

RODA: preservação digital na administração pública portuguesa

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

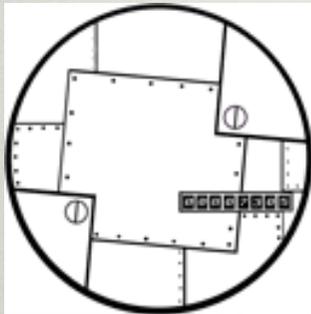
Luis Corujo
lcorujo@iantt.pt

CONTEXT



Digitarq (2003-now)

- metadata management (EAD based)
- digital object management (NISO MIX)



RODA (2006-2009)

- metadata management (EAD based)
- digital object management (...)
- digital preservation protocols and policies

CRAV: Readers Virtual Room (2006-2007)

- request management
- document workflow

RODA: MOTIVAÇÃO

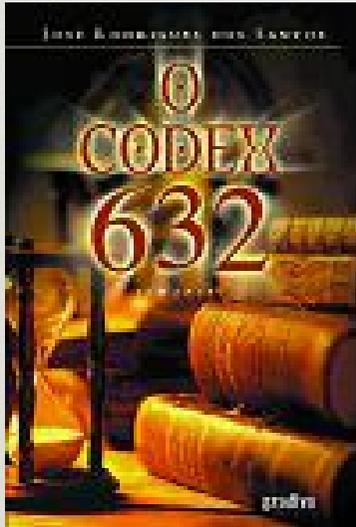
- Hoje a história é digital;
- A produção de objectos digitais cresce a cada dia: desmaterialização na administração pública;
- Não há estruturas capazes de suportar a incorporação, a gestão e a preservação a longo prazo dos objectos digitais;
- É necessário preservar a memória digital, a herança e o testemunho das instituições.
 - Exemplo: SGU, processos judiciais, registos paroquiais, etc.

REQUISITOS

- Como garantir a autenticidade?
- Como descrever e classificar ODs?
- Como implementar a preservação digital?
- ...
- Como manter a solução depois de desenvolvida?



AUTENTICIDADE



“O Codex 632” de José Rodrigues dos Santos

Assunto: Quem foi Cristovão Colombo?

Um italiano? Espanhol? Ou um português de uma família judaica?

AUTENTICIDADE

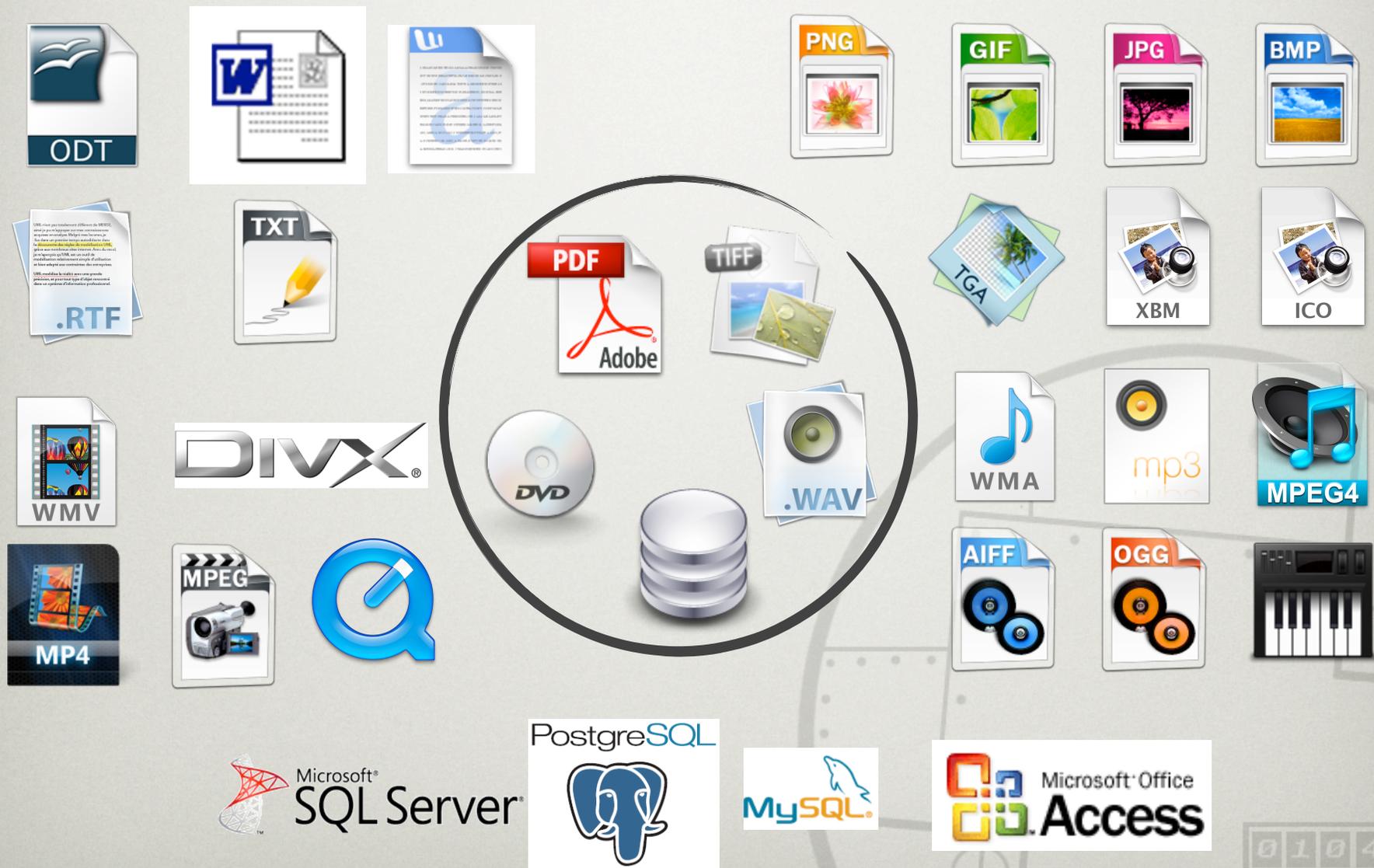
Temos de confiar nas nossas fontes: na História Antiga não há discurso directo ou evidências.

EX: a bíblia

Como nos tornamos dignos de confiança?

- Reputação
- Documentando todo o ciclo de vida de um OD
- Implementando o TRAC (Trustworthy Repositories Audit & Certification)

FORMATOS SUPORTADOS



Anatomia de um OD

Nível
Conceptual

Nível
Lógico

Nível
Físico

Se um destes níveis se tornar obsoleto deixamos de ter acesso ao OD

B.Dados
Doc.Texto
Imagem

PDF Doc.

Ms Word Doc.

PNG image

Fita

...



Estratégias de Preservação

- Focada no **objecto físico**
 - o Centrada na preservação da informação no seu **formato lógico e/ou no suporte físico**;
 - o Usa a tecnologia original associada ao objecto para assegurar o acesso ao mesmo;
 - o **Preservação tecnológica.**
- Focada no **objecto conceptual**
 - o Centrada na preservação das propriedades do objectos de **forma independente** do hardware e software;
 - o **Preservação do objecto conceptual.**

Emulação

Emulador: aplicação capaz de reproduzir o comportamento de uma plataforma de hardware/software: ZX Spectrum, GBA, ...

- Vantagens
 - o Reprodução do contexto tecnológico original;
 - o Preservação do *look & feel* do objecto.
- Desvantagens
 - o Os emuladores também se tornam obsoletos;
 - o Os utilizadores têm de trabalhar com sistemas obsoletos;
 - o Criar um emulador é uma tarefa complexa;
 - o Problemas de Copyright;
 - o Preservar todo um sistema para poder visualizar um documento pode ser demasiado!
 - o A reutilização da informação não é garantida.

Encapsulamento

Preservação da **bit stream original** juntamente com metainformação suficiente para assegurar a sua interpretação e acesso no futuro

- Vantagens
 - Permite adiar as responsabilidades da preservação;
 - Deve ser utilizada em objectos cujo acesso será desejado num futuro a longo prazo;
- Desvantagens
 - **Objectos complexos têm especificações complexas;**
 - Uma **especificação incompleta** pode ter efeitos indesejados.

Preservação do Objecto Conceptual

Migração: transferência periódica do OD de uma configuração de hw/sw para outra mais actualizada (preservando as propriedades significativas do objecto em lugar da sequência de bits original).

Vantagens

- Os ODs são disseminados em formatos familiares aos utilizadores;
- Não há necessidade de manter a plataforma original de hw/sw;
- A estratégia mais usada e a única que funcionou até hoje.

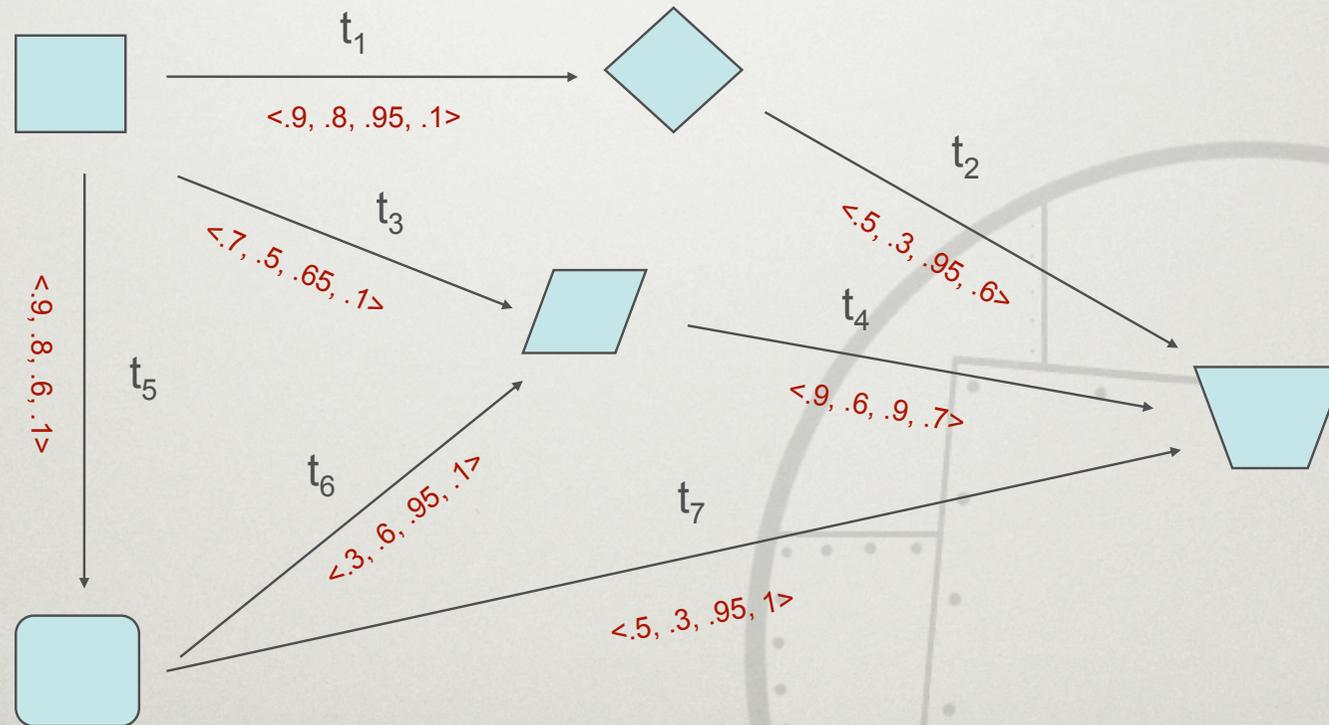
Desvantagens

- Possível perda de informação nas conversões;
- Manutenção constante e continuada;
- **A longo prazo os custos poderão ser altos.**

