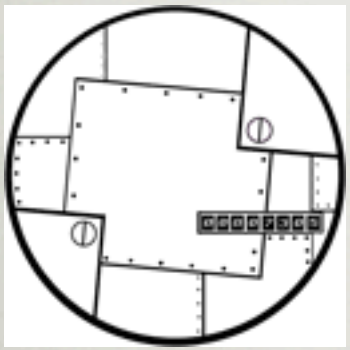
RODA: repositório de objectos digitais autênticos

José Carlos Ramalho
Dep. Informática
Universidade do Minho
jcr@di.uminho.pt

Contexto

digitArq

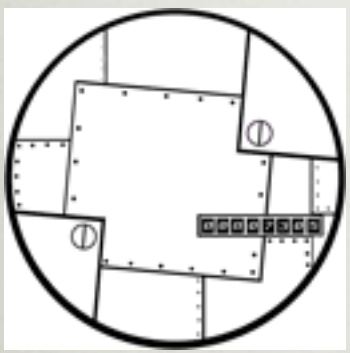
Digitarq (2003-2008)
- gestão de metainformação (EAD)
- gestão de ODs (NISO MIX)



RODA (2006-2009)
- gestão de metainformação (EAD)
- gestão de ODs (NISO MIX, METS, DBML, etc)
- políticas e protocolos de preservação digital

Contexto

Digitarq (2003-2008)
- gestão de metainformação (EAD)
- gestão de ODs (NISO MIX)



RODA (2006-2009)
- gestão de metainformação (EAD)
- gestão de ODs (NISO MIX, METS, DBML, etc)
- políticas e protocolos de preservação digital

CRAV
consulta real em ambiente virtual

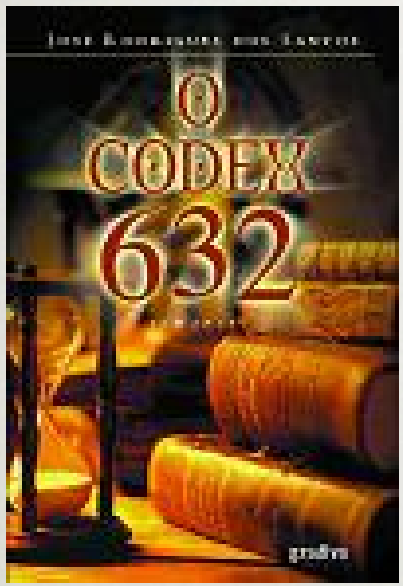
RODA: Motivação

- 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

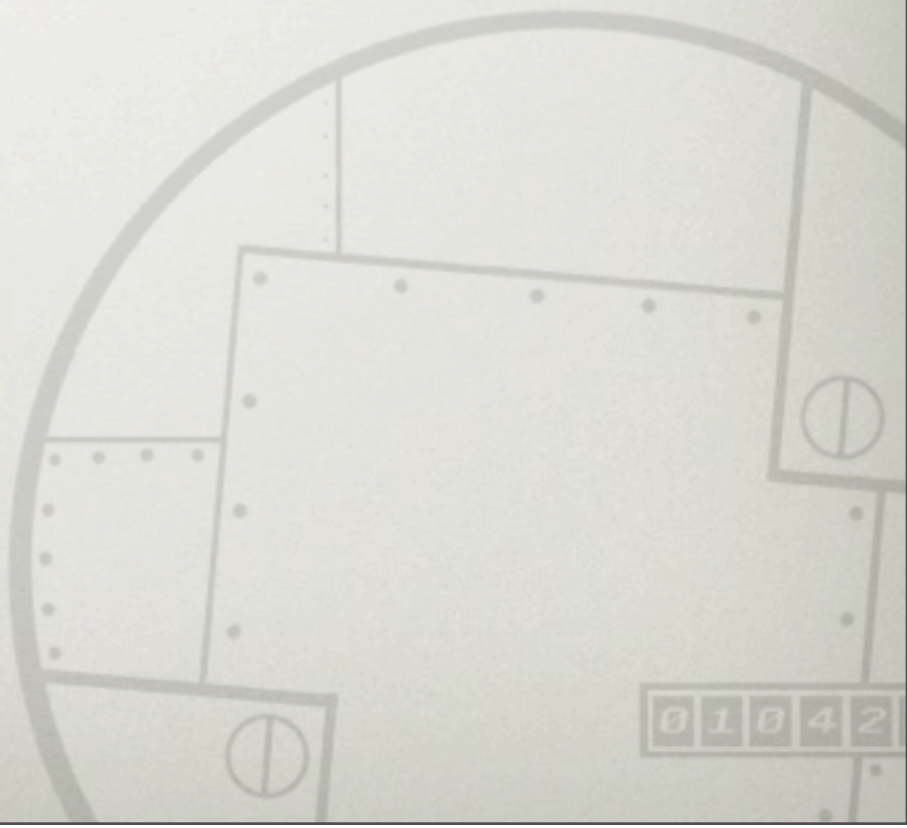
Assunto: Quem foi Cristovão Colombo?

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



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?

