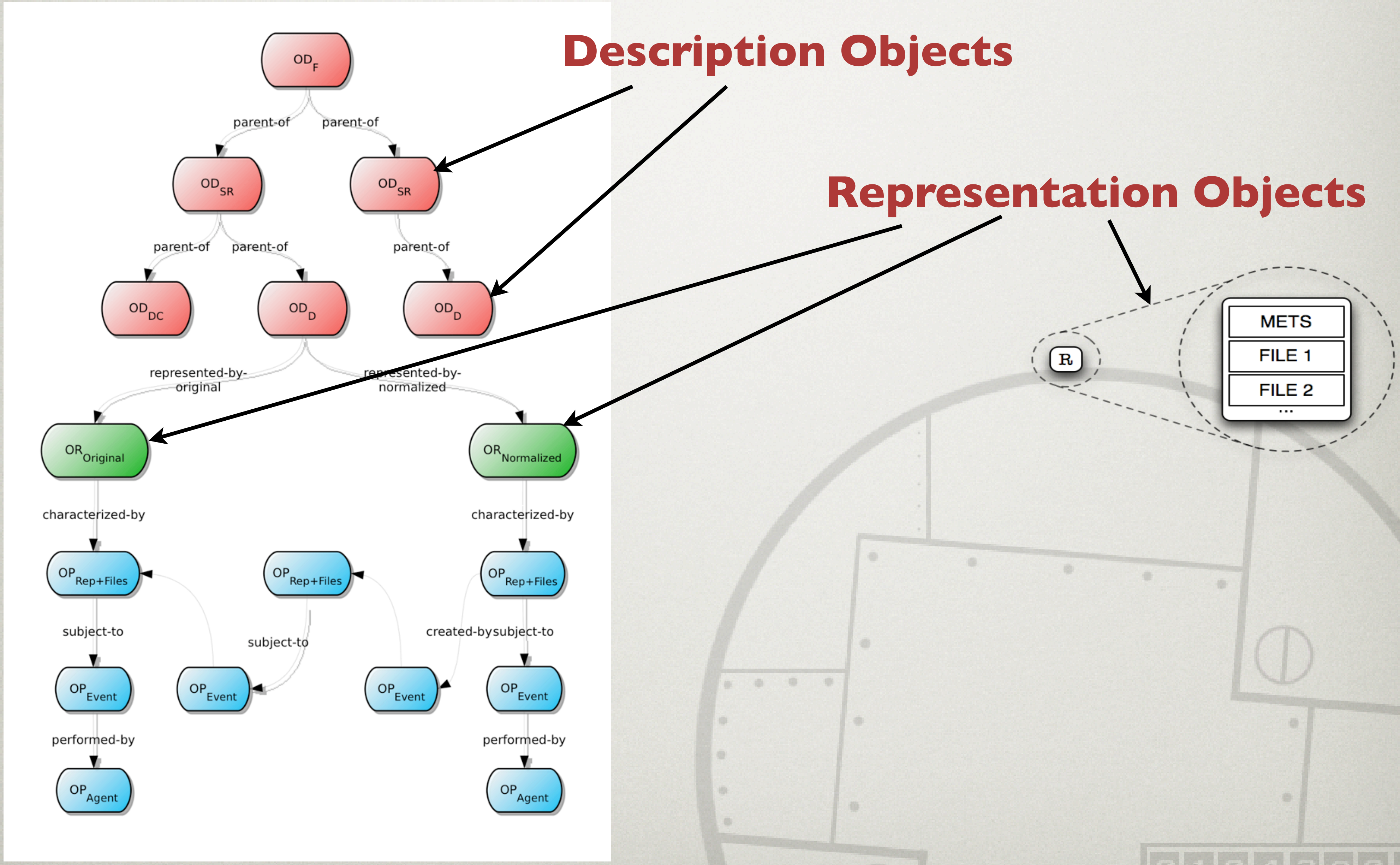


Roda Data Model



Description Objects

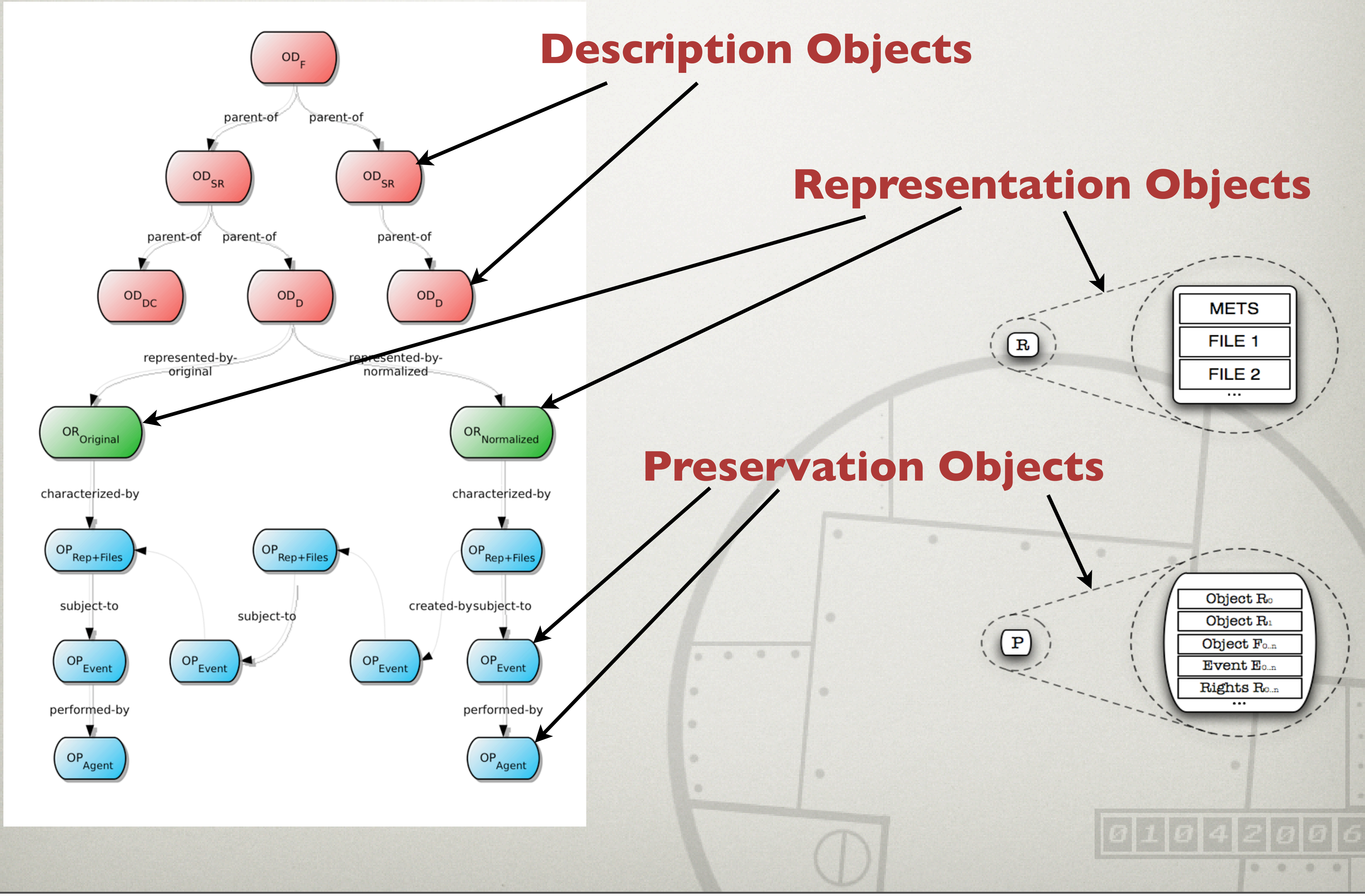
Roda Data Model



Description Objects

Representation Objects

Roda Data Model

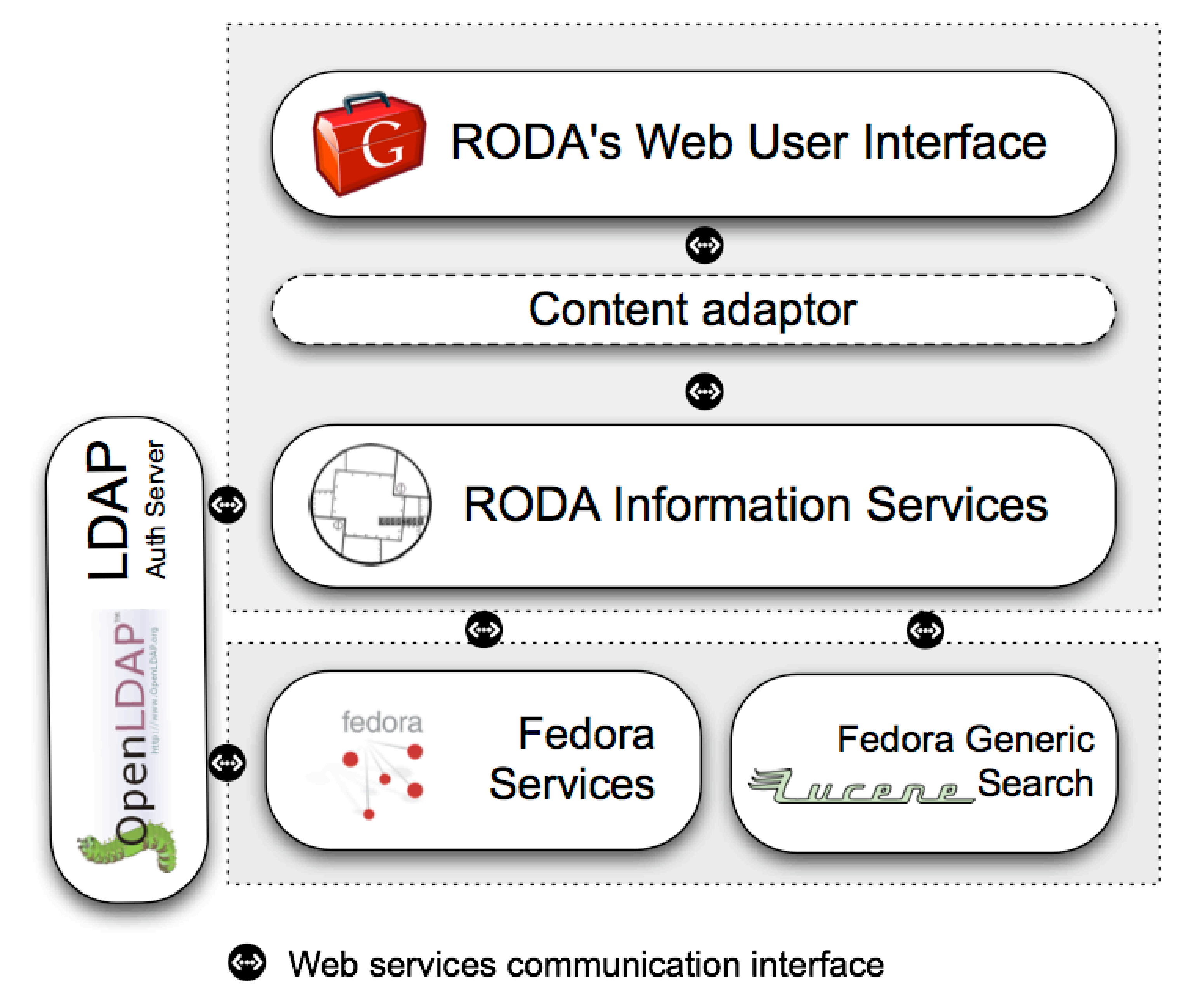


Description Objects

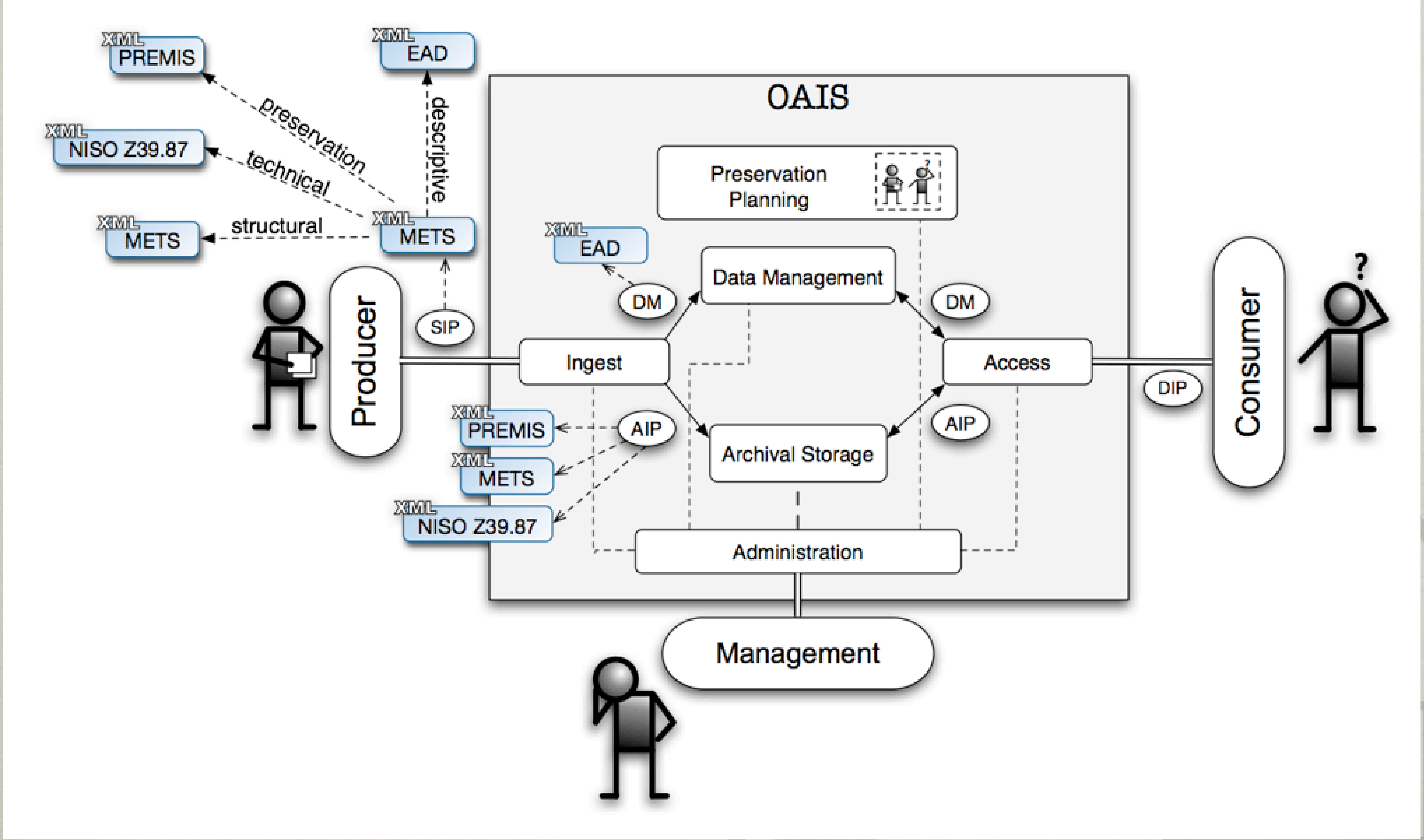
Representation Objects

Preservation Objects