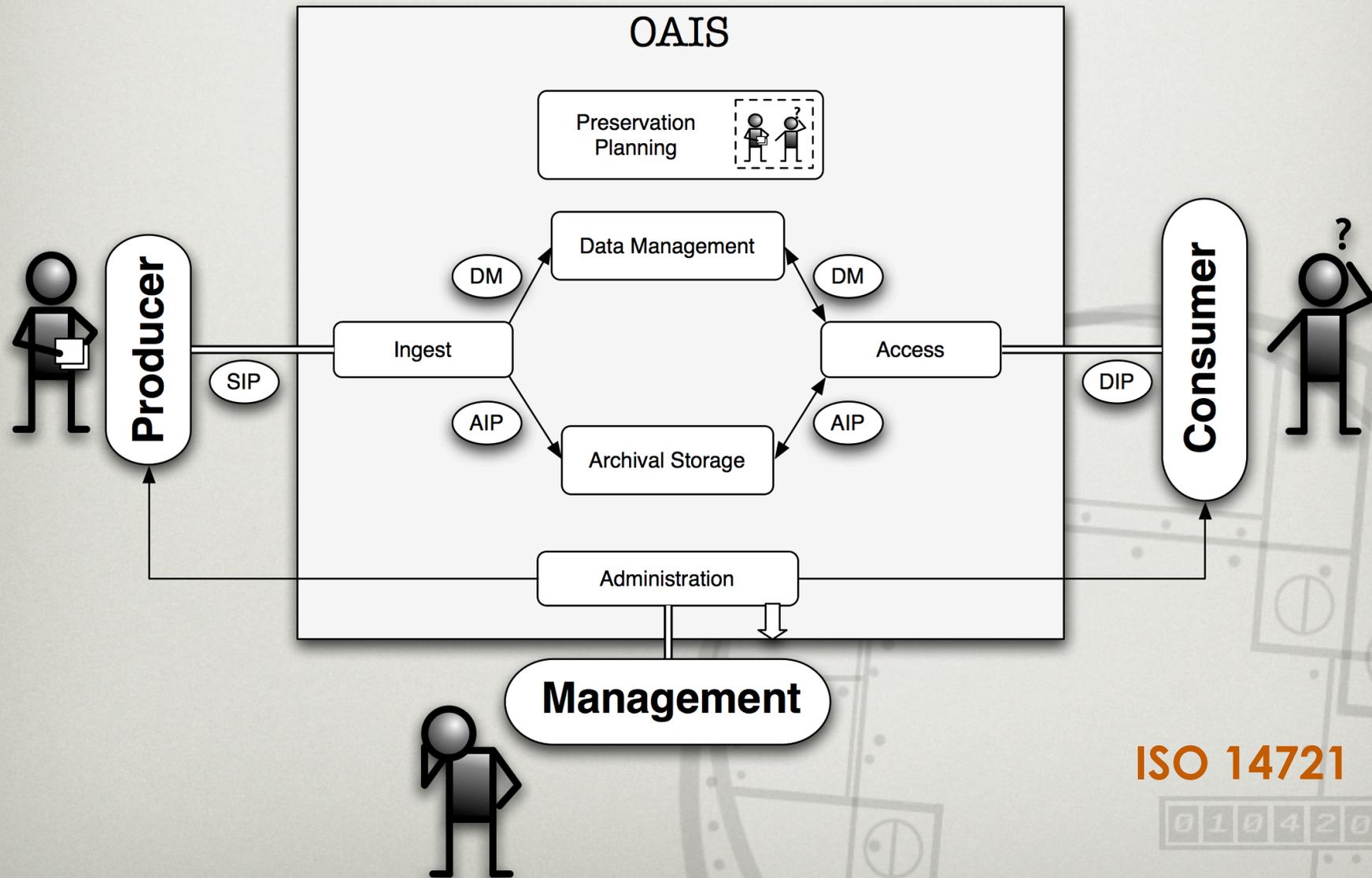
O que são propriedades significativas?

Serviços de Preservação

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



OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)



ISO 14721

OAIS (COMPONENTES FUNCIONAIS)

❖ Ingestão

- **Recepção, validação, transformação/normalização**, descrição do pacote enviado pelo produtor;

❖ Armazenamento

- Assegura a preservação da informação ao nível físico/lógico (e.g. refrescamento, migração, verificação de integridade, tolerância a falhas, etc.)

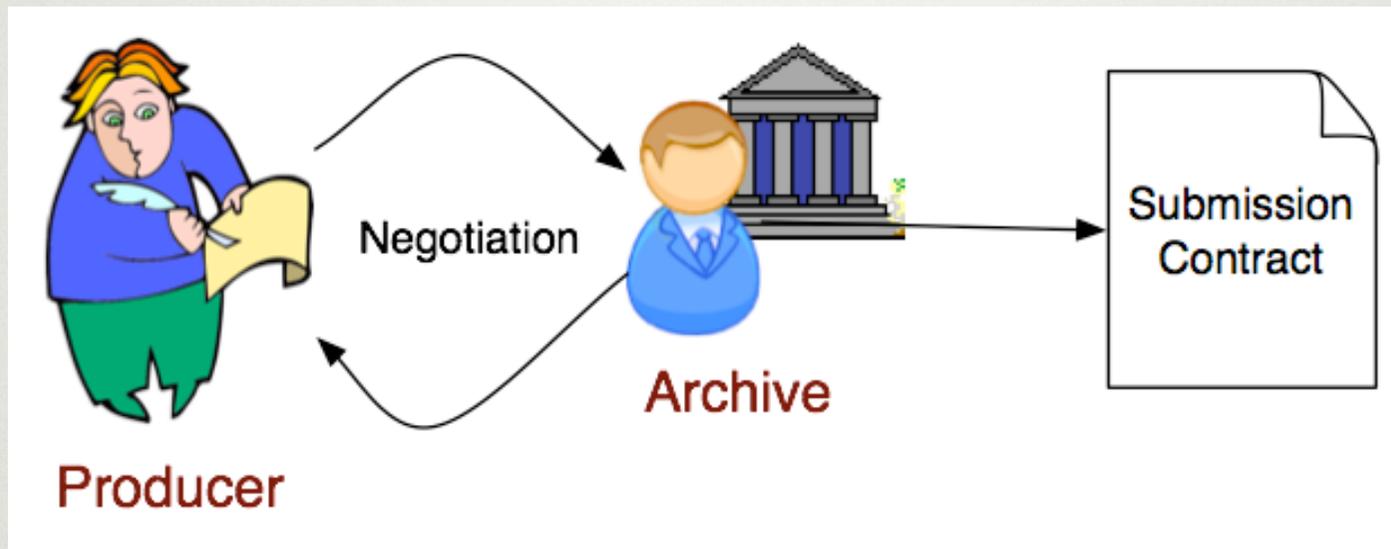
❖ Gestão de metainformação

- Responsável pela gestão dos ODs armazenados

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**

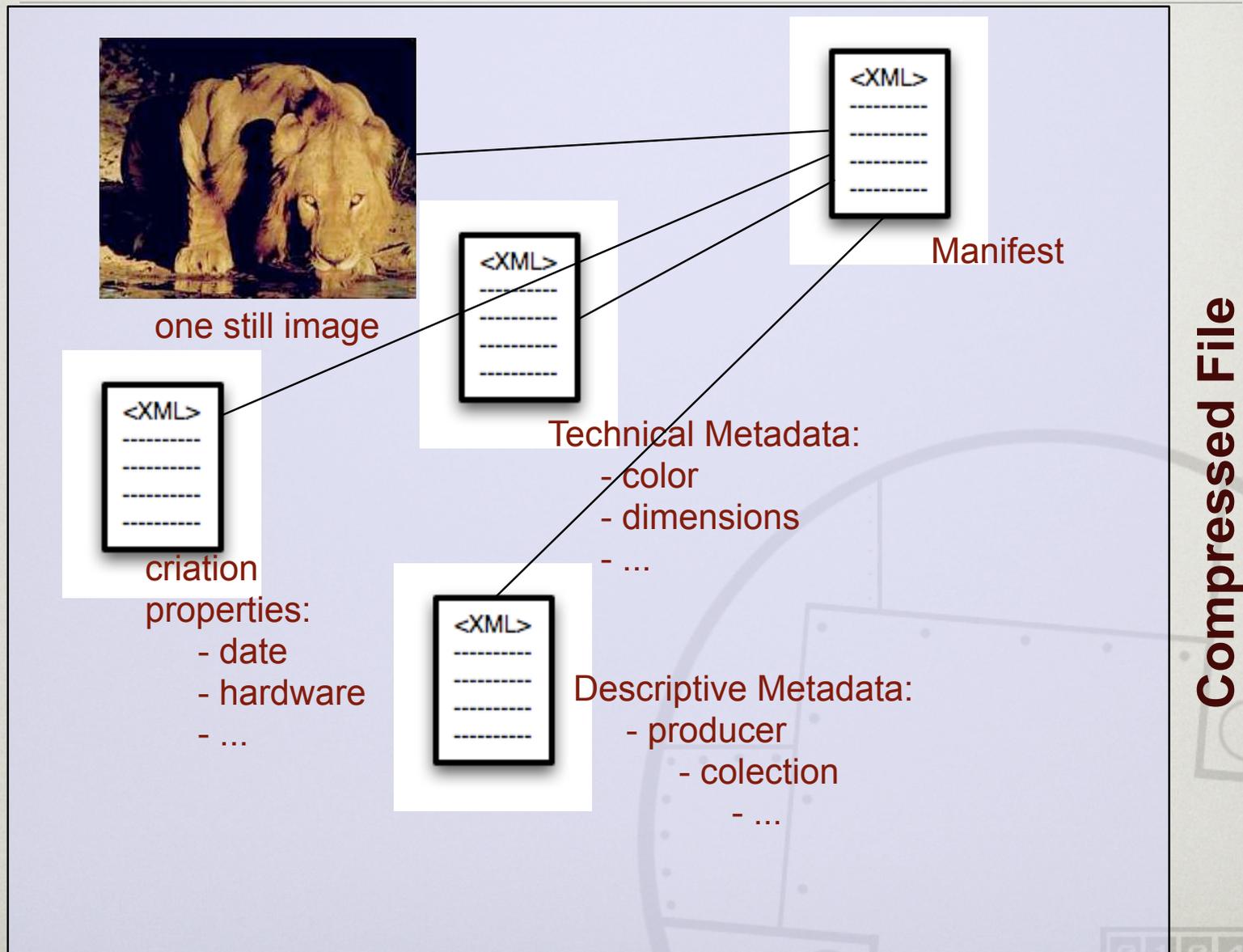
INGESTION



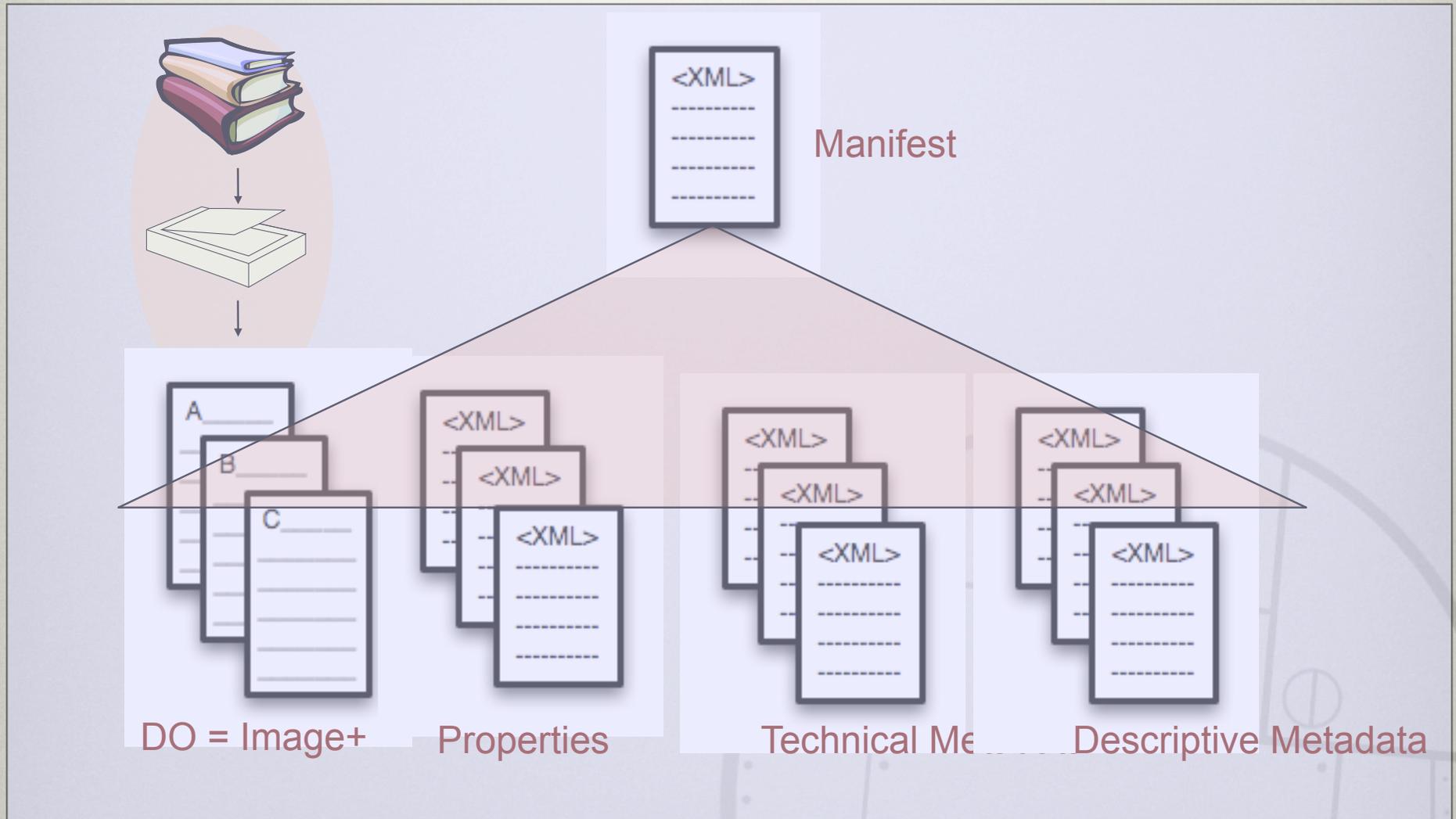
Submission Contract

- SIP specification
- Ingestion workflow specification

SIP STRUCTURE (EXAMPLE)