Autenticidade

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

EX: a bíblia

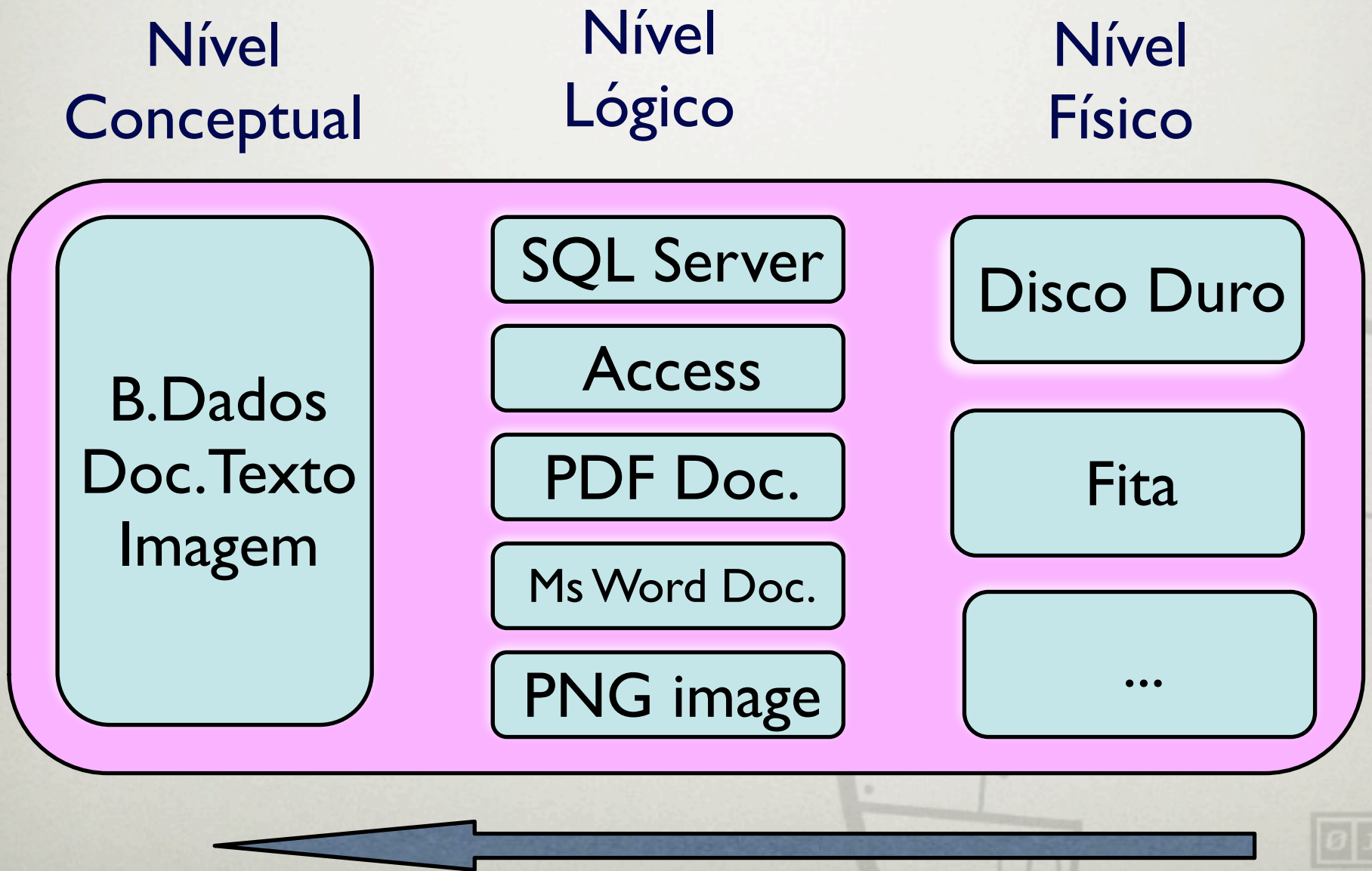
Como nos tornamos dignos de confiança?