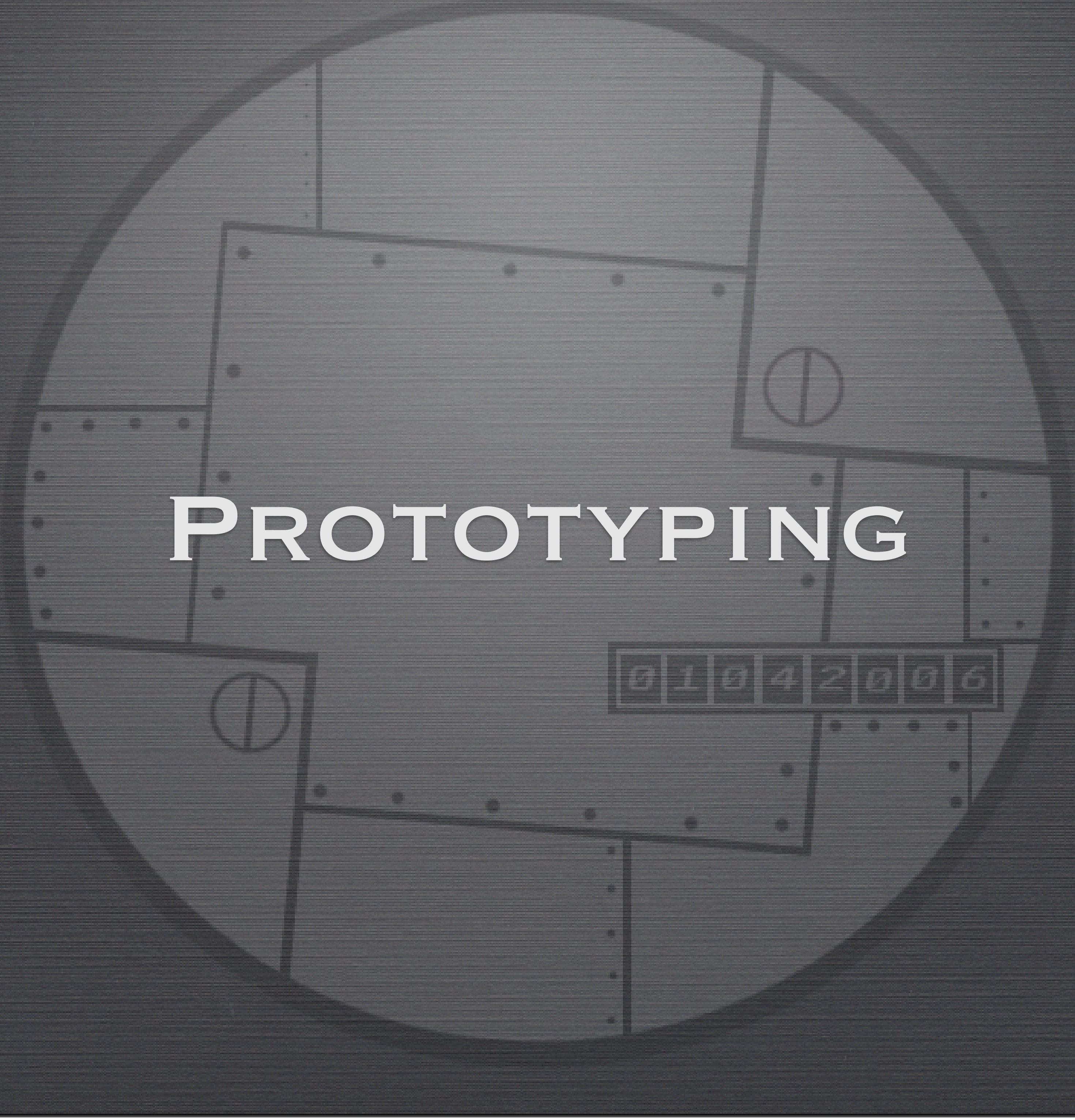
Architecture



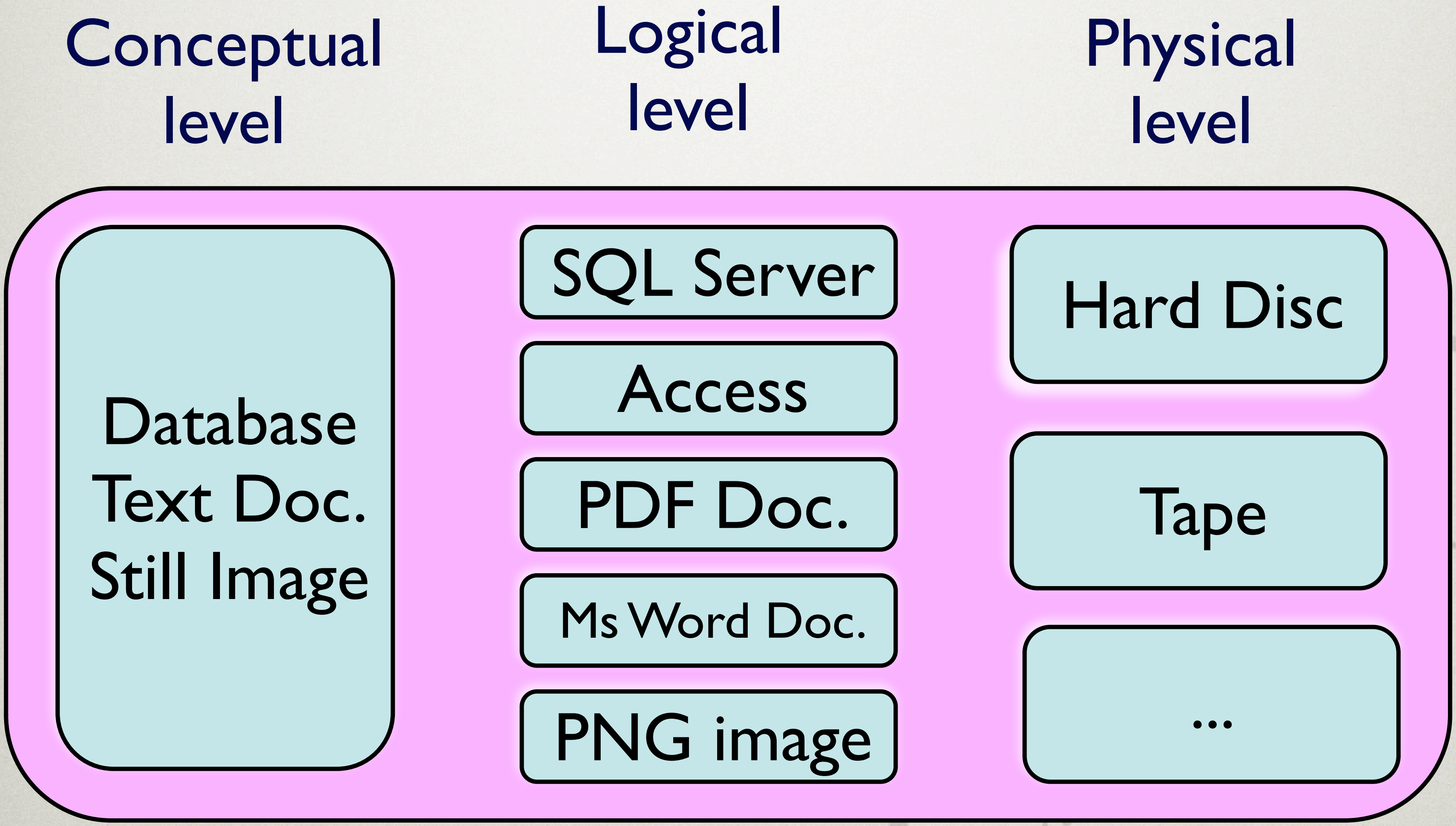
RODA Schemas



Prototyping



Preserving Conceptual Object



Conceptual level

Logical level

Physical level

Database
Text Doc.
Still Image

SQL Server

Access

PDF Doc.

Ms Word Doc.

PNG image

Hard Disc

Tape

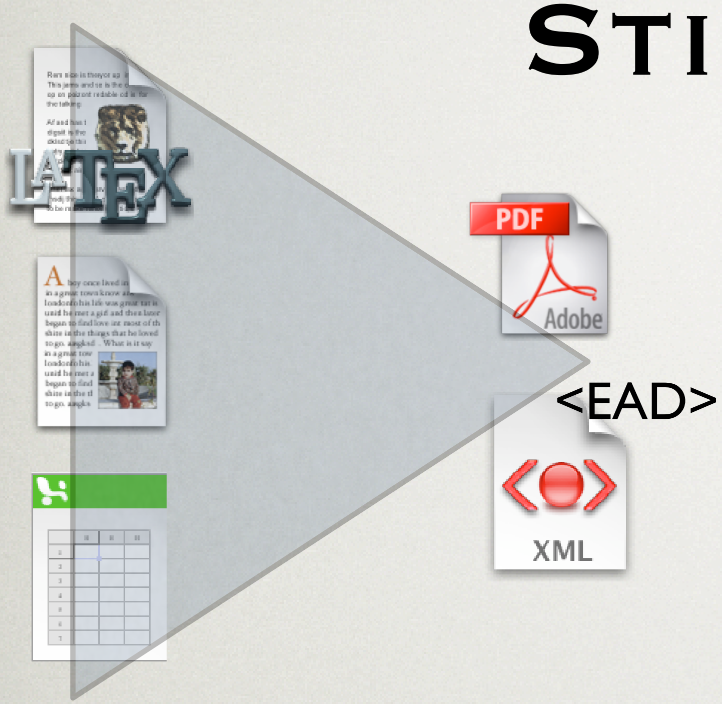