SIP STRUCTURE (+COMPLEX)



Compressed File

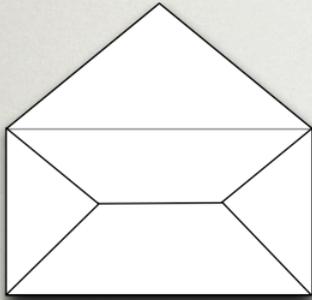
01042006

INGESTION WORKFLOW

SIP

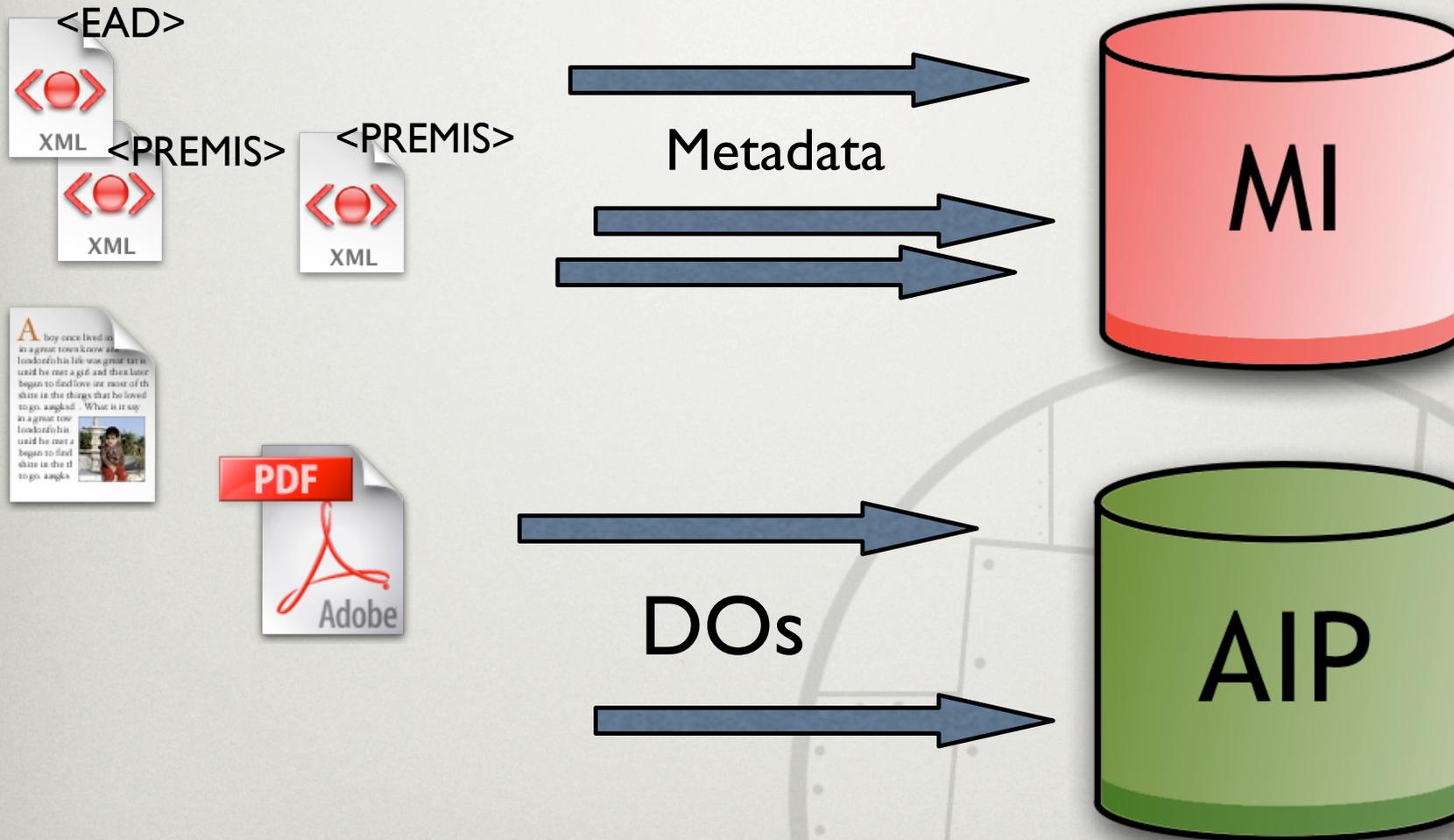


AIP

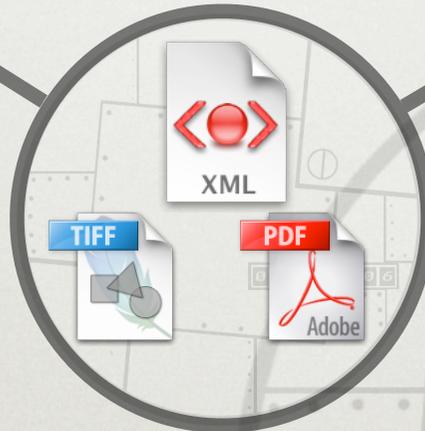
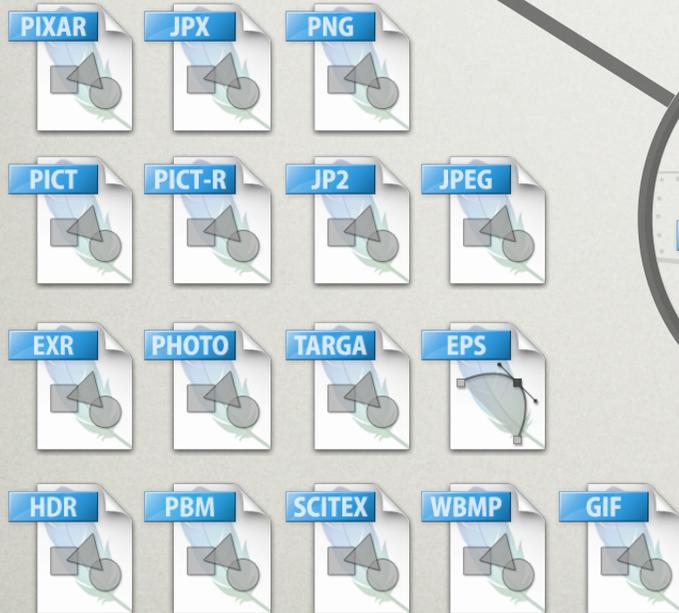
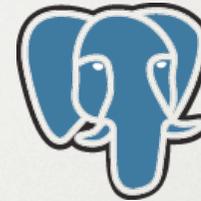


- integrity check
- virus check
- generation of preservation metadata (PREMIS)
- conversion to a normalized format
- generation of technical metadata
- generation of preservation metadata (PREMIS)