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

Formatos suportados

Formatos suportados

Anatomia de um OD



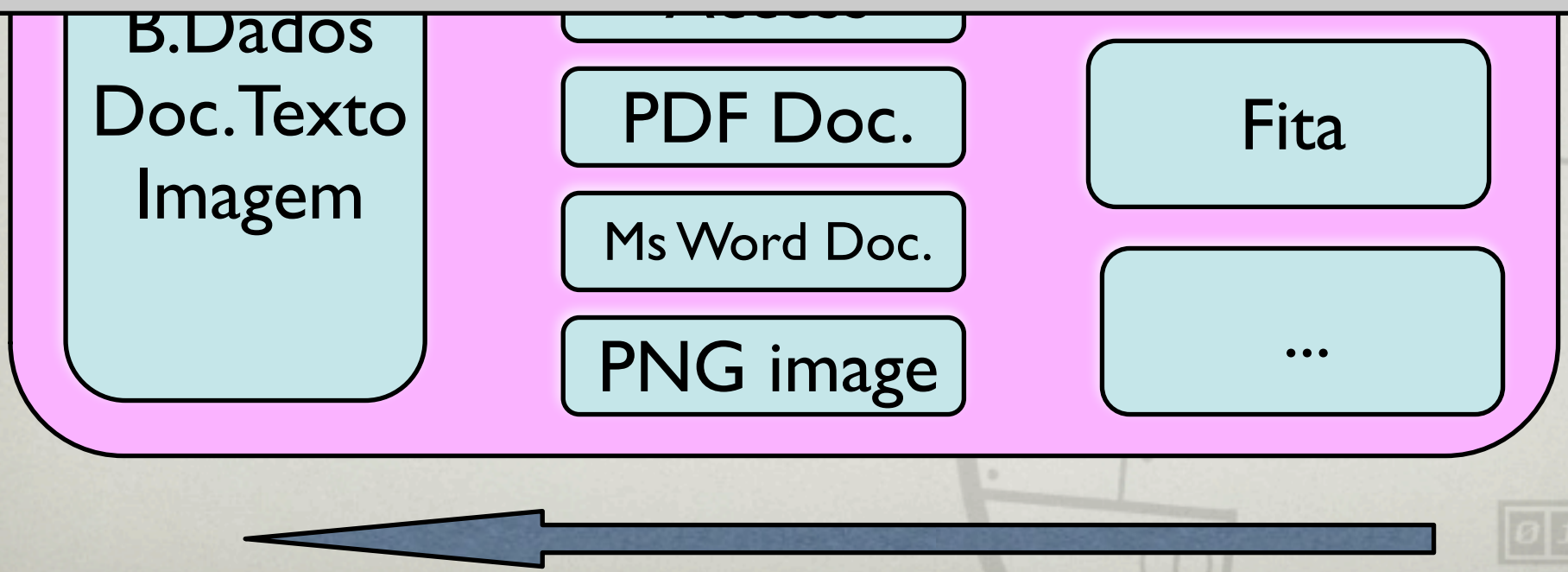
Anatomia de um OD

Nível Conceptual

Nível Lógico

Nível Físico

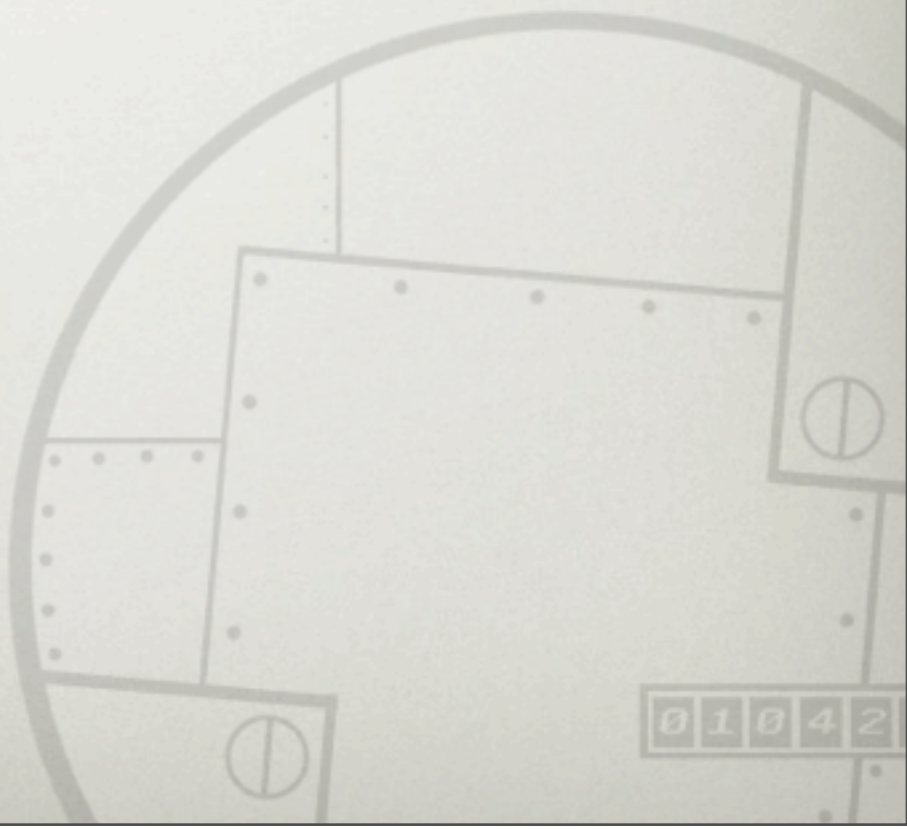
Se um destes níveis se tornar obsoleto deixamos de ter acesso ao OD



Estratégias de Preservação

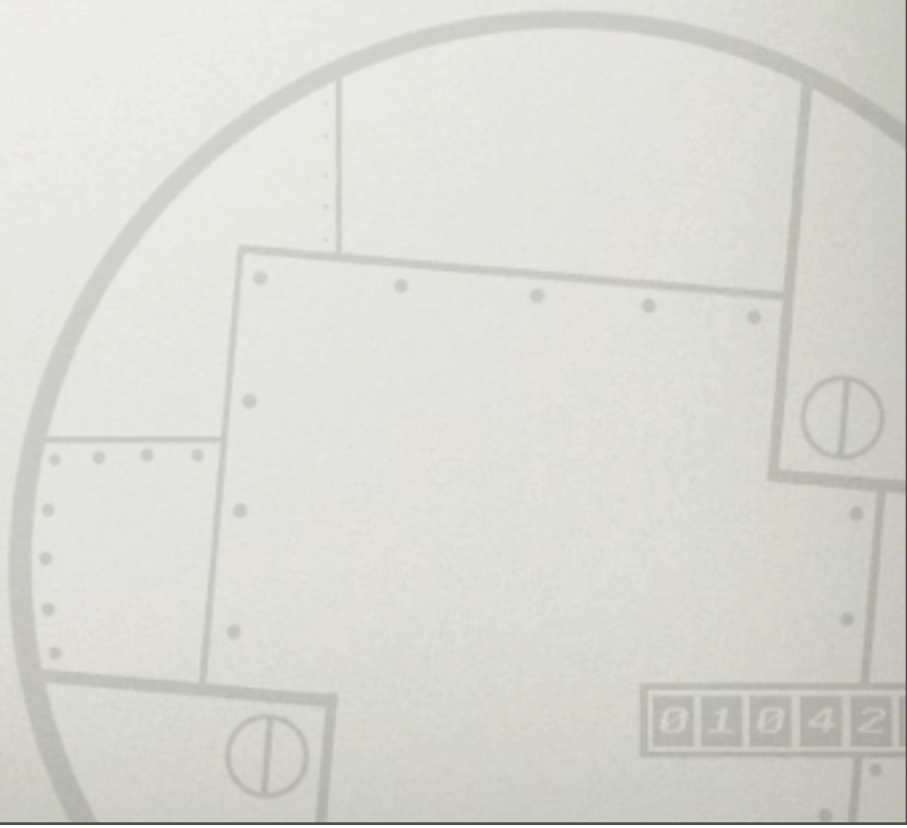
- Focada no **objecto físico**
  - Centrada na preservação da informação no seu **formato lógico** e/ou no **suporte físico;**
  - Usa a tecnologia original associada ao objecto para assegurar o acesso ao mesmo;
  - **Preservação tecnológica.**
- Focada no **objecto conceptual**
  - Centrada na preservação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