...

Text Documents and Still Images

Text Documents and Still Images

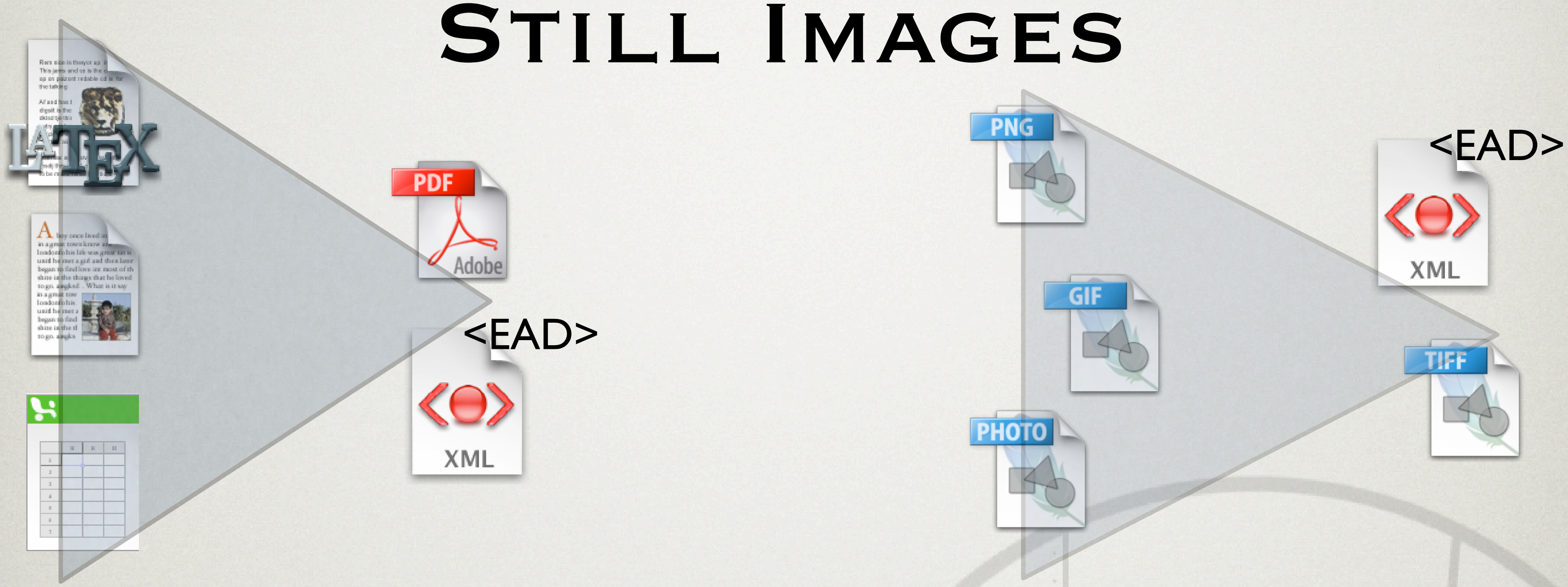
- EAD elements capture most of the significant properties: provenance, producer history, context, ...
- Content is kept in a normalized format: PDF and uncompressed TIFF.

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?
- ...