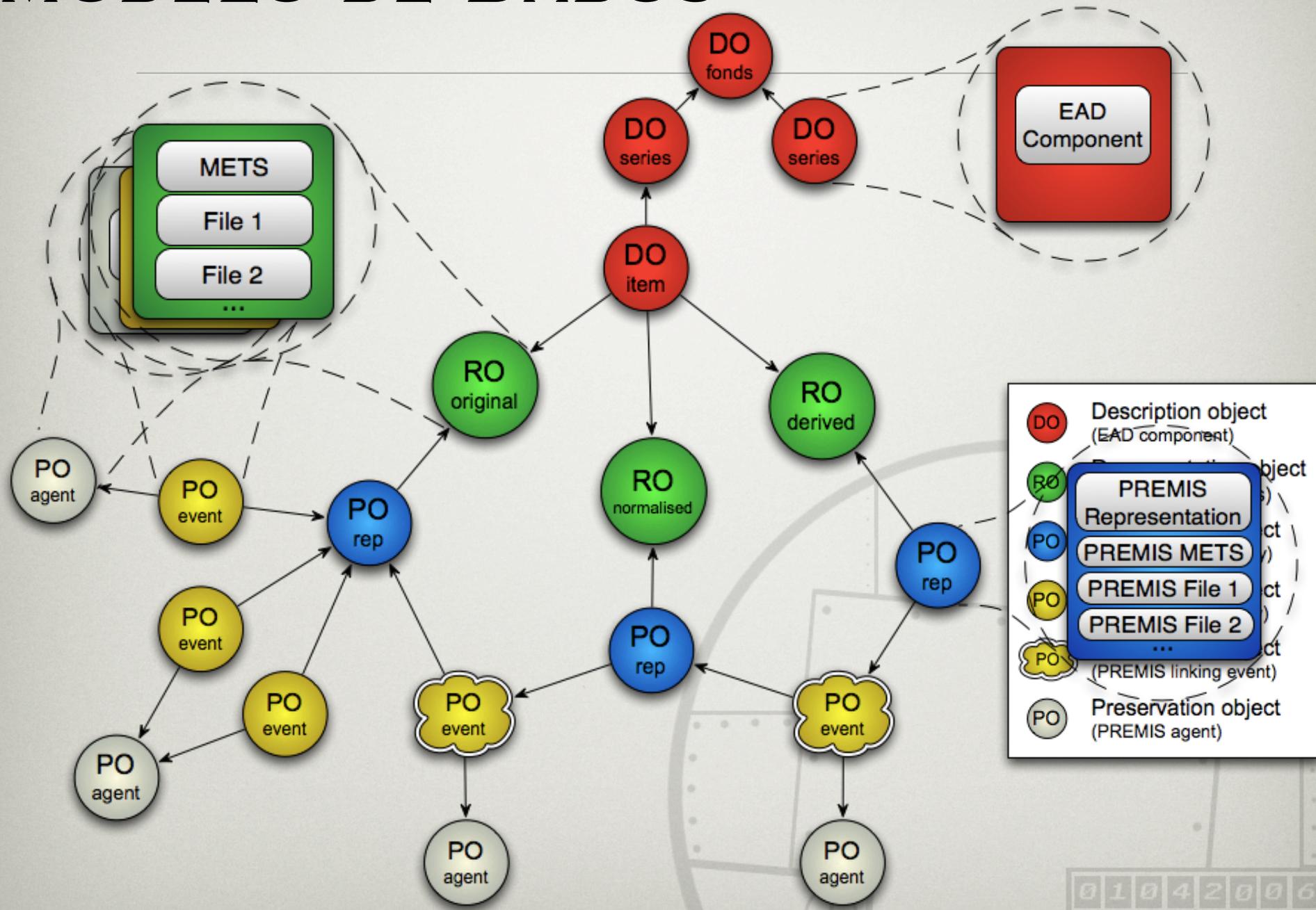
AIP STORAGE



NORMALIZATION

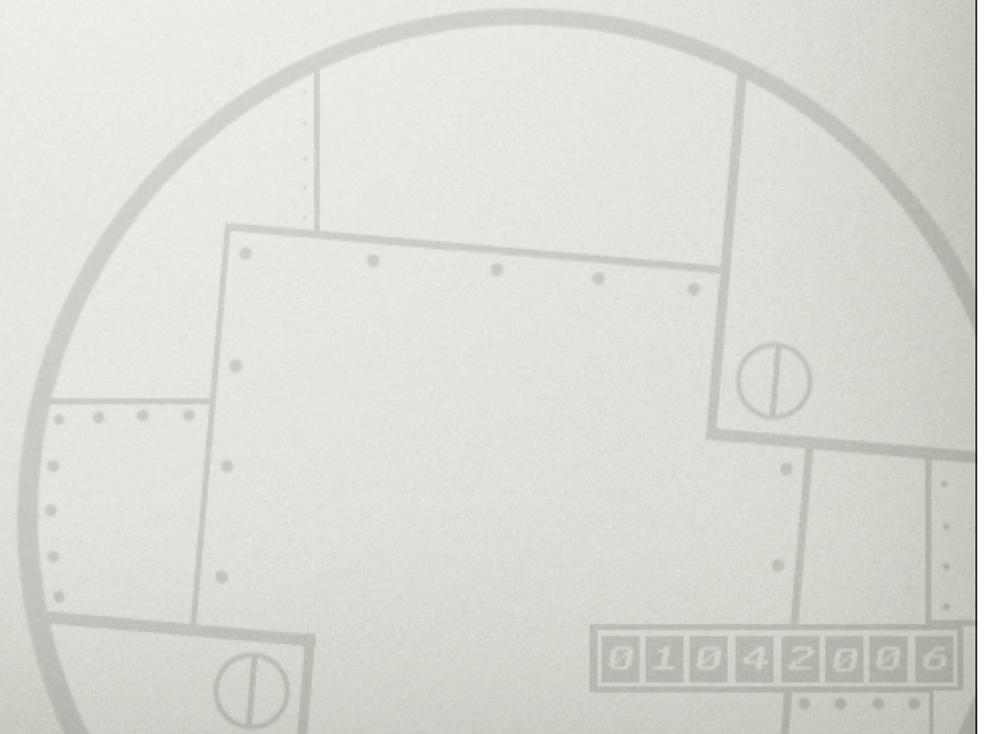


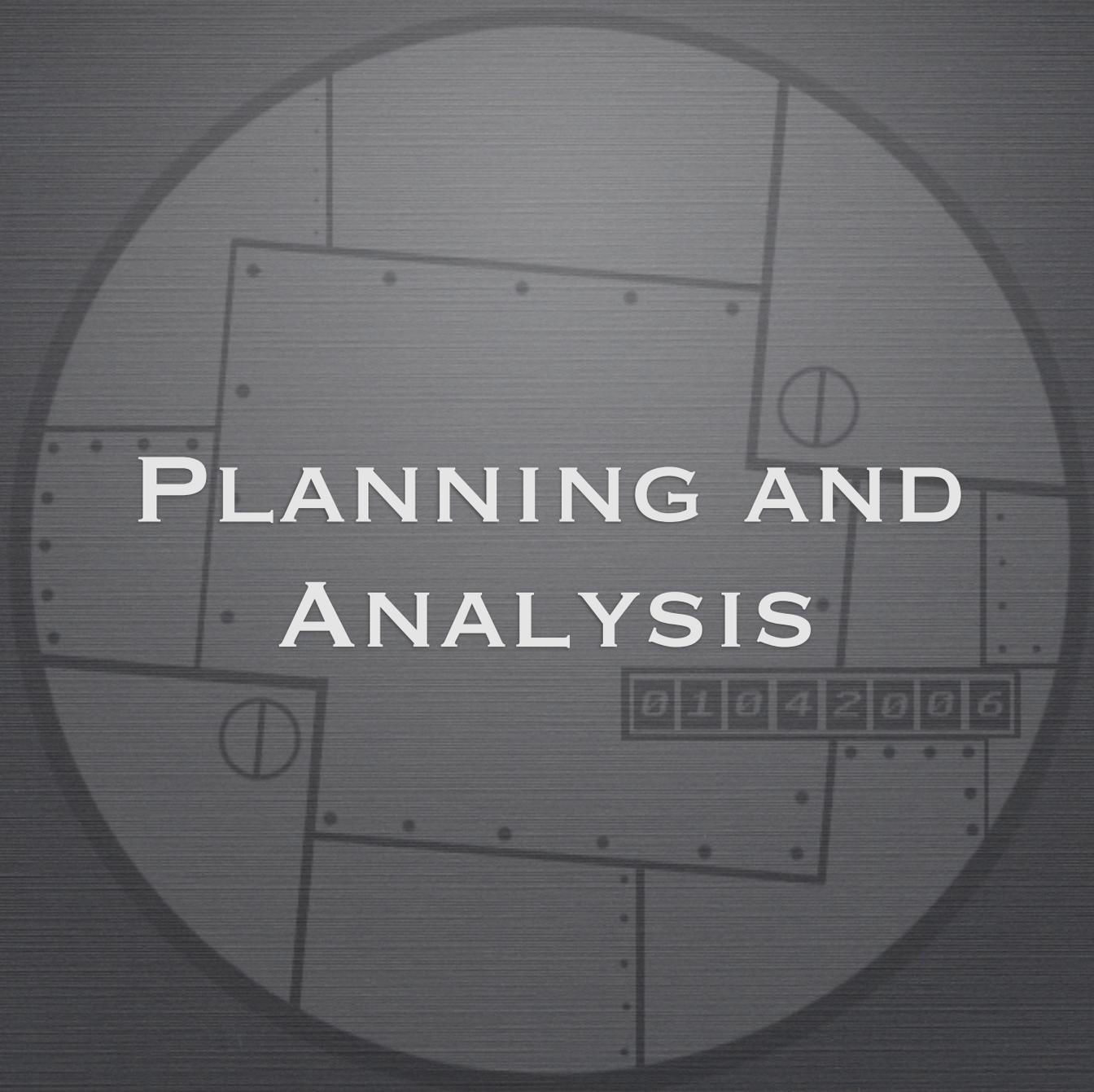
MODELO DE DADOS



STAGES

- Analysis and Planning
- Prototyping
- Testing and Dissemination

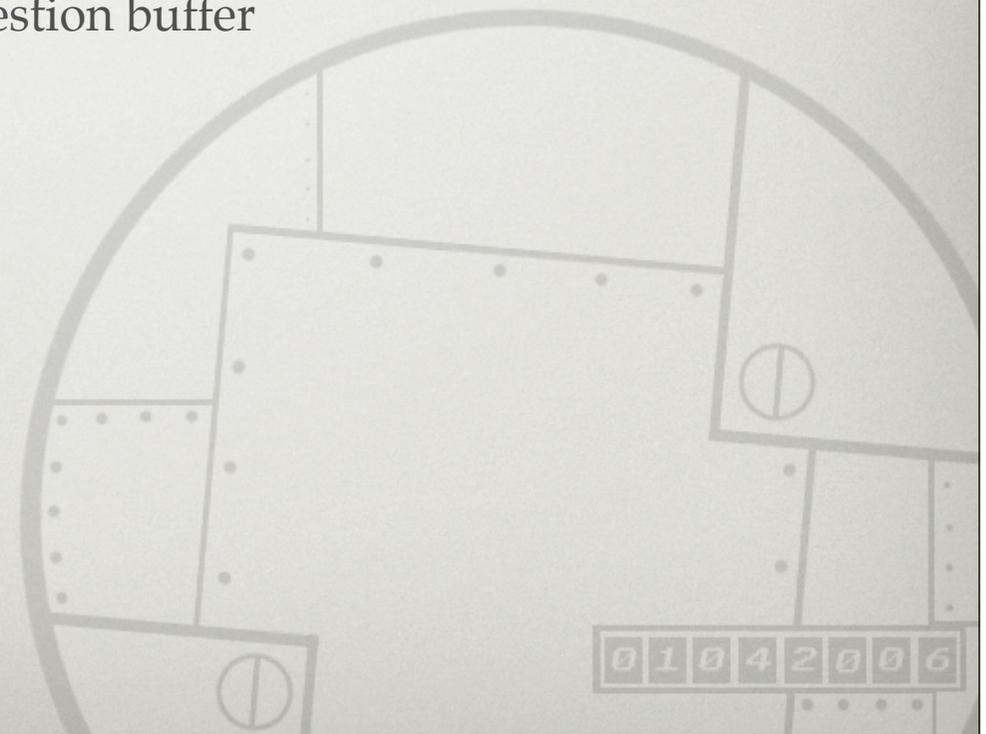




PLANNING AND ANALYSIS

REQUISITES

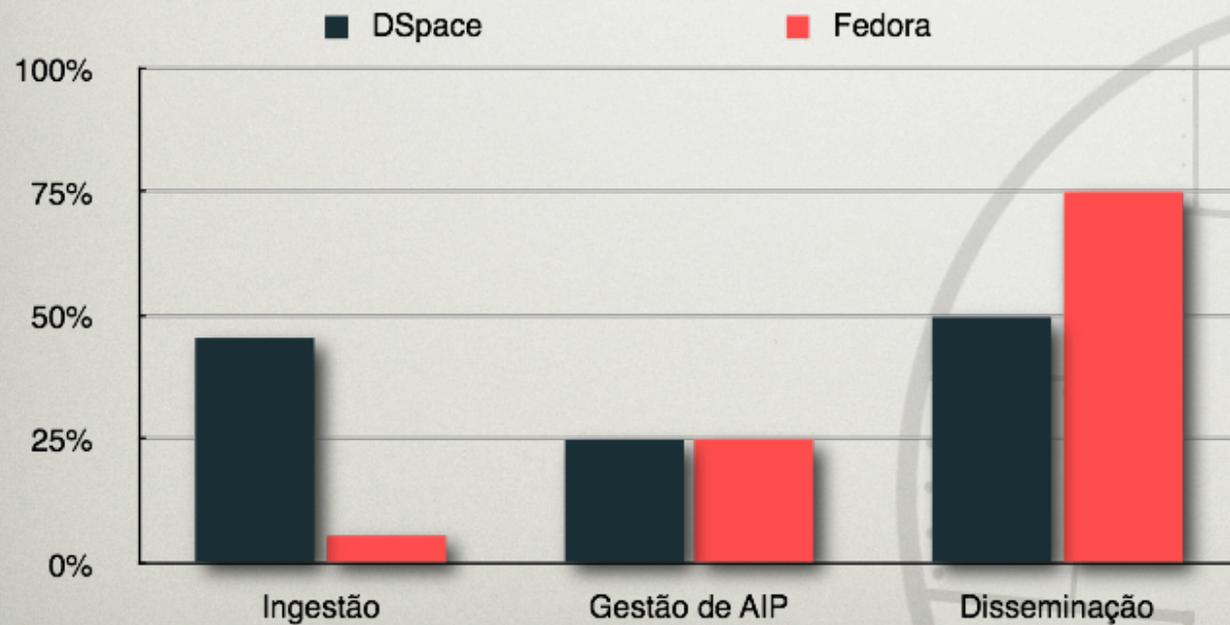
- Graphical Interface for Ingestion process
- Producer registry
- SIP production tool
- Ingestion feedback
- Partial Ingestion
- “Quarantine” zone: cache, ingestion buffer
- SIP validation
- Error reporting
- Persistent identifiers
- PREMIS event generation
- DIP digital signature
- ...