Emulação



Emulação

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

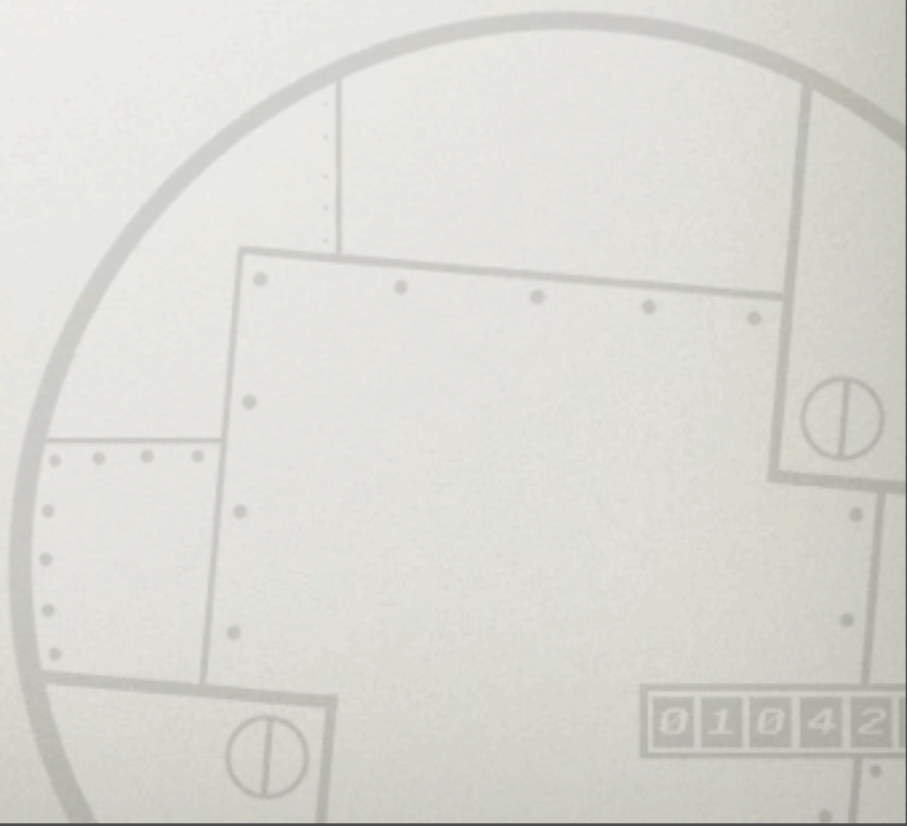


Emulação

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

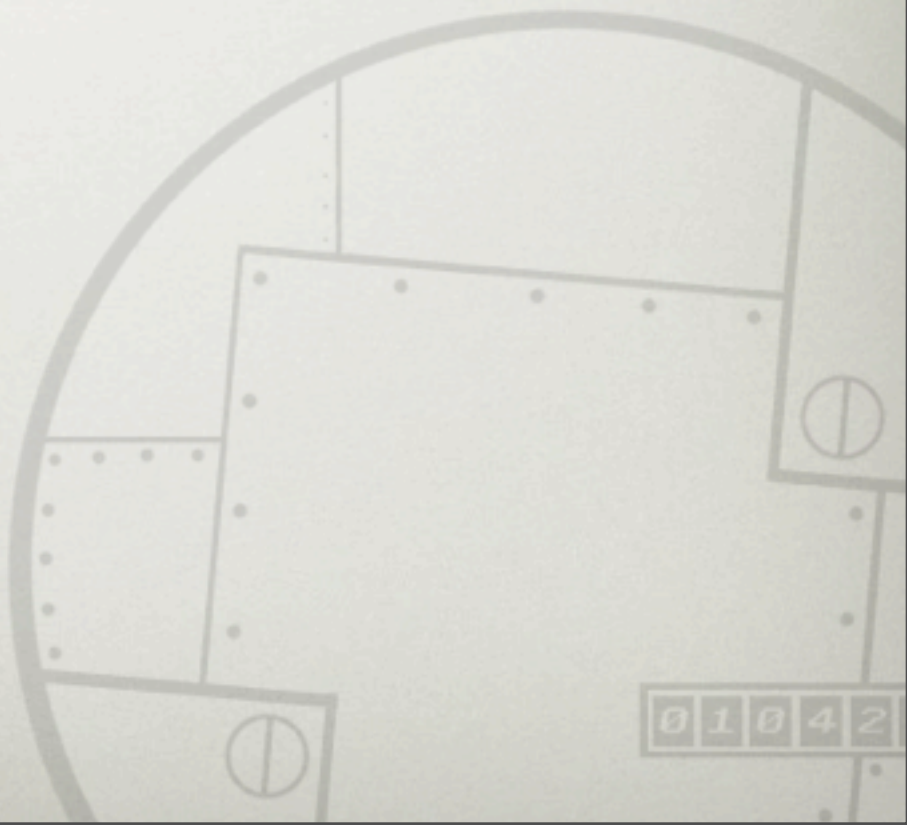
- Vantagens
  - Reprodução do contexto tecnológico original;
  - Preservação do ***look & feel*** do objecto.
- Desvantagens
  - Os emuladores também se tornam obsoletos;
  - Os utilizadores têm de trabalhar com sistemas obsoletos;
  - Criar um emulador é uma tarefa complexa;
  - Problemas de Copyright;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

Encapsulamento



Encapsulamento

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



Encapsulamento

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

- Vantagens
  - Permite adiar as responsibilidades 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