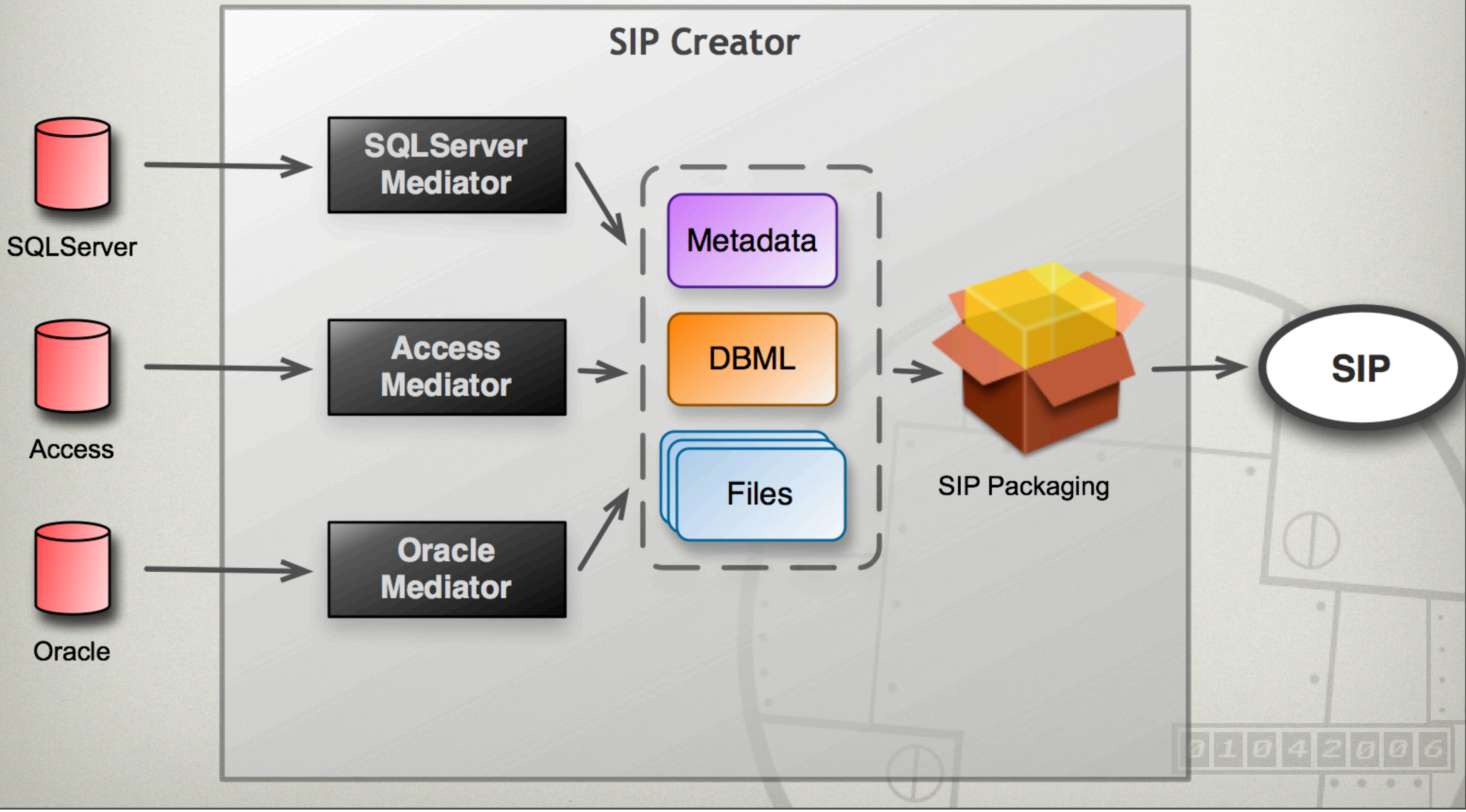
Databases

- 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

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
- BLOB
  standa
- Transf
  define

```
<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>
                      <KEY NAME="" REF=""/>
                      ...
                      </KEYS>
          </TABLE>
```

DBML

- P MS independent
- S
- B
  s
- Transf
  define

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

```
                <COLUMN NAME="code" TYPE="int" NULL="no"/>

UMNS>
>
TYPE="simple">
                <FIELD NAME=""/>
>
TYPE="compound">
                <FIELD NAME=""/>
                <FIELD NAME=""/>
          </PKEY>
          <KEY NAME="" REF=""/>
          ...
          </KEYS>
     </TABLE>
```

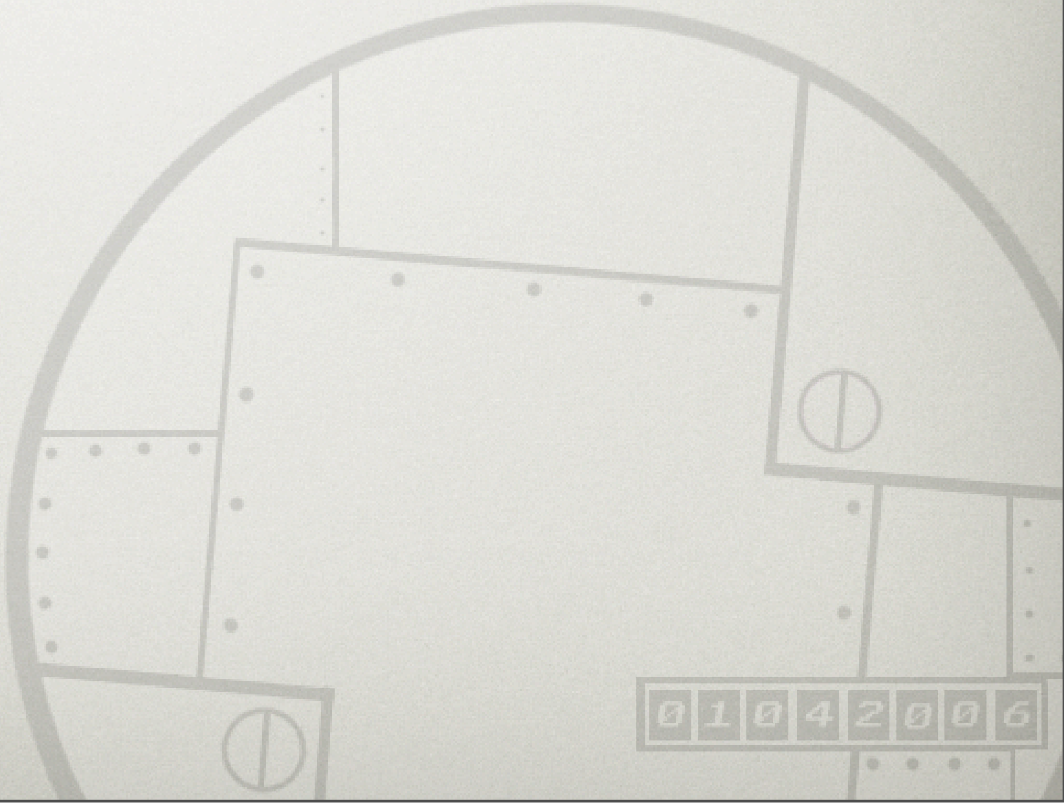
DBML

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

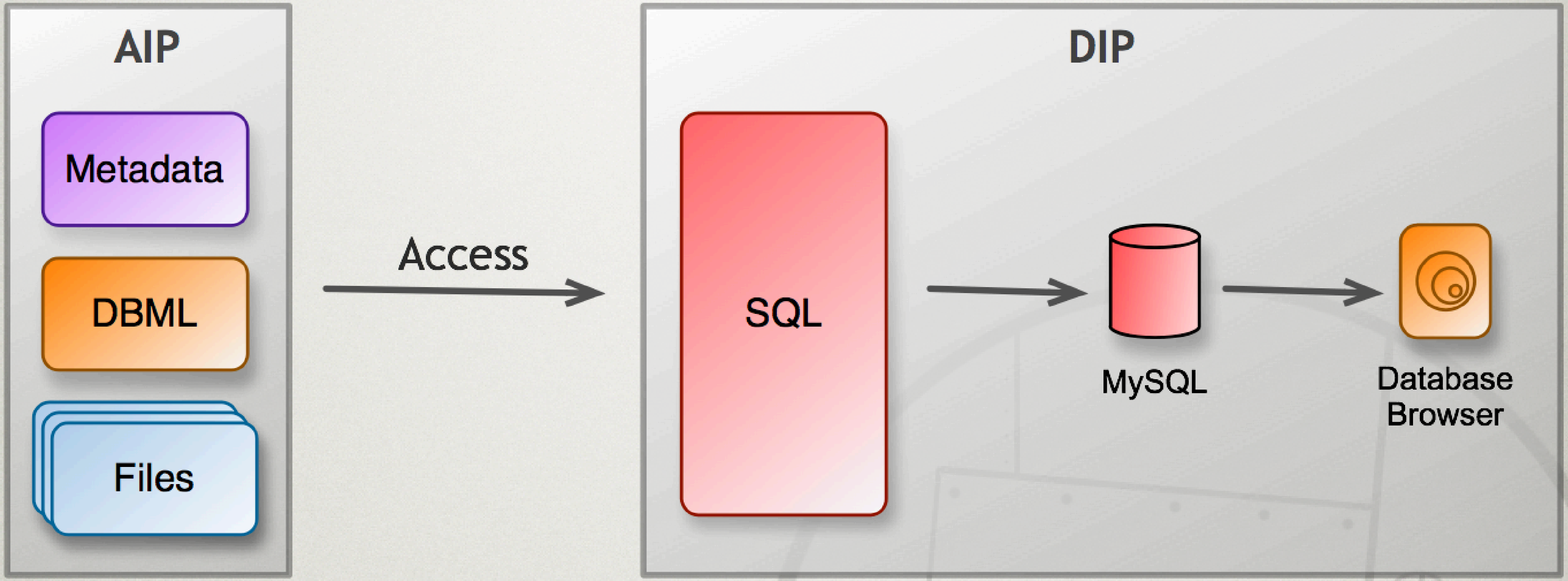
- Check and validation ...
- Generate SQL file
- Generate PREMIS