DEVELOPMENT FRAMEWORK

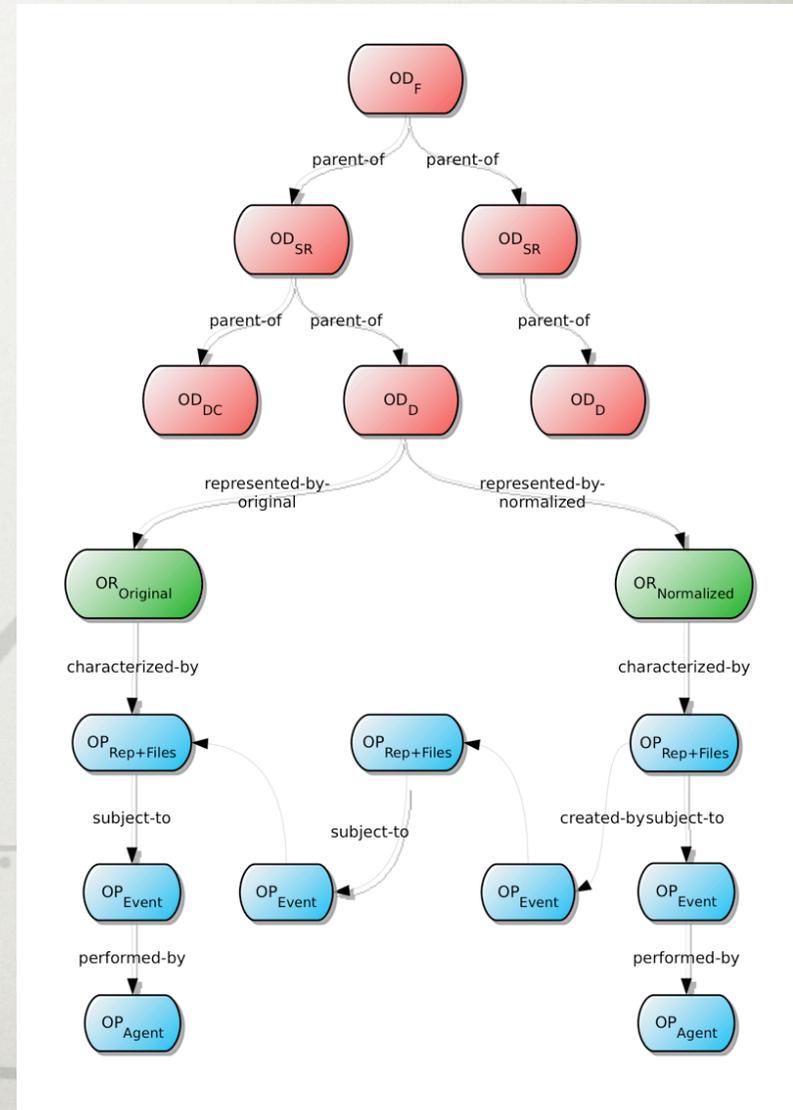
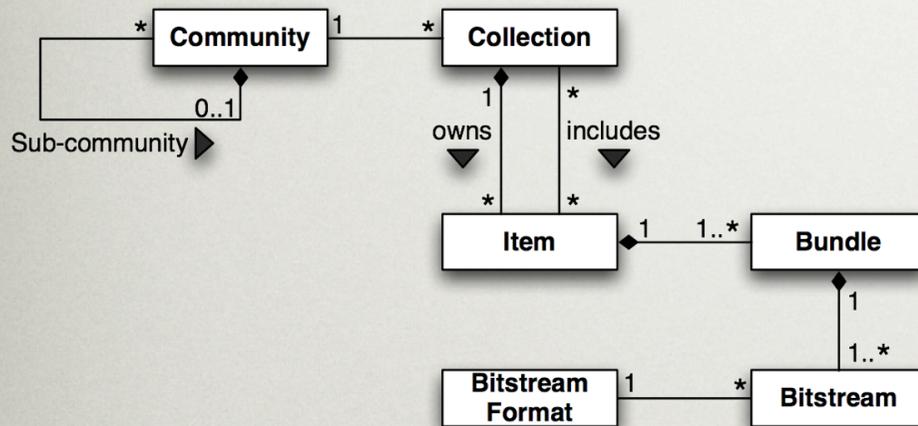


REQUISITES BASED COMPARAISON



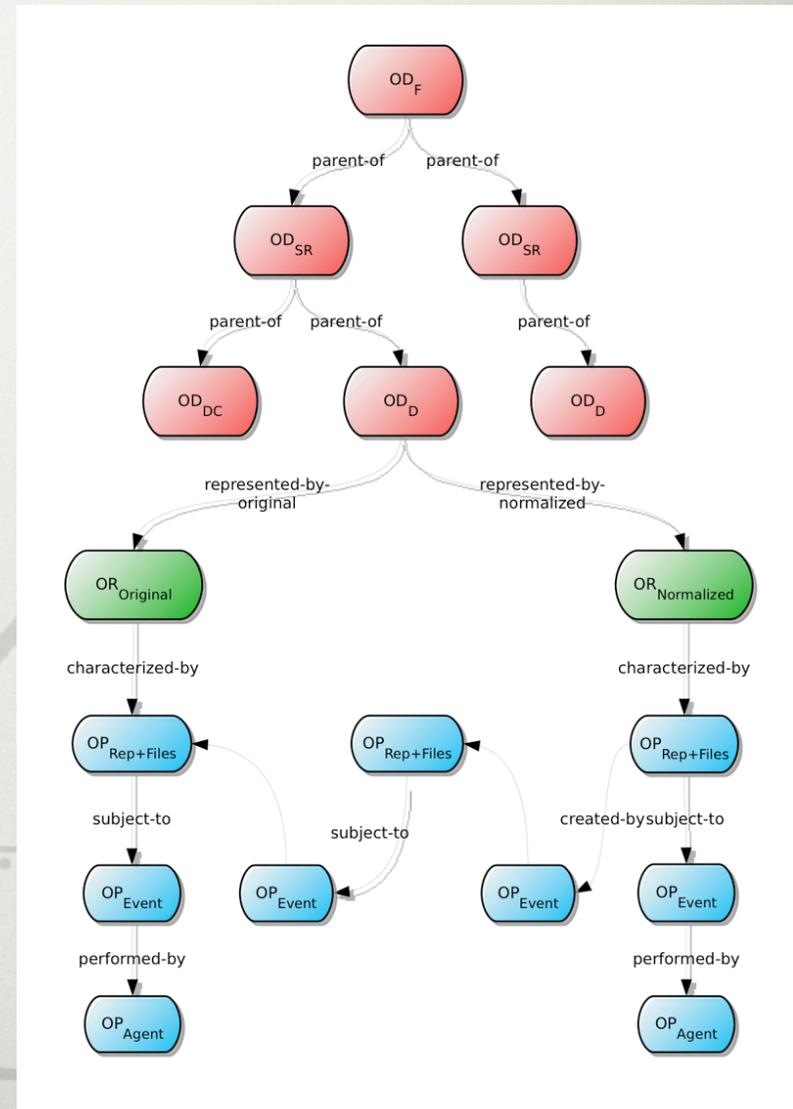
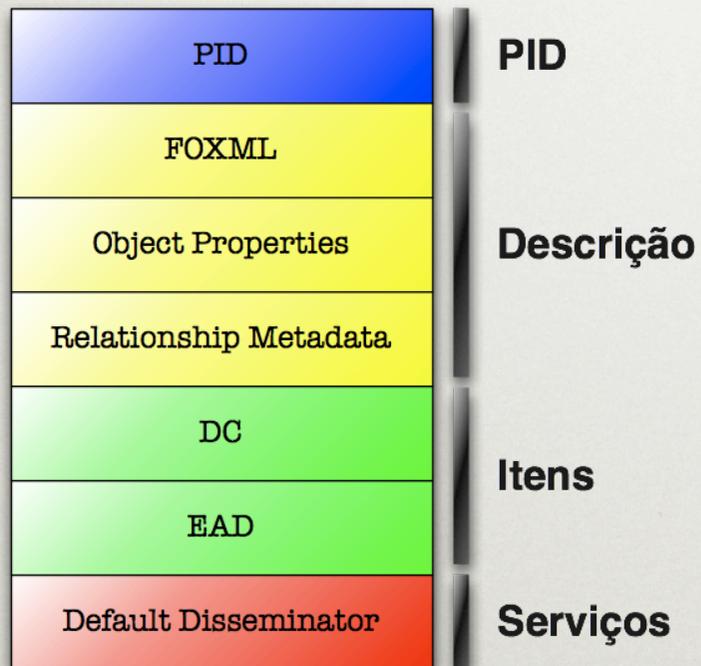
MATCHING DATA MODELS