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.**

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

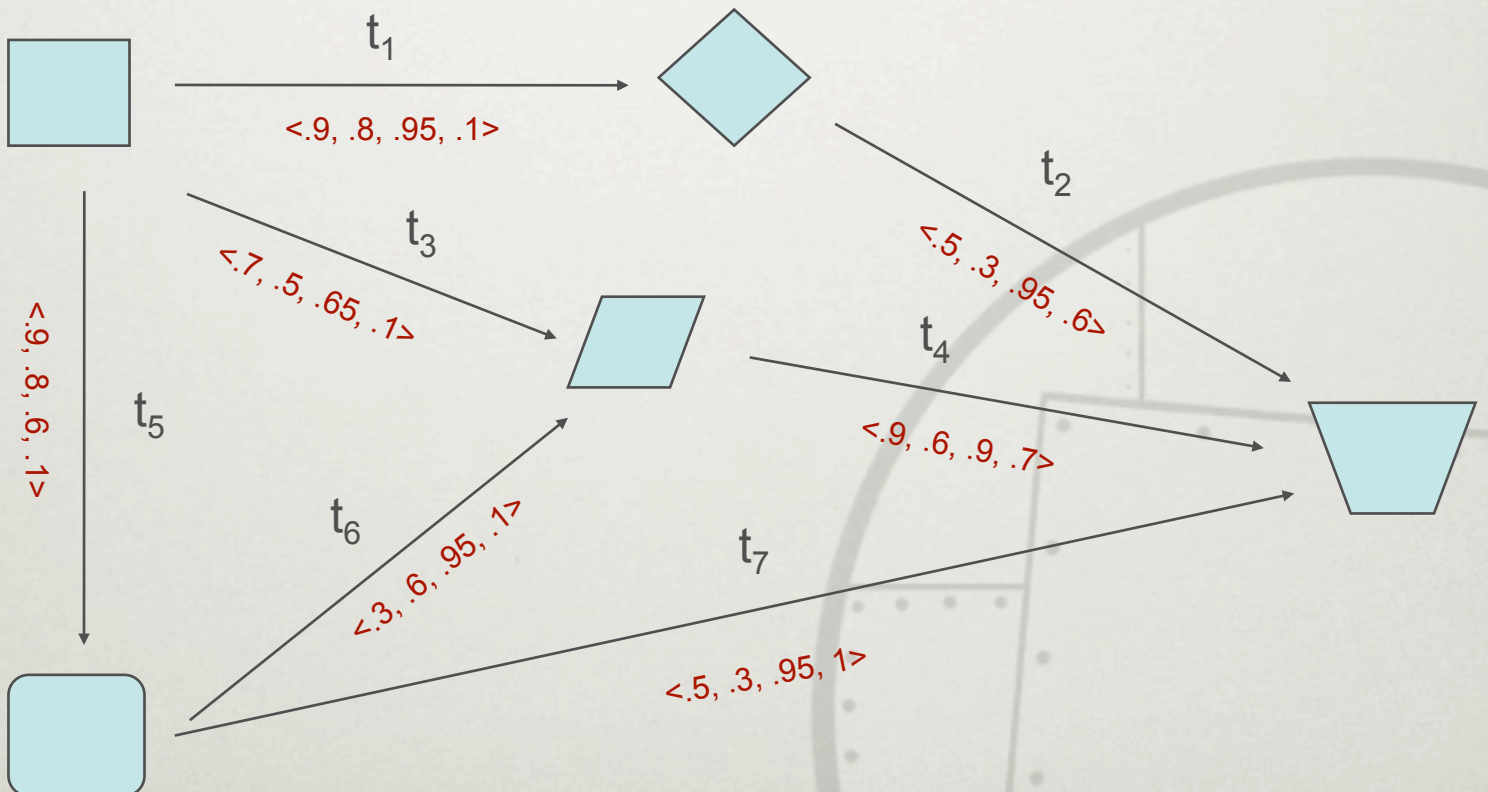
- Os ODs são disseminados em formatos familiares aos utilizadores;
- Não há necessidade de manter a plataforma original de hw/sw;
- A estratégia mais usada e a única que funcionou até hoje.

Desvantagens

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

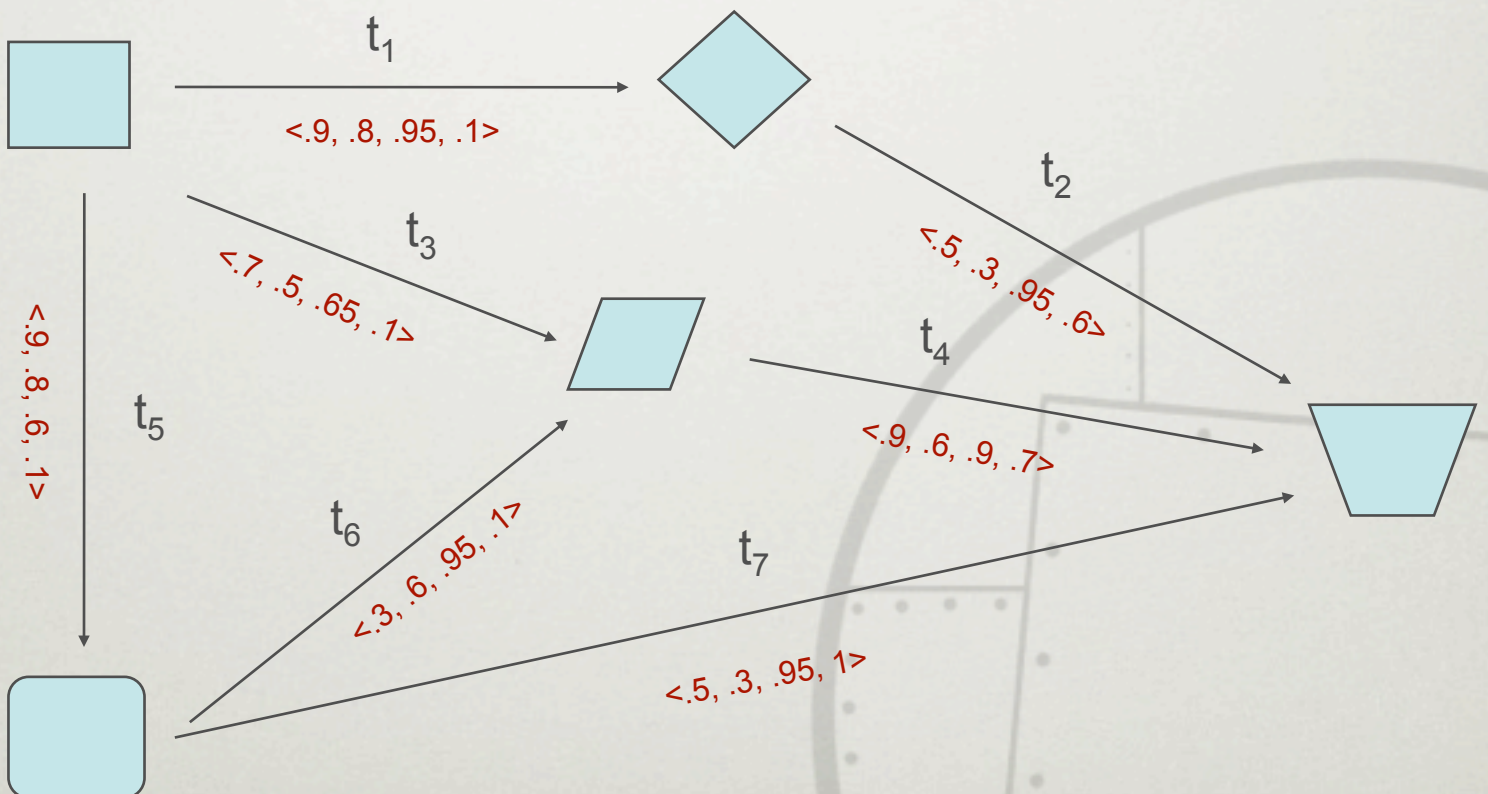
O que são propriedades significativas?