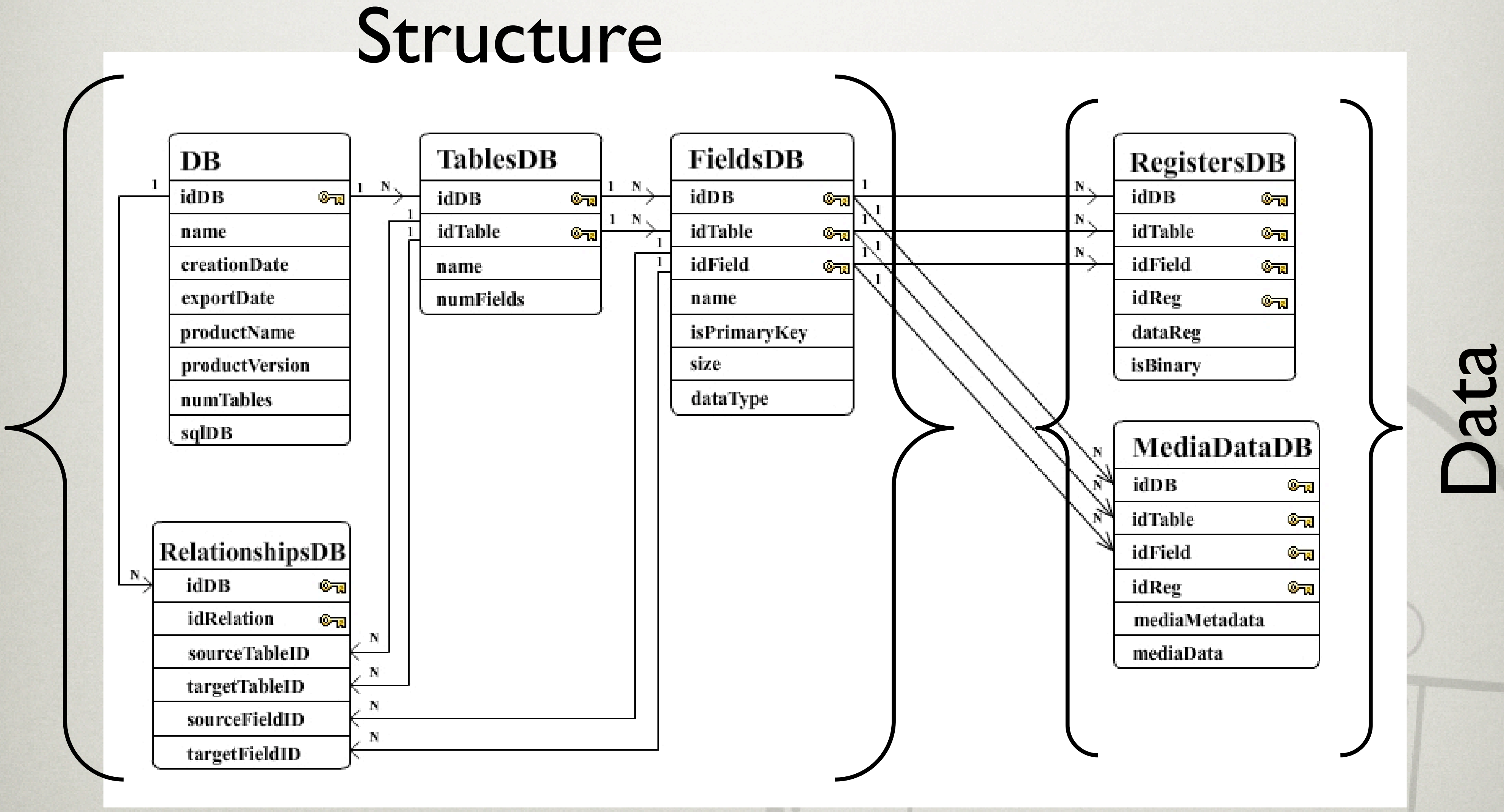
Dissemination

- 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



DB Abstract Schema



Structure

DB

- idDB
- name
- creationDate
- exportDate
- productName
- productVersion
- numTables
- sqlDB

TablesDB

- idDB
- idTable
- name
- numFields

FieldsDB

- idDB
- idTable
- idField
- name
- isPrimaryKey
- size
- dataType

RelationshipsDB

- idDB
- idRelation
- sourceTableID
- targetTableID
- sourceFieldID
- targetFieldID

Data

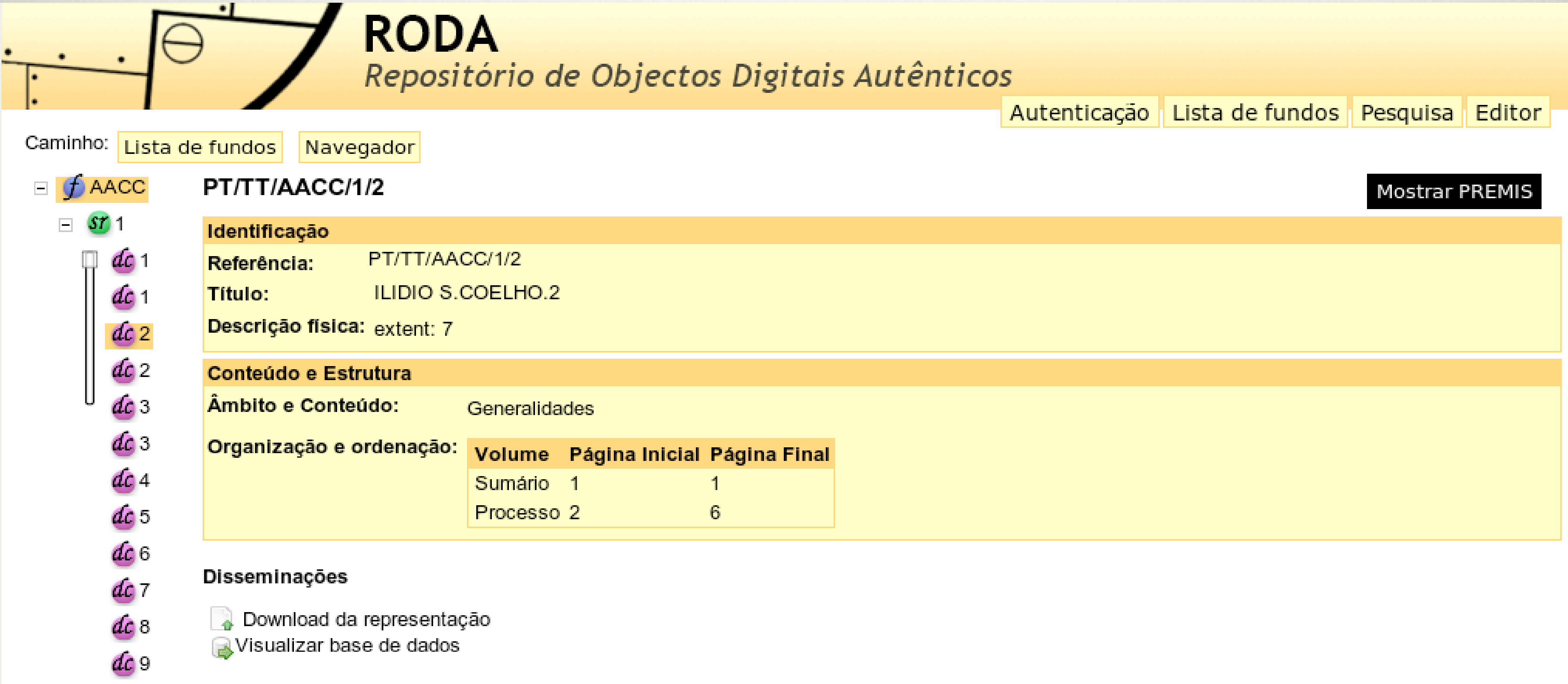
RegistersDB

- idDB
- idTable
- idField
- idReg
- dataReg
- isBinary

MediaDataDB

- idDB
- idTable
- idField
- idReg
- mediaMetadata
- mediaData

Browser



PT/TT/AACC/1/2

Identificação

Referência: PT/TT/AACC/1/2

Título: ILIDIO S.COELHO.2

Descrição física: extent: 7

Conteúdo e Estrutura

Âmbito e Conteúdo: Generalidades

Organização e ordenação:

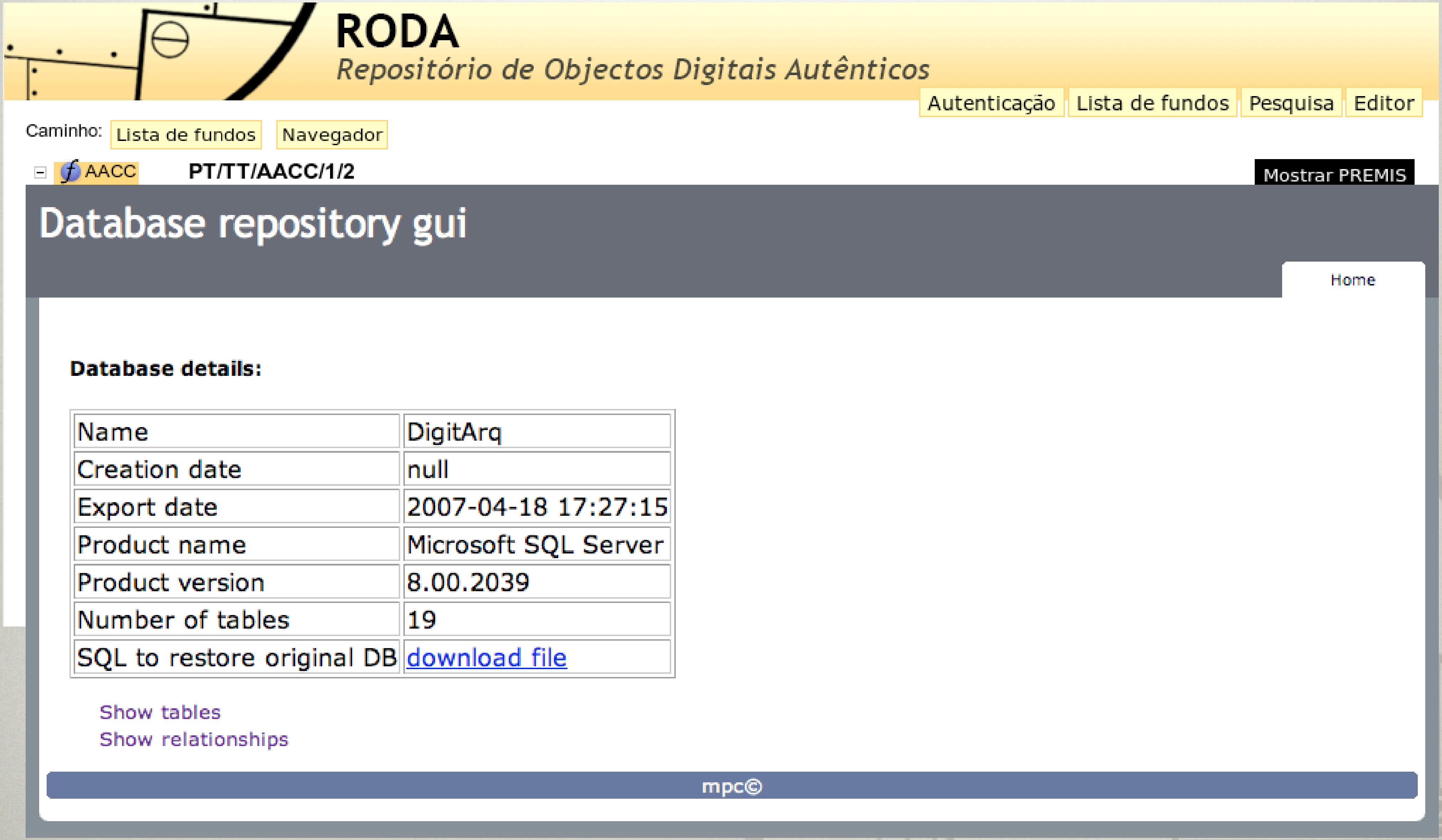
Volume	Página Inicial	Página Final
Sumário	1	1
Processo	2	6

Disseminações

Download da representação

Visualizar base de dados

Browser



PT/TT/AACC/1/2

Database repository gui

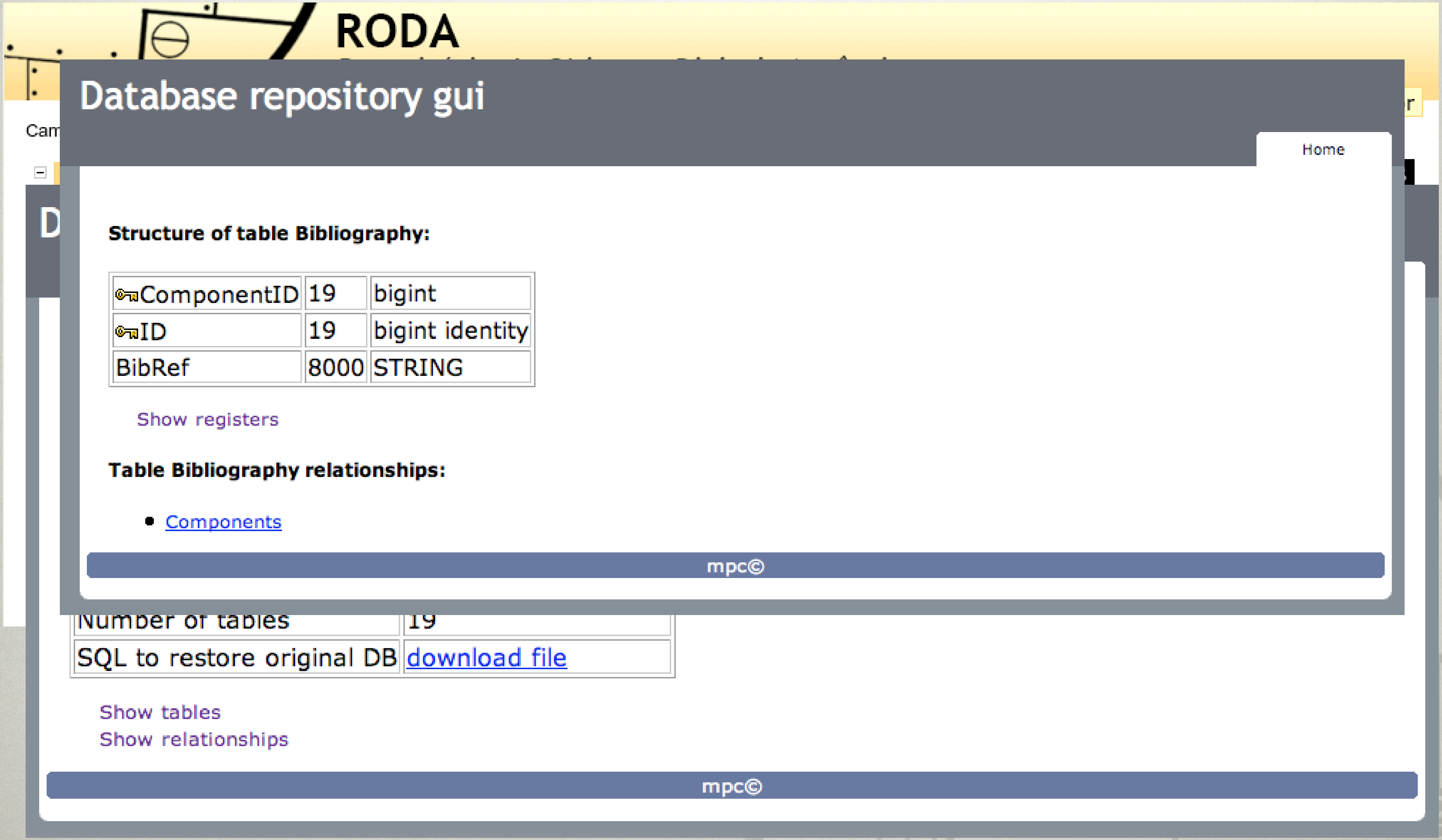
Database details:

Name	DigitArq
Creation date	null
Export date	2007-04-18 17:27:15
Product name	Microsoft SQL Server
Product version	8.00.2039
Number of tables	19
SQL to restore original DB	download file

Show tables
Show relationships

mpc©

Browser



Database repository gui

Structure of table Bibliography:

🔑ComponentID	19	bigint
🔑ID	19	bigint identity
BibRef	8000	STRING

Show registers

Table Bibliography relationships:

- Components

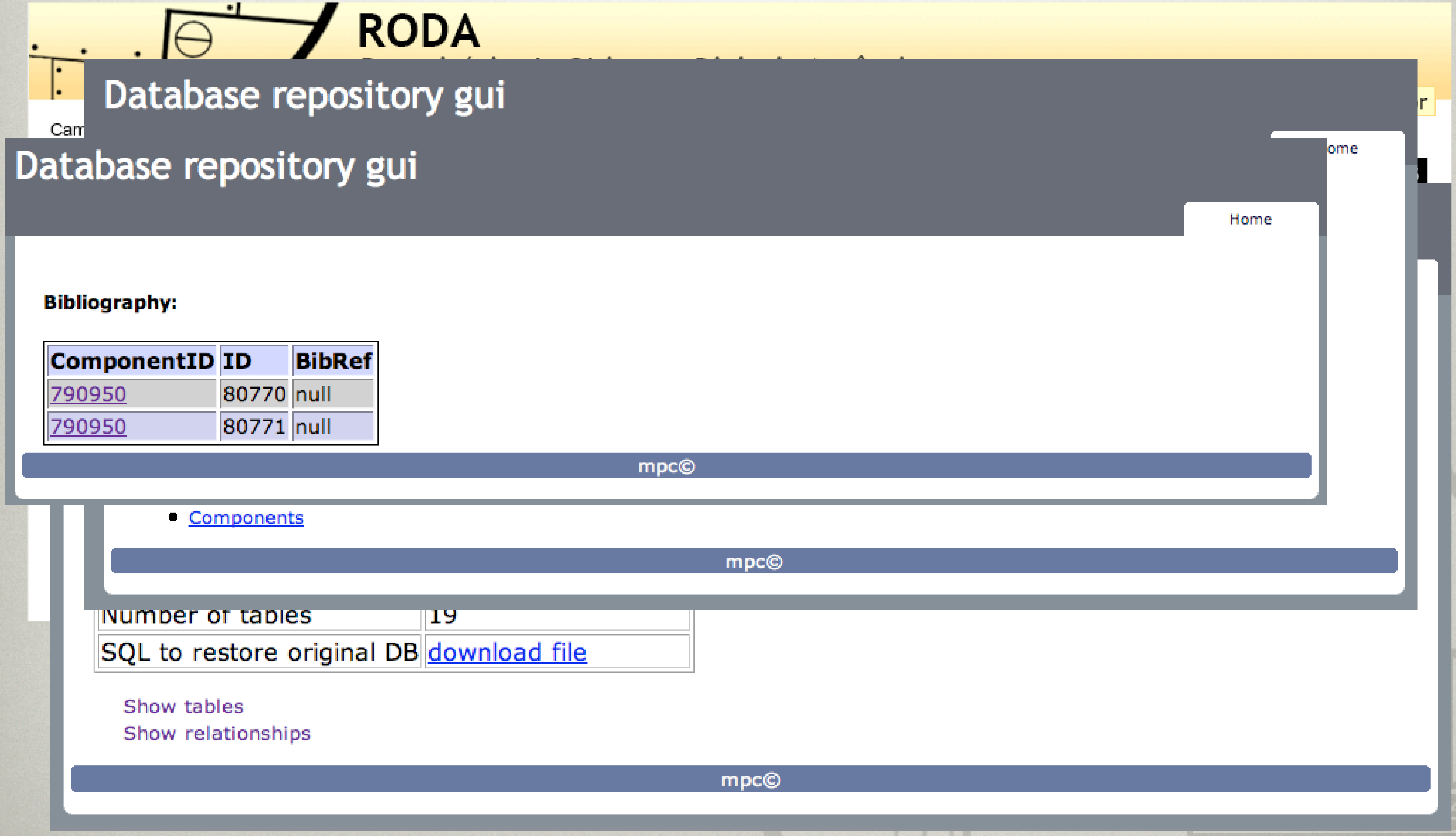
mpc©

Number of tables	19
SQL to restore original DB	download file

Show tables
Show relationships

mpc©

Browser



RODA

Cam

Database repository gui

ome

Database repository gui

Home

Bibliography:

ComponentID	ID	BibRef
790950	80770	null
790950	80771	null

mpc©

- Components

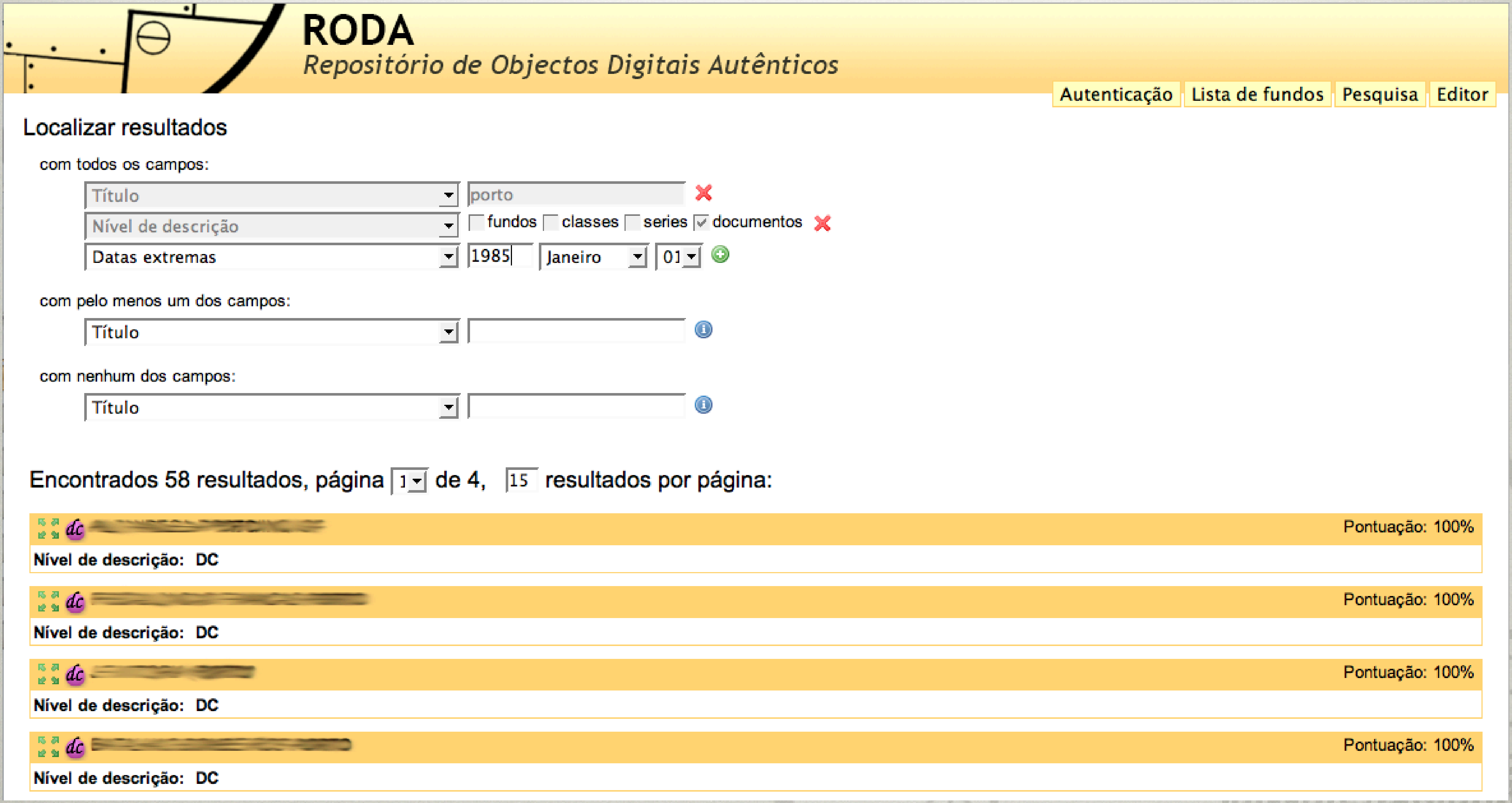
mpc©

Number of tables	19
SQL to restore original DB	download file

Show tables
Show relationships

mpc©

Search Engine



RODA

Repositório de Objectos Digitais Autênticos

Localizar resultados

com todos os campos:

Título ▾ porto ✖

Nível de descrição ▾ ☐ fundos ☐ classes ☐ series ☑ documentos ✖

Datas extremas ▾ 1985 Janeiro ▾ 01 ▾ ⊕

com pelo menos um dos campos:

Título ▾ ⓘ

com nenhum dos campos:

Título ▾ ⓘ

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

dc [redacted] Pontuação: 100%

Nível de descrição: DC

dc [redacted] Pontuação: 100%

Nível de descrição: DC

dc [redacted] Pontuação: 100%

Nível de descrição: DC

dc [redacted] Pontuação: 100%

Nível de descrição: DC

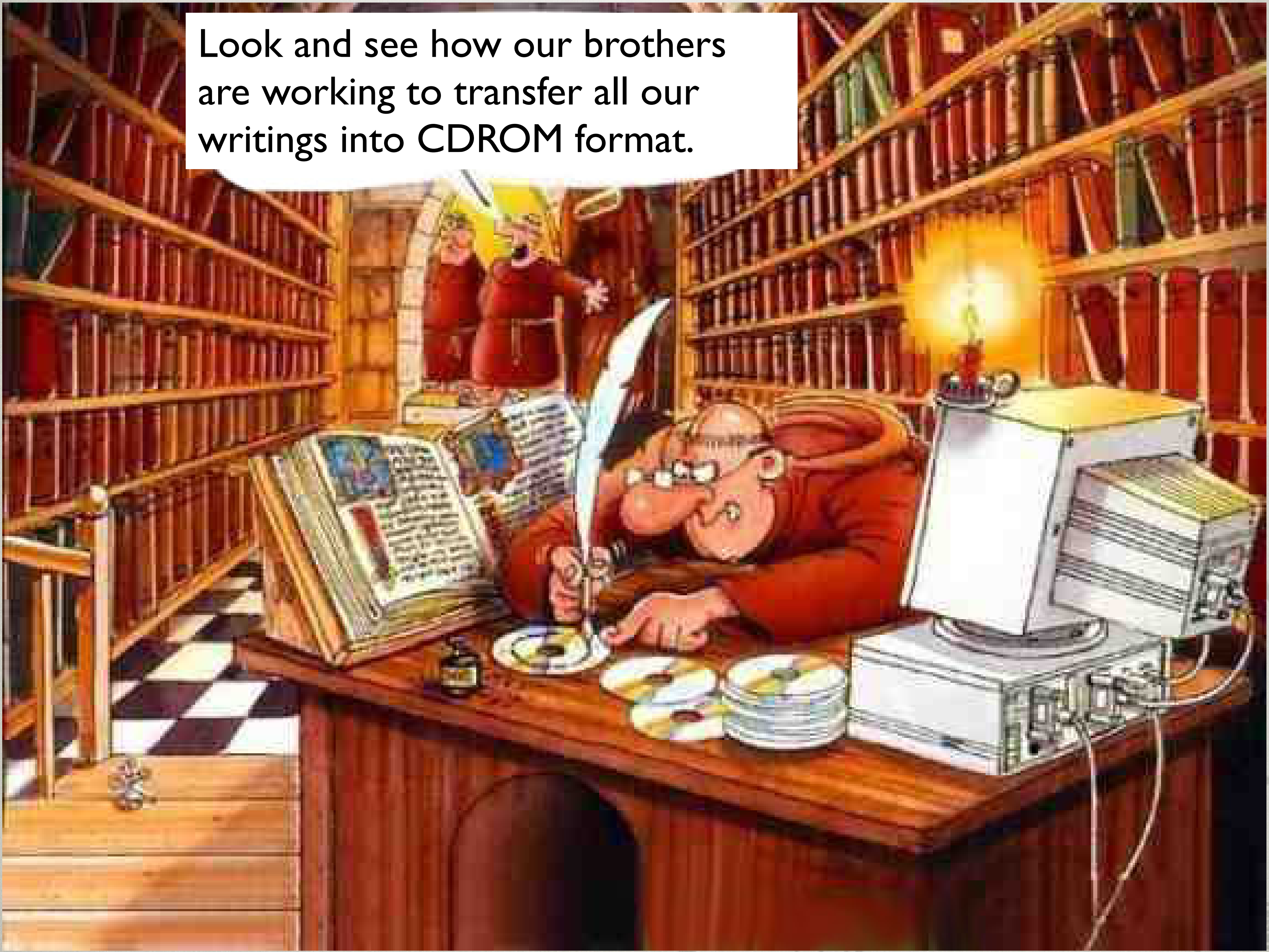
Final thoughts

"Data Preservation is a people problem"
Michael Lesk

Final thoughts

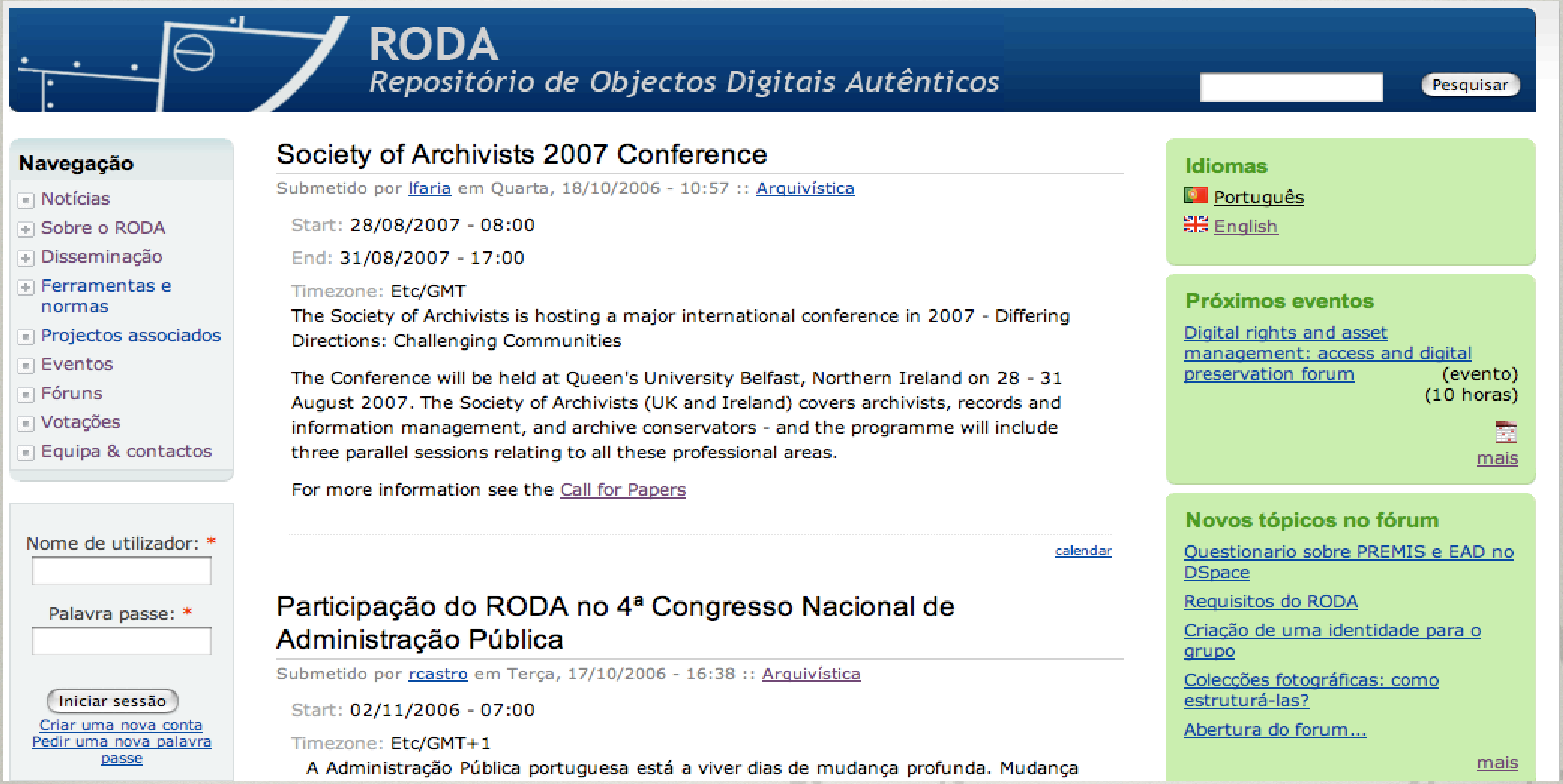
“Data Preservation is a people problem”
Michael Lesk

- 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.

RODA Homepage



RODA

Repositório de Objectos Digitais Autênticos

Navegação

- Notícias
- Sobre o RODA
- Disseminação
- Ferramentas e normas
- Projectos associados
- Eventos
- Fóruns
- Votações
- Equipa & contactos

Nome de utilizador: *

Palavra passe: *

Criar uma nova conta

Pedir uma nova palavra passe

Society of Archivists 2007 Conference

Submetido por lfaria 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

calendar

Participação do RODA no 4ª Congresso Nacional de Administração Pública

Submetido por rcastro 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

Idiomas

Português

English

Próximos eventos

Digital rights and asset management: access and digital preservation forum (evento) (10 horas)

mais

Novos tópicos no fórum

Questionario sobre PREMIS e EAD no DSpace

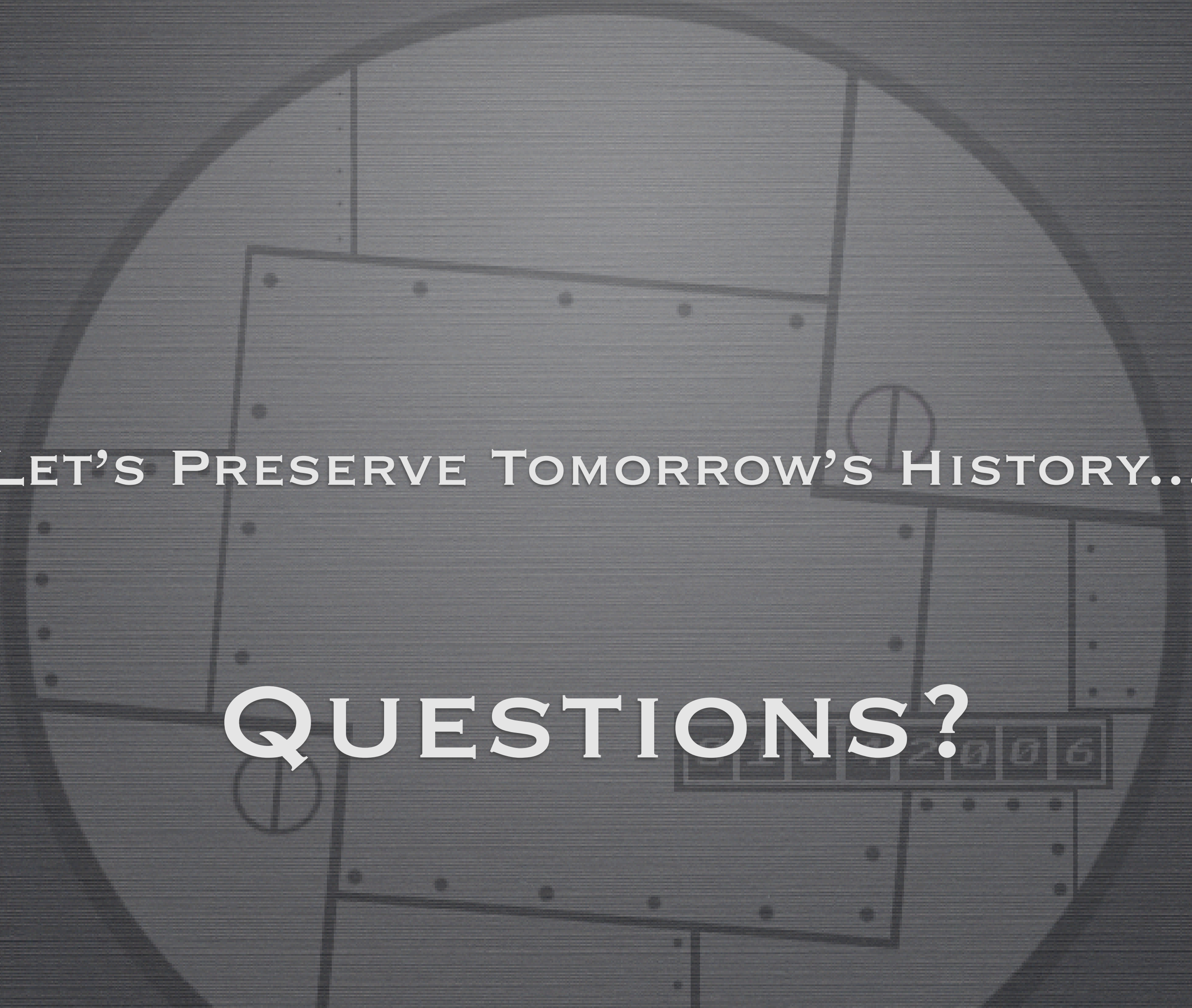
Requisitos do RODA

Criação de uma identidade para o grupo

Colecções fotográficas: como estruturá-las?

Abertura do forum...

mais



Let's Preserve Tomorrow's History...

Questions?