DSpace

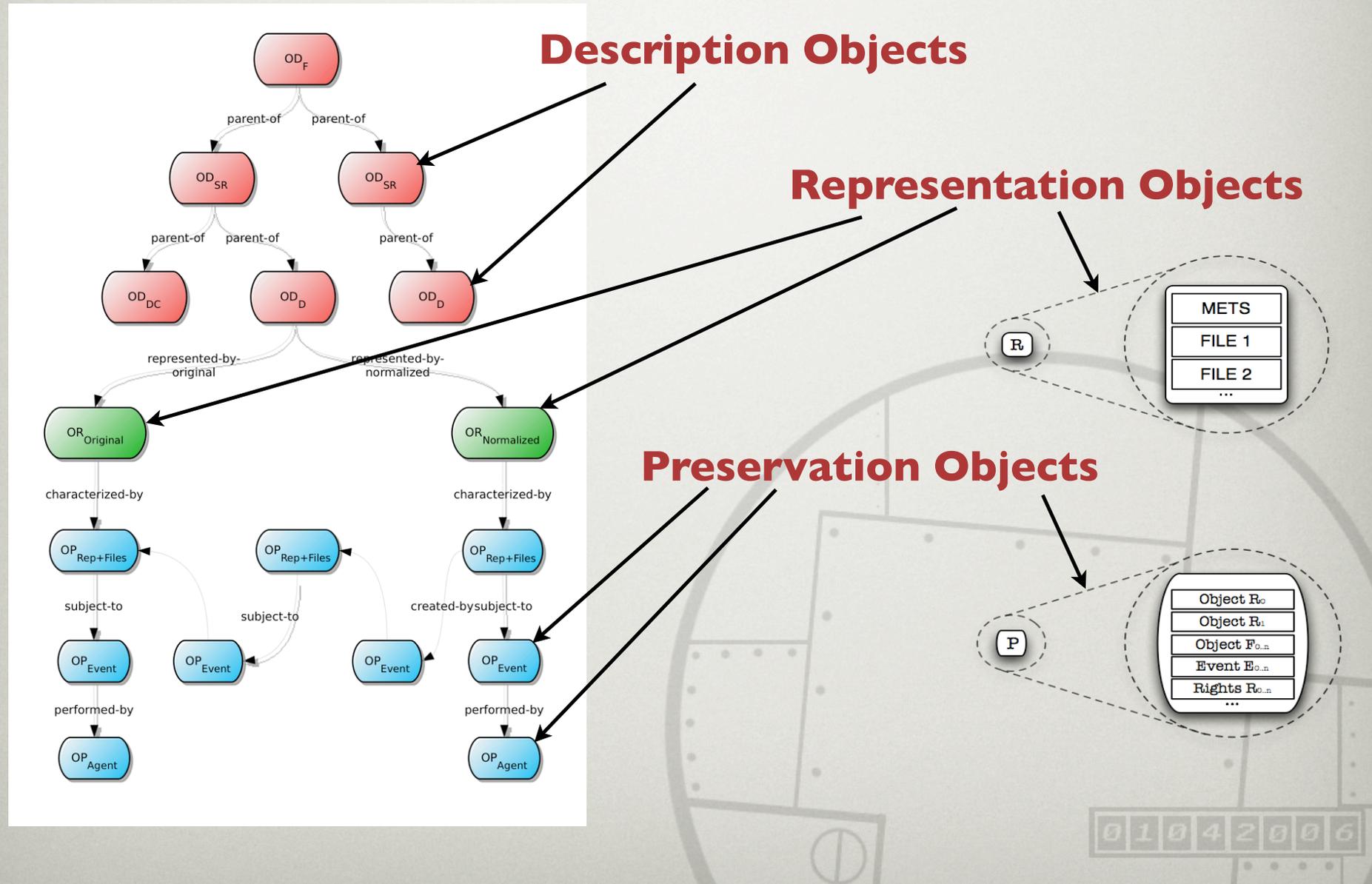


MATCHING DATA MODELS

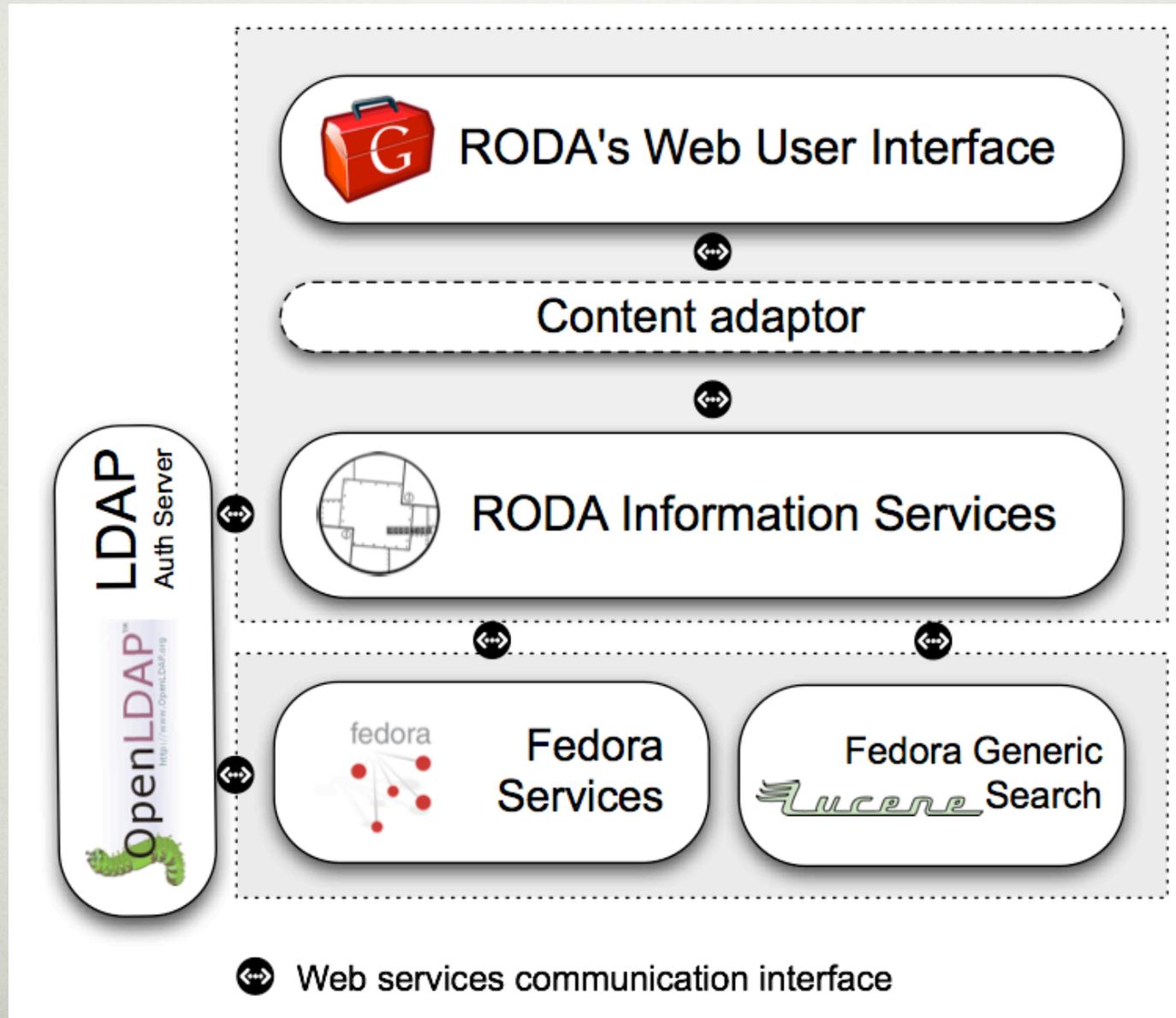
Fedora



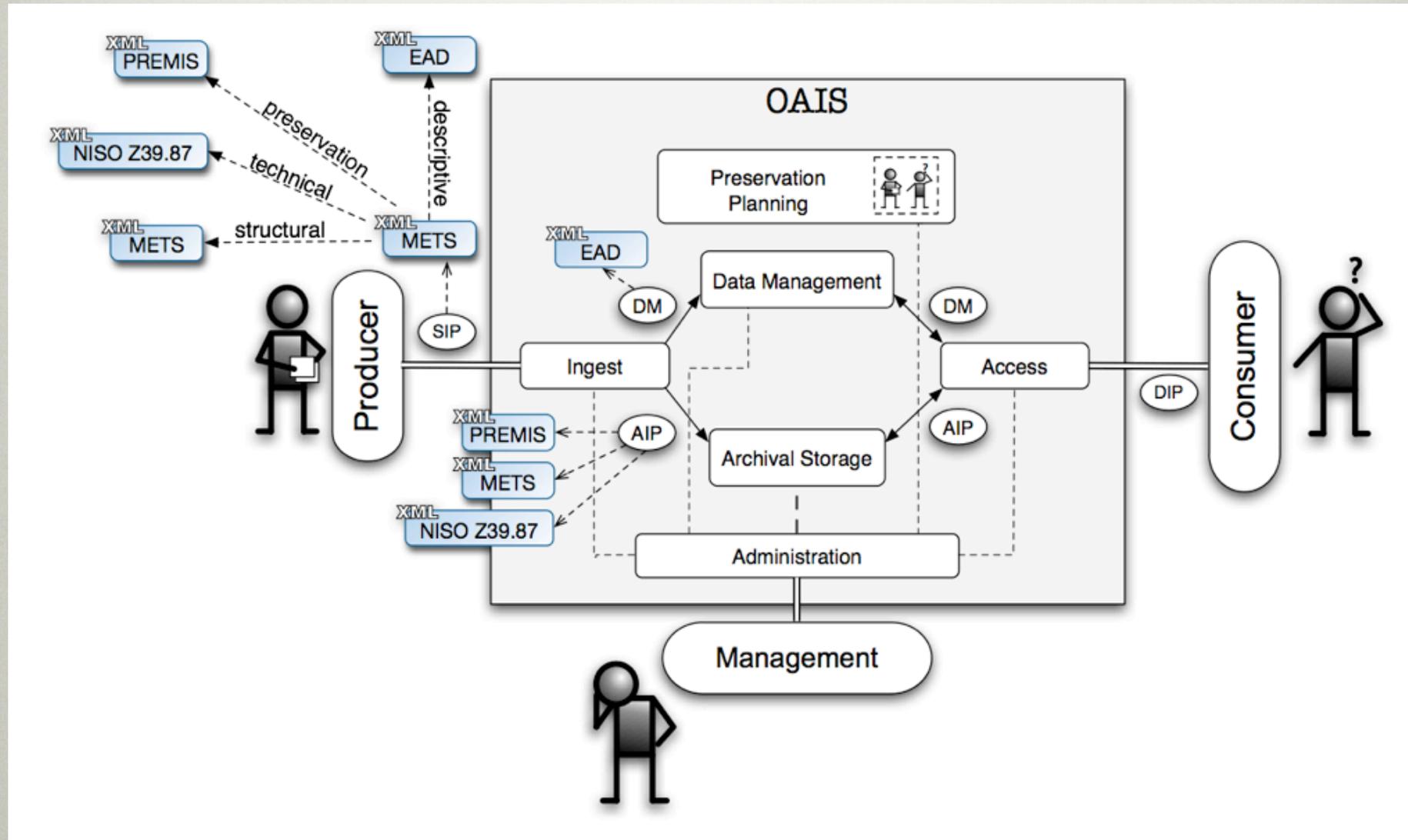
RODA DATA MODEL

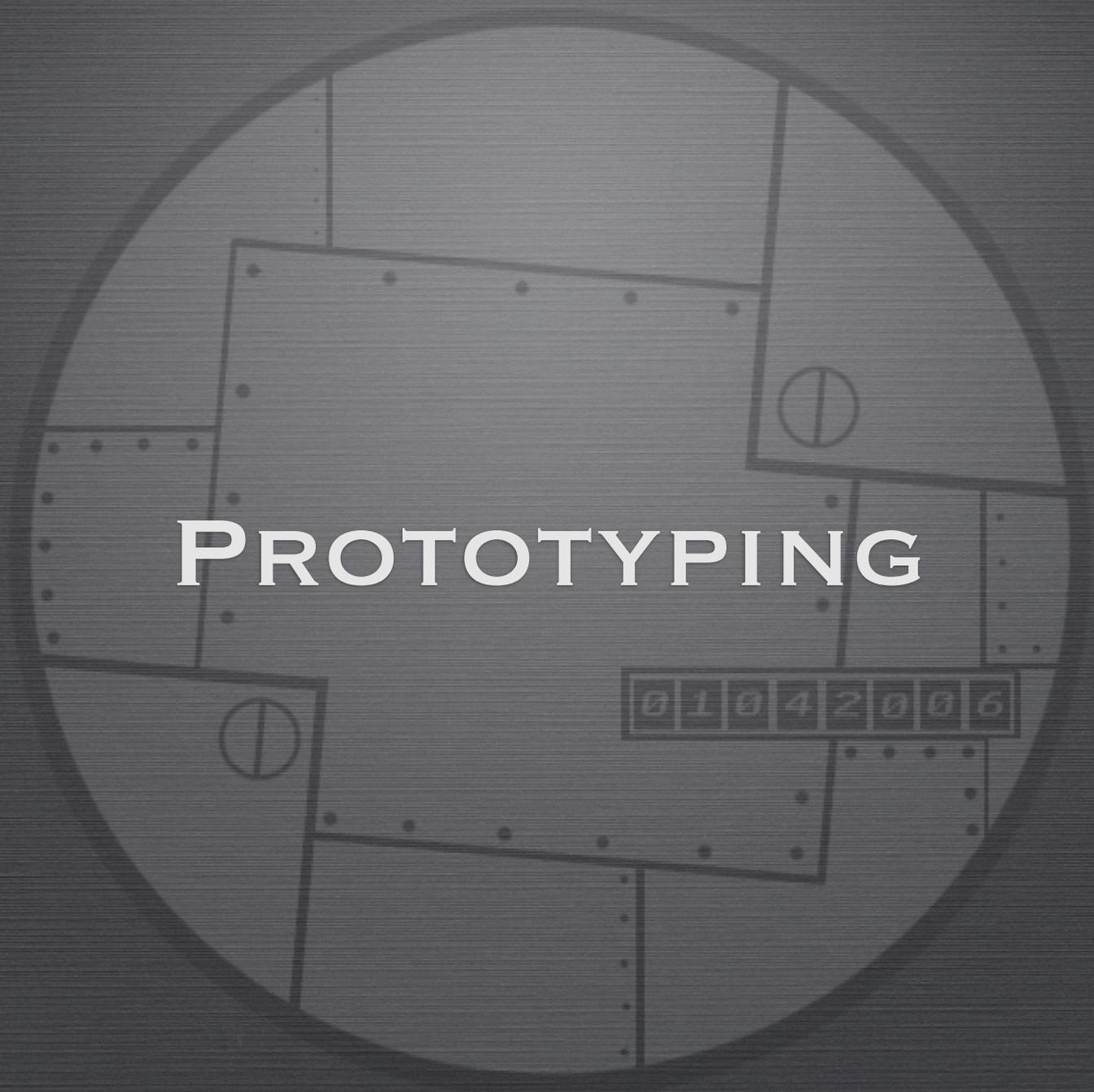


ARCHITECTURE



RODA SCHEMAS





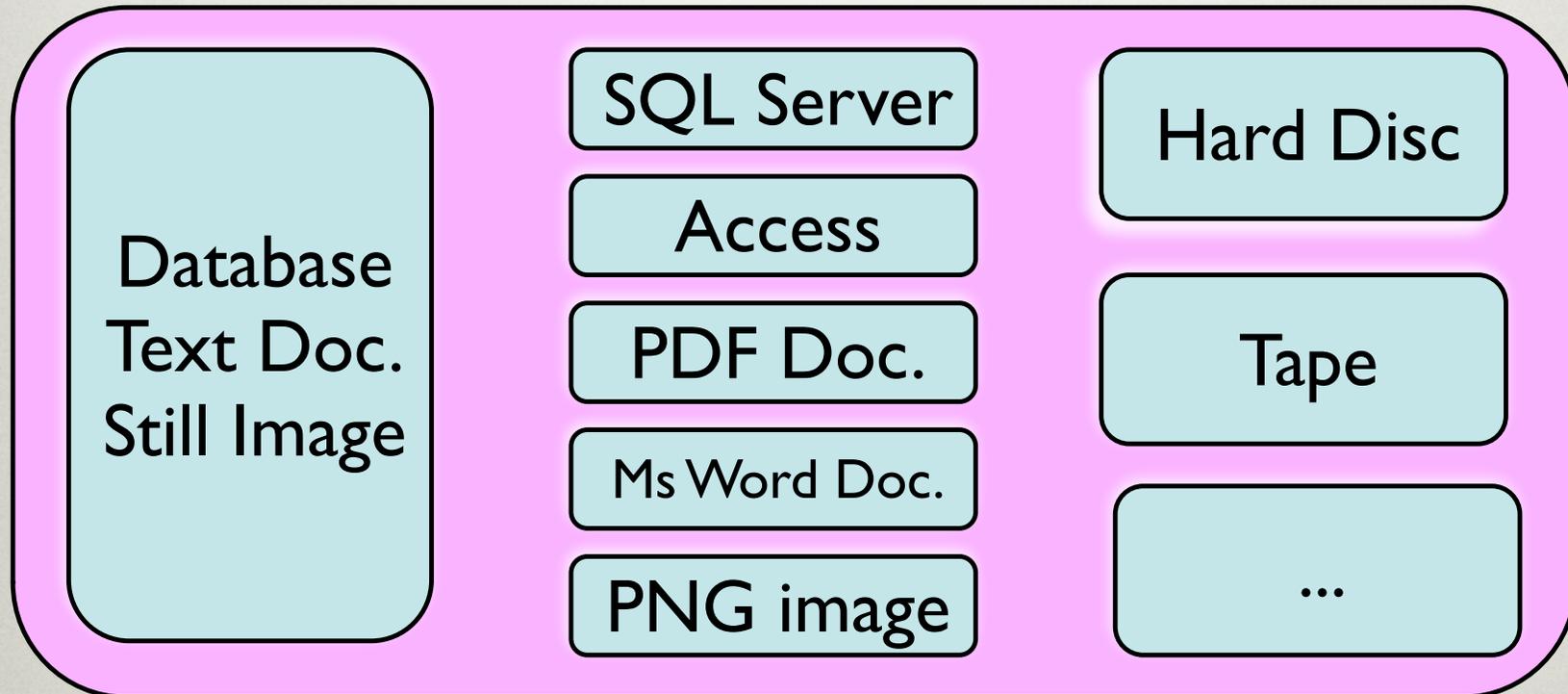
PROTOTYPING

Preserving Conceptual Object

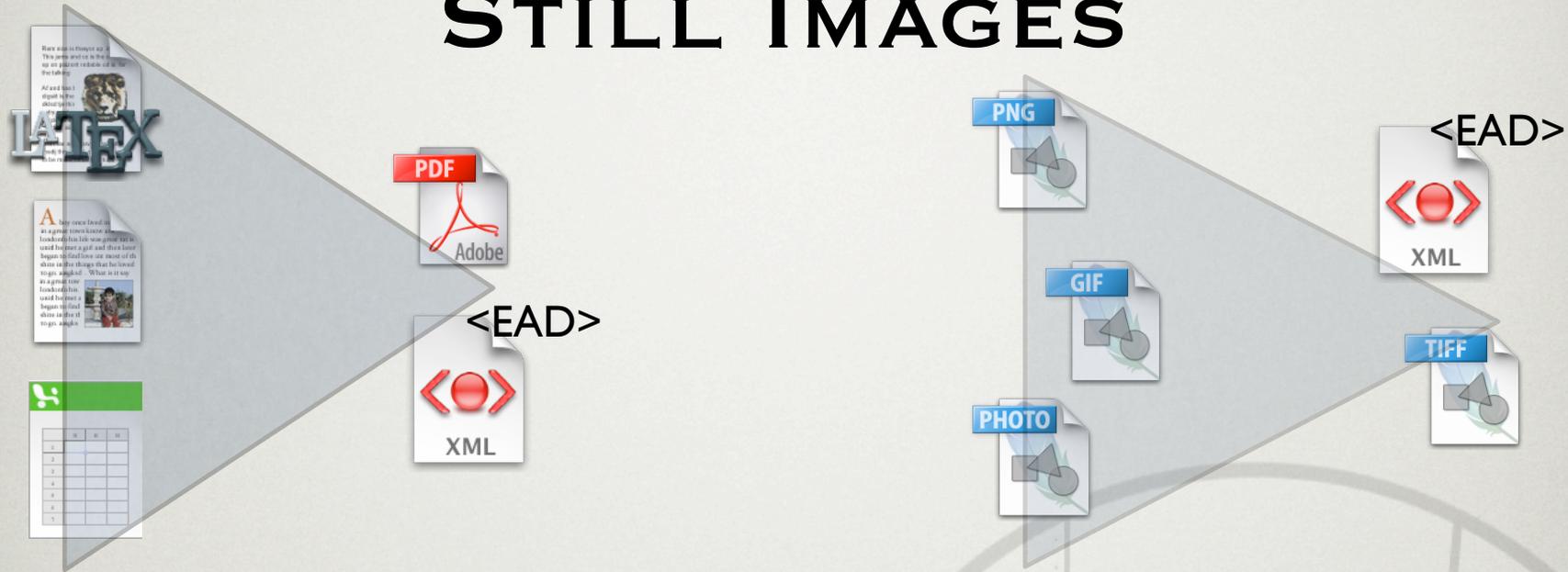
Conceptual
level

Logical
level

Physical
level