Serviços de Preservação

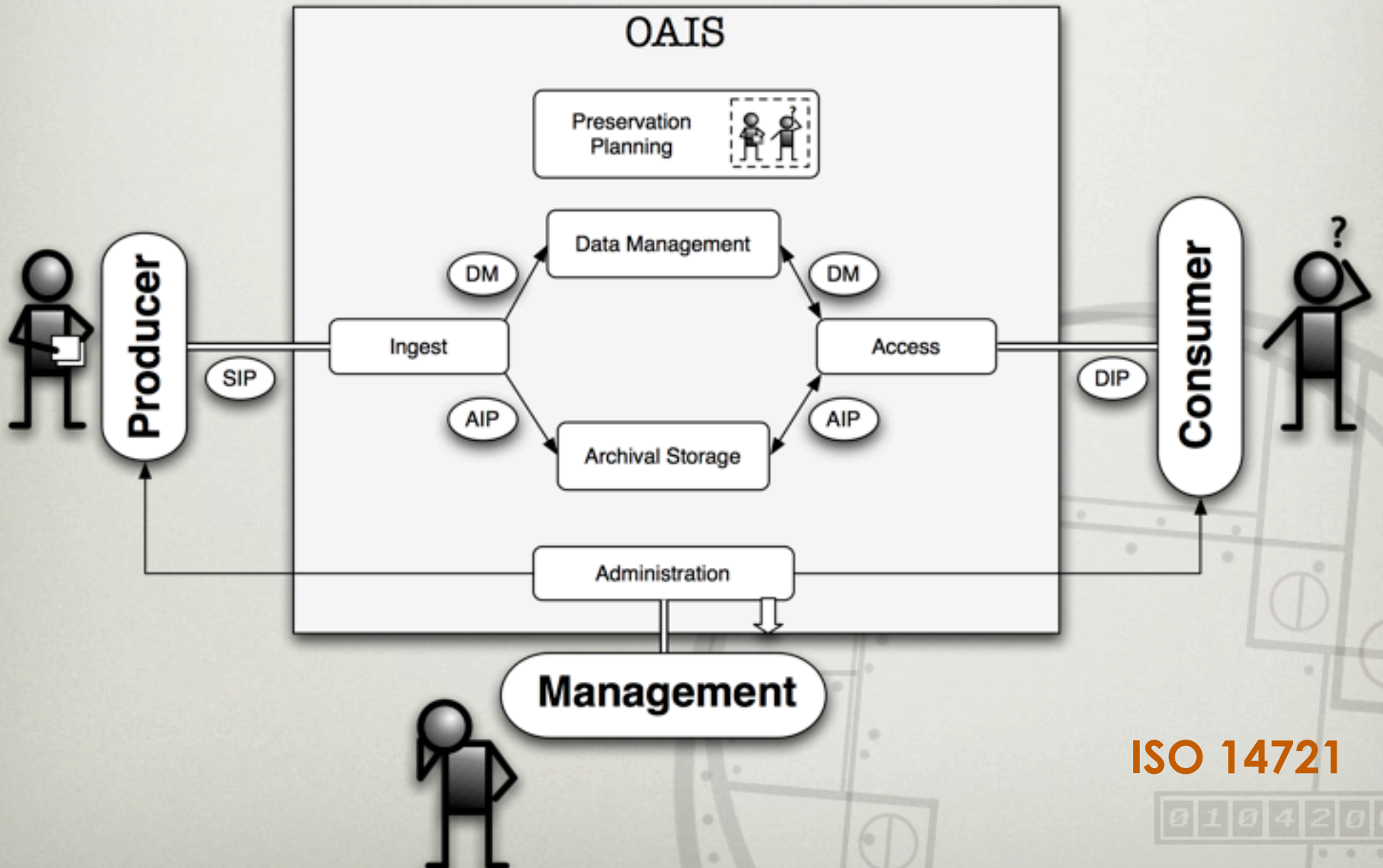


Serviços de Preservação

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



Open Archival Information System (OAIS)



ISO 14721

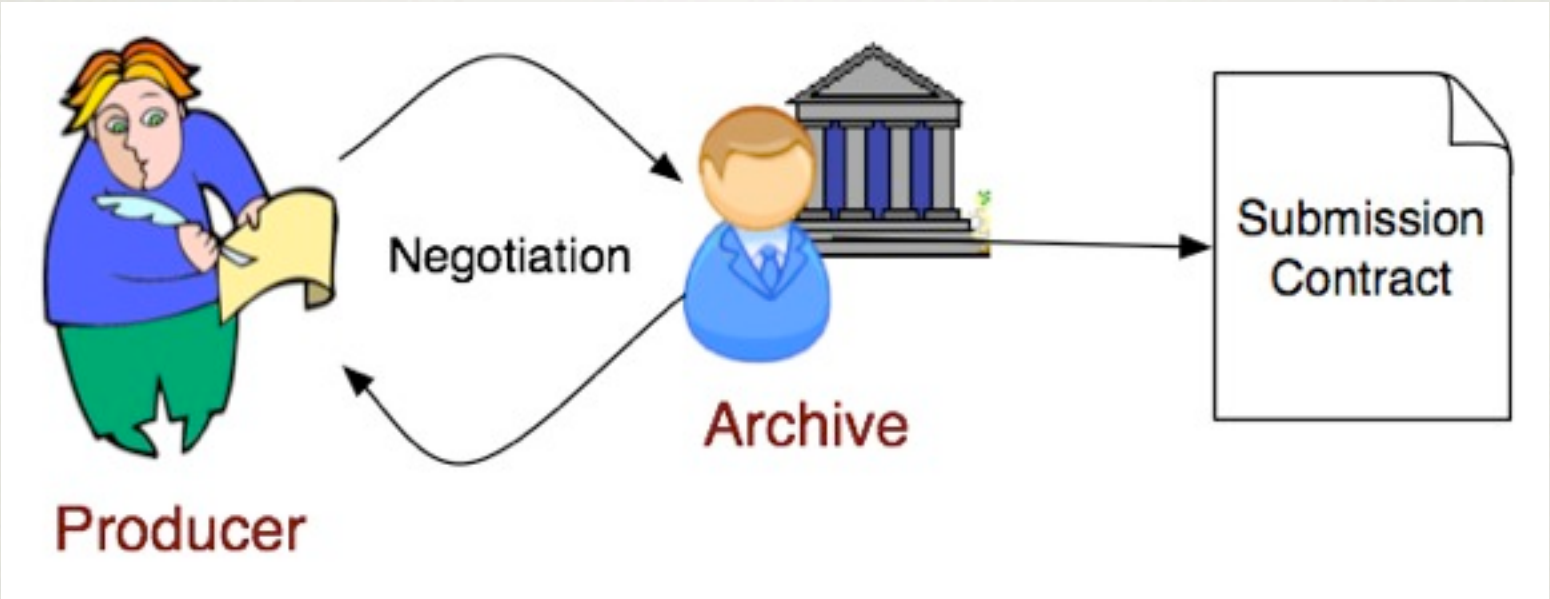
OAIS (Componentes Funcionais)

- 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

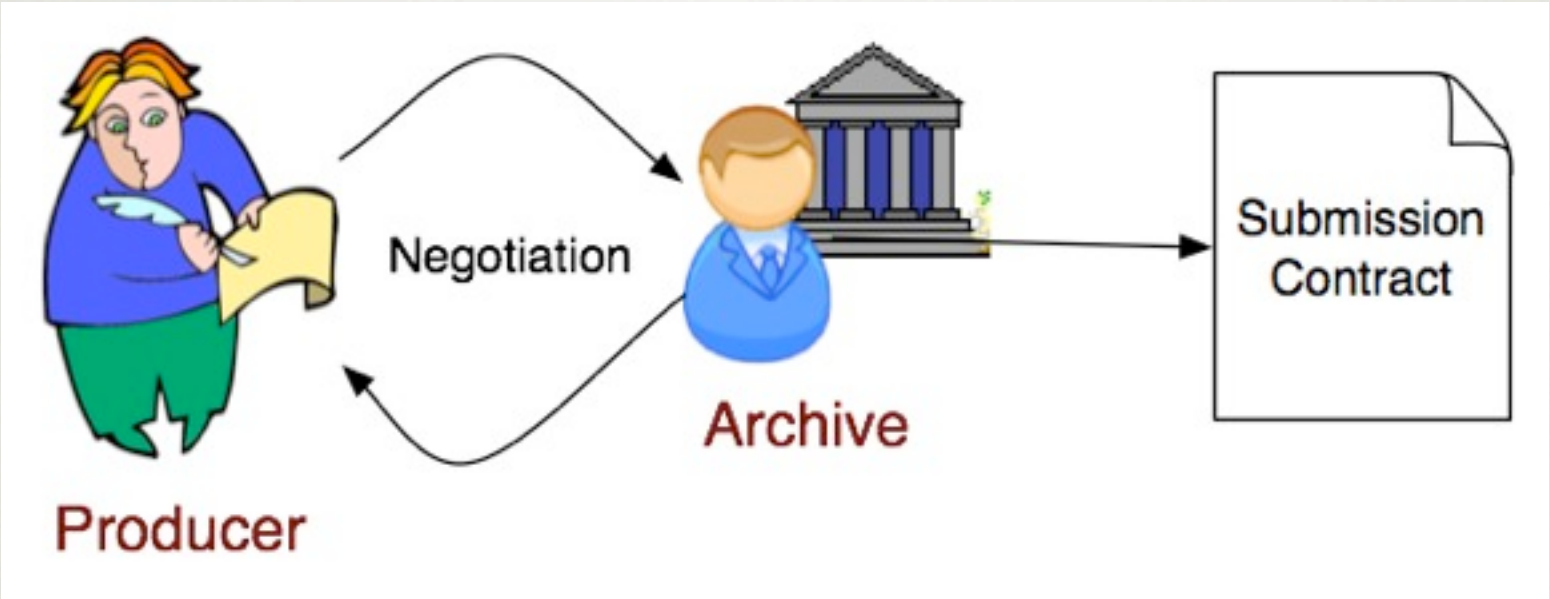
OAIS ("pacotes" de informação)

- Submission Information Package (SIP)
  - **Objecto Digital (representação ou representações);**
  - **Metainformação criada pelo produtor**
    - **demasiado livre...**
- Archival Information Package (AIP)
  - **Objecto digital arquivado (processado);**
  - **Metainformação:** a suficiente para assegurar a preservação e o acesso aos ODs;
    - modelo definido pelo PREMIS;
- Dissemination Information Package (DIP)
  - OD transformado no **formato** a ser **entregue** ao **consumidor;**
  - **Metainformação.**

Ingestão



Ingestão



Contrato de Submissão/Envio

- Especificação do SIP;
- Especificação do workflow de ingestão.

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

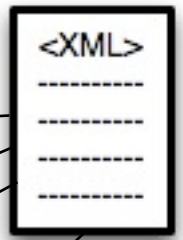
Technical Metadata:
- color
- dimensions
- ...

criation properties:
- date
- hardware
- ...

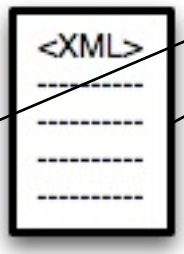
Descriptive Metadata:
- producer
- colection
- ...