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.

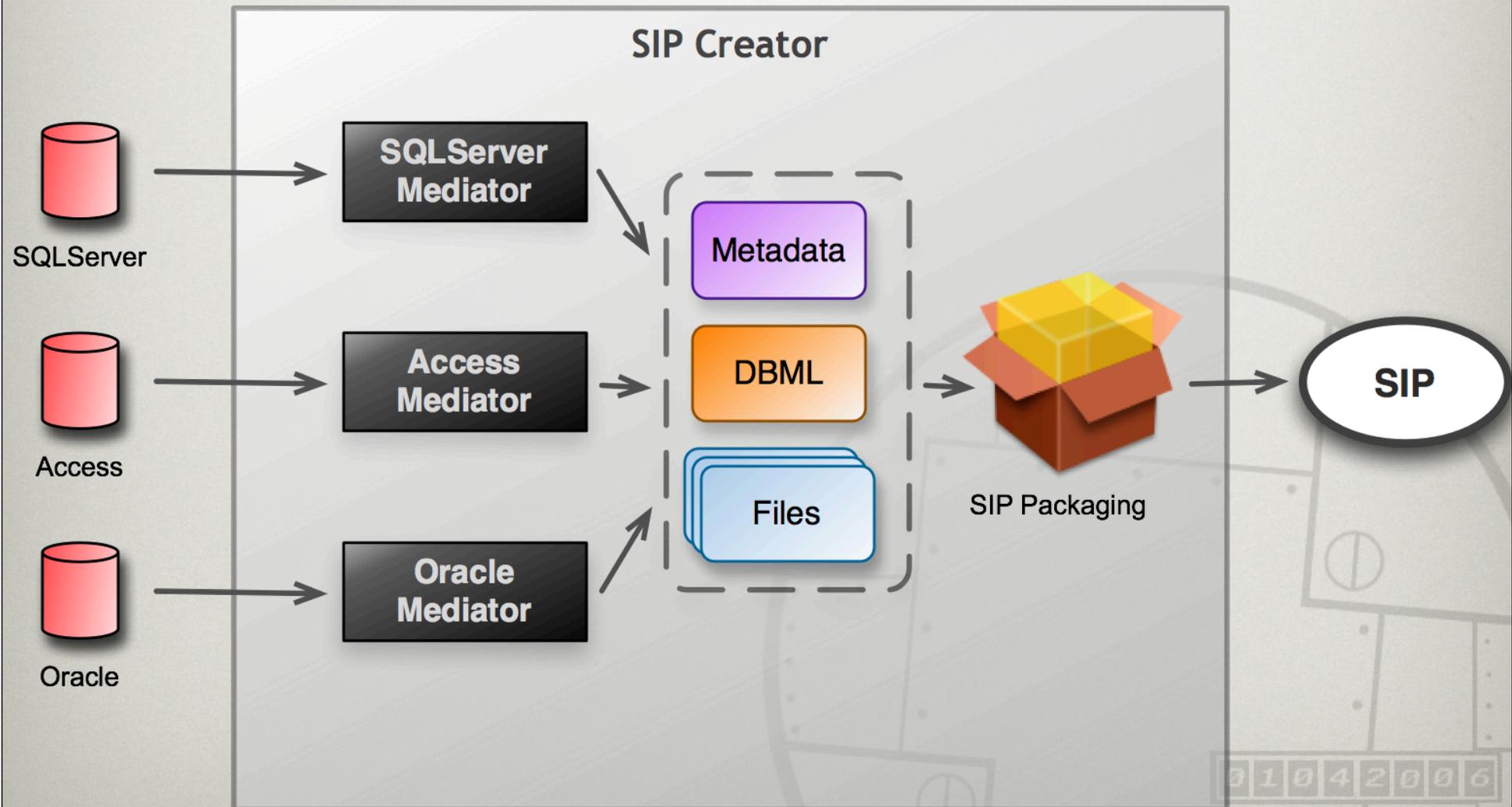
DATABASES

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

First prototype:

- Data
- Structure

SIP BUILDER



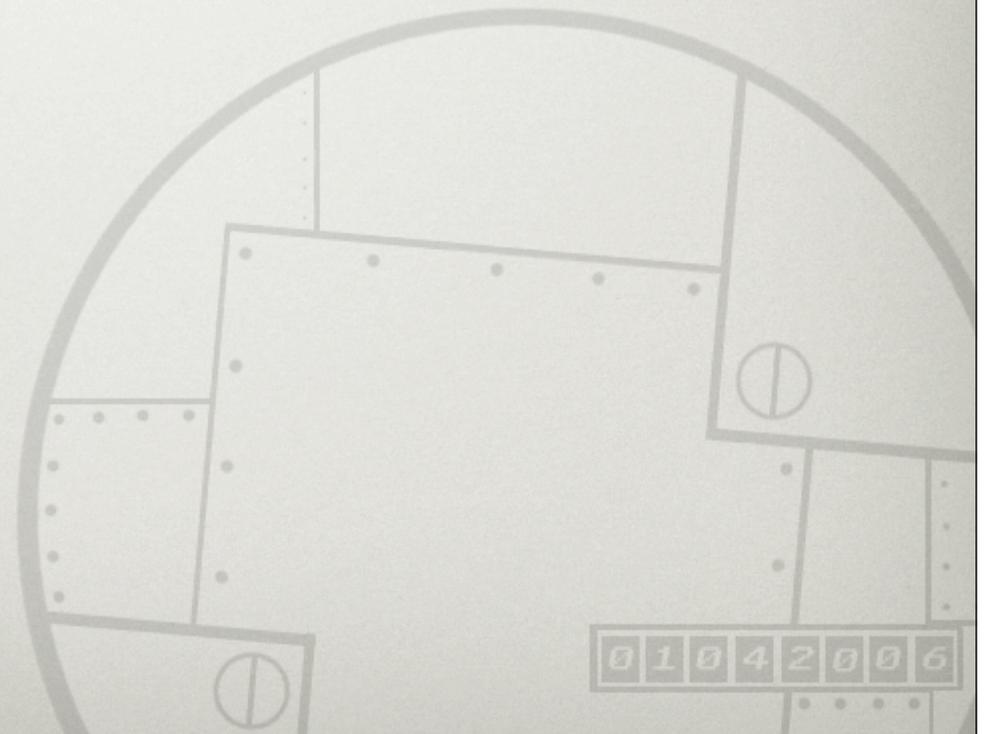
DBML

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

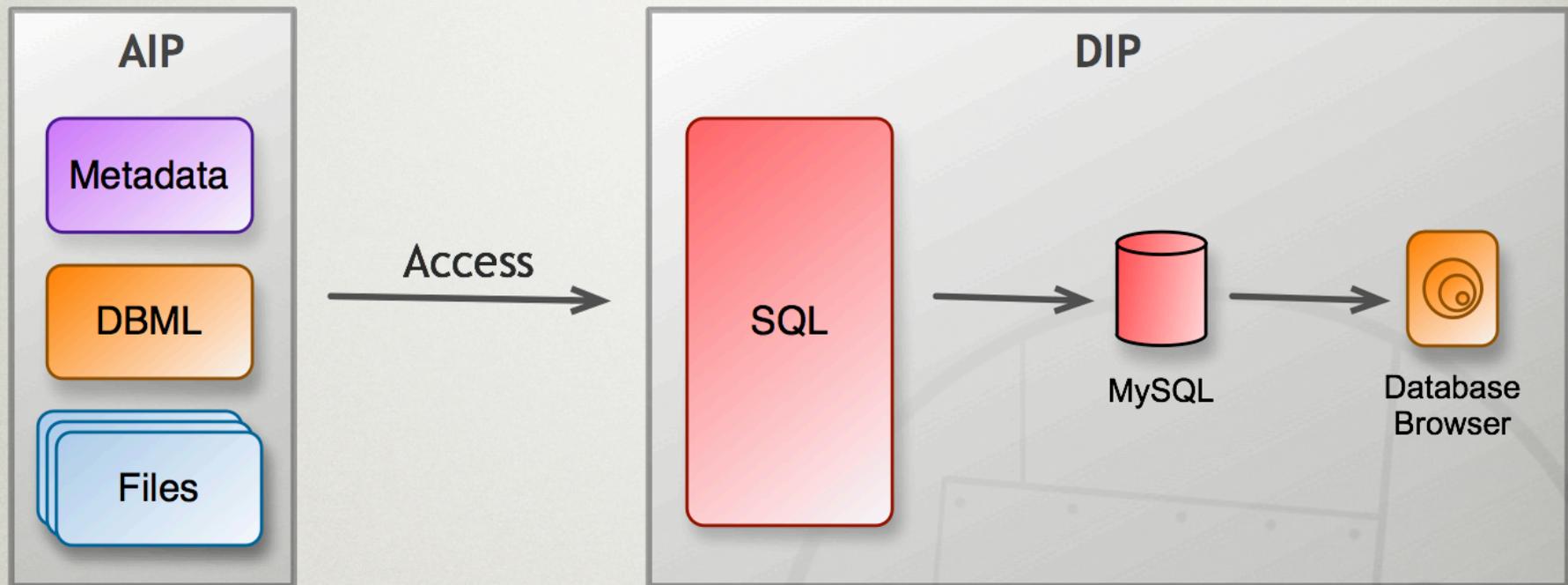
- Check and validation ...
- Generate SQL file
- Generate PREMIS