SIP Structure (example)

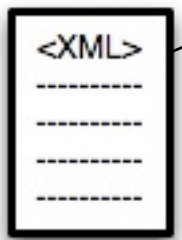
one still image



Manifest



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>

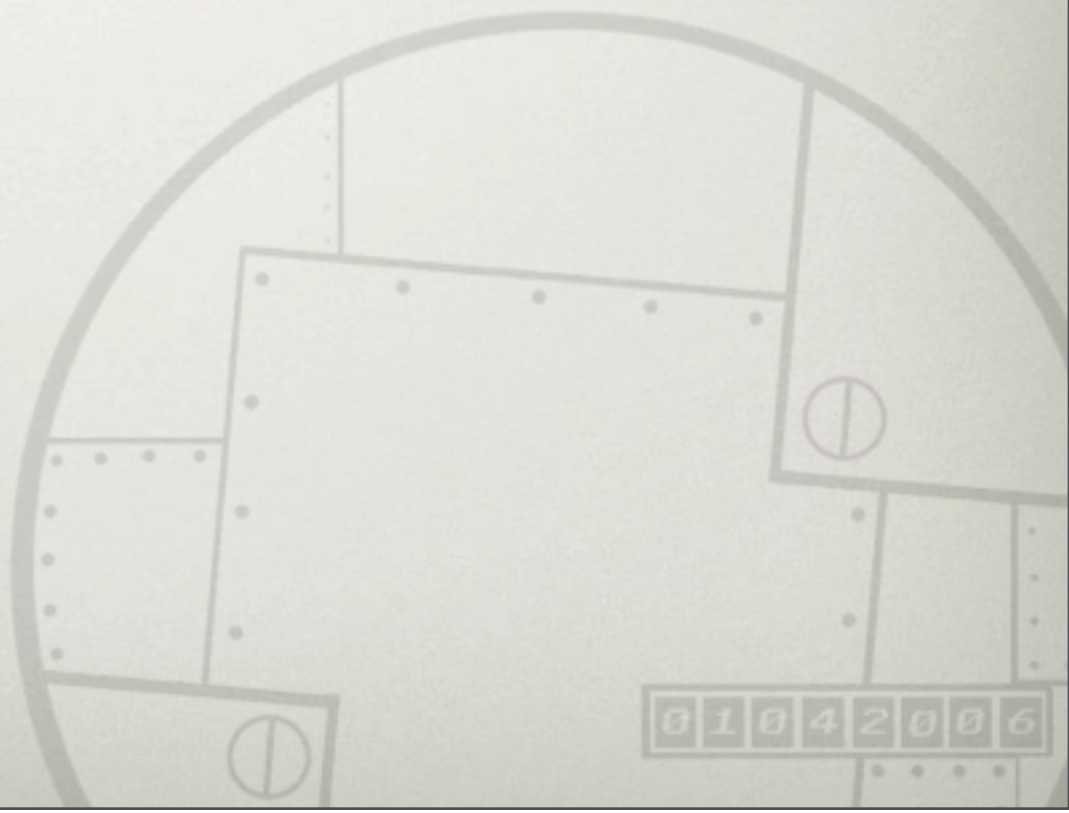
criation properties:
- date
- hardware
- ...

<XML>

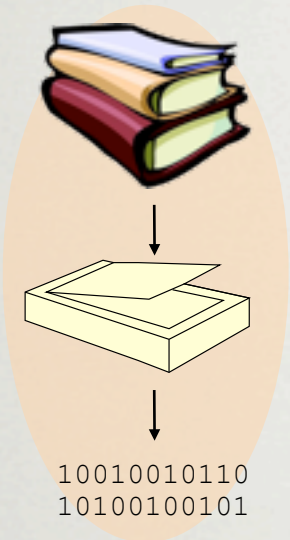
Descriptive Metadata:
- producer
- colection
- ...

Compressed File

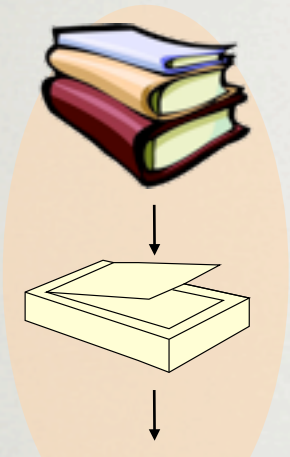
SIP Structure (+complex)

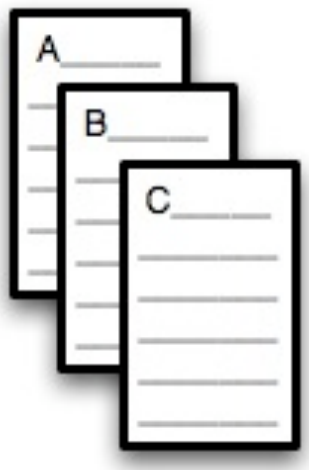


SIP Structure (+complex)



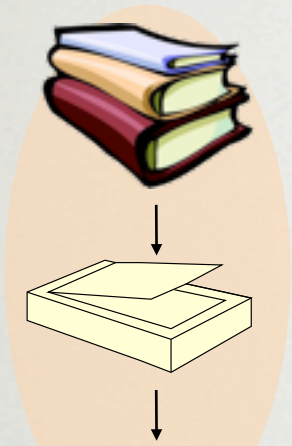
SIP Structure (+complex)

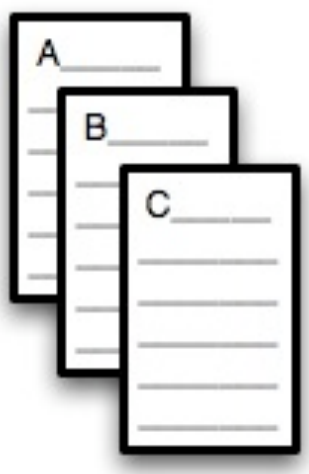




DO = Image+

SIP Structure (+complex)

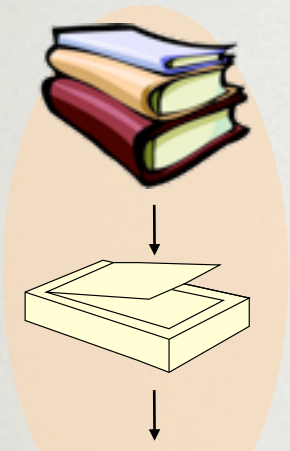