DISSEMINATION

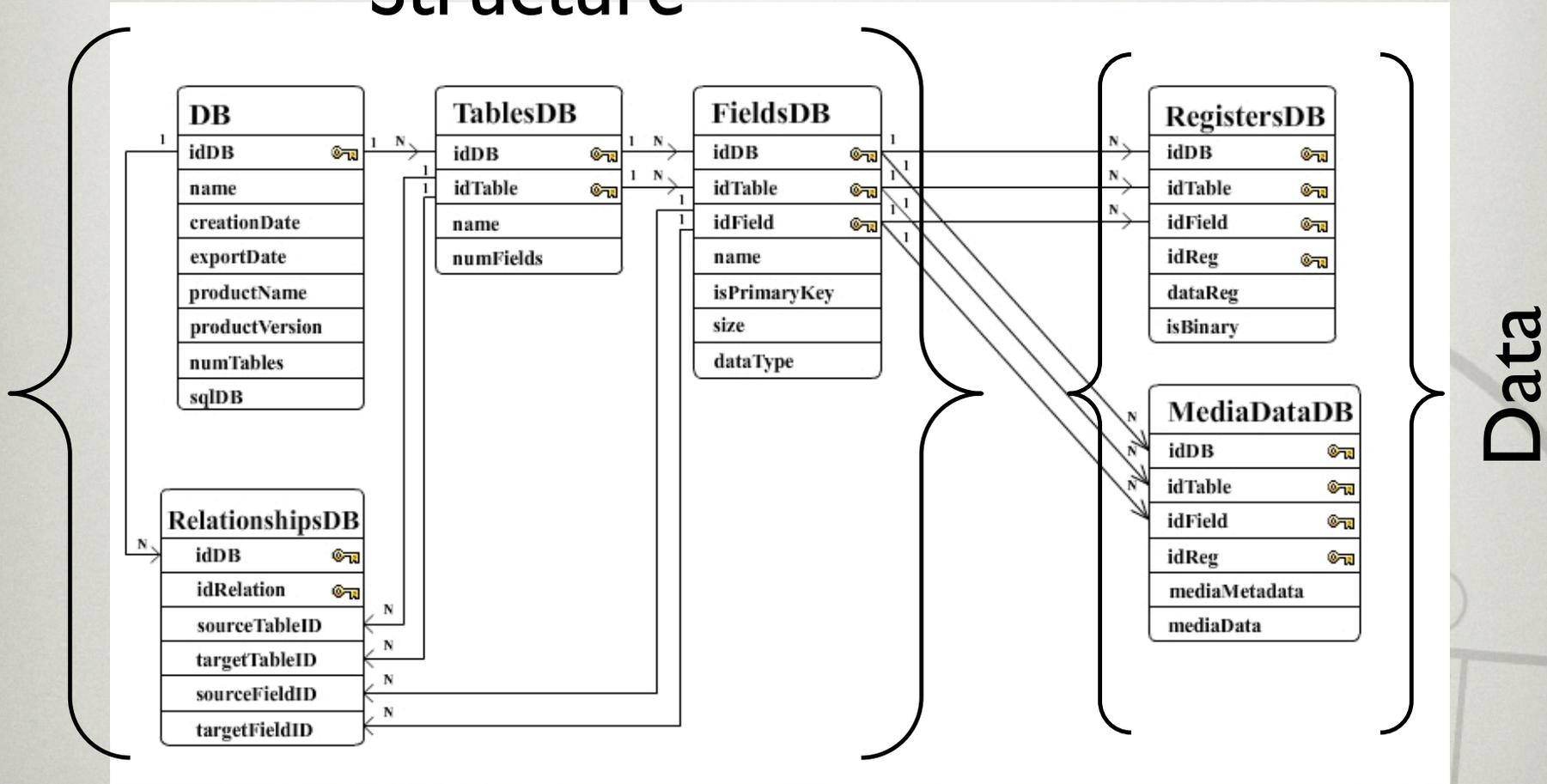
- Abstract Database Creation: a database of databases... Ingests databases from DBML (DBML-->SQL_{adb});
- Specific Database Creation: execute the SQL file in the selected RDMS

DISSEMINATION



DB ABSTRACT SCHEMA

Structure



Data

01042006

BROWSER

The screenshot displays the RODA Database repository GUI. At the top, the text "RODA" is visible. Below it, the page title is "Database repository gui". A navigation menu includes a "Home" link. The main content area features a "Bibliography:" section with a table:

ComponentID	ID	BibRef
790950	80770	null
790950	80771	null

Below the table, there is a "Components" link. Further down, a summary table shows "Number of tables" as 19 and a link to "SQL to restore original DB" with a "download file" button. At the bottom of the content area, there are links for "Show tables" and "Show relationships".

SEARCH ENGINE

**RODA**
Repositório de Objectos Digitais Autênticos

[Autenticação](#) [Lista de fundos](#) [Pesquisa](#) [Editor](#)

Localizar resultados

com todos os campos:

Título	porto	×
Nível de descrição	<input type="checkbox"/> fundos <input type="checkbox"/> classes <input type="checkbox"/> series <input checked="" type="checkbox"/> documentos	×
Datas extremas	1985 Janeiro 01	+

com pelo menos um dos campos:

Título		i
--------	--	---

com nenhum dos campos:

Título		i
--------	--	---

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

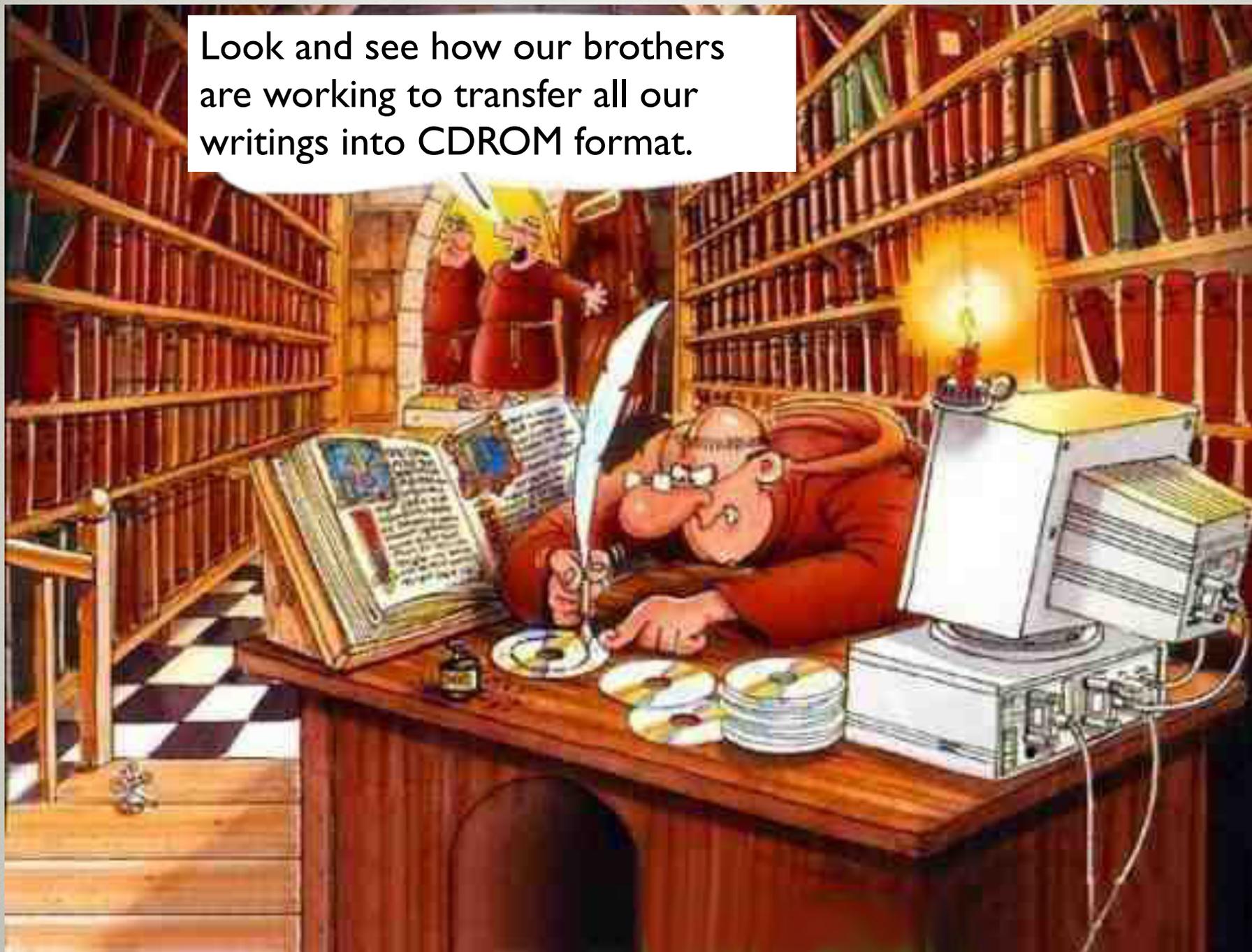
  		Pontuação: 100%
Nível de descrição: DC		
 		Pontuação: 100%
Nível de descrição: DC		
 		Pontuação: 100%
Nível de descrição: DC		
 		Pontuação: 100%
Nível de descrição: DC		

FINAL THOUGHTS

*“Data Preservation is a people problem”
Michael Lesk*

- 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 4^a Congresso Nacional de Administração Pública

Submetido por [rcaastro](#) 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

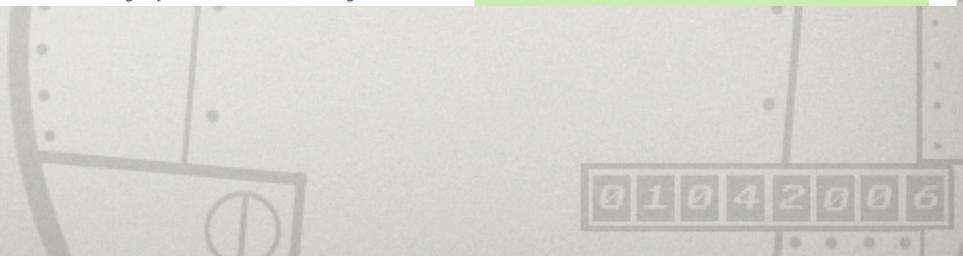
[Digital rights and asset management: access and digital preservation forum](#) (evento) (10 horas)

 [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 fórum...](#)

[mais](#)





LET'S PRESERVE TOMORROW'S HISTORY...

QUESTIONS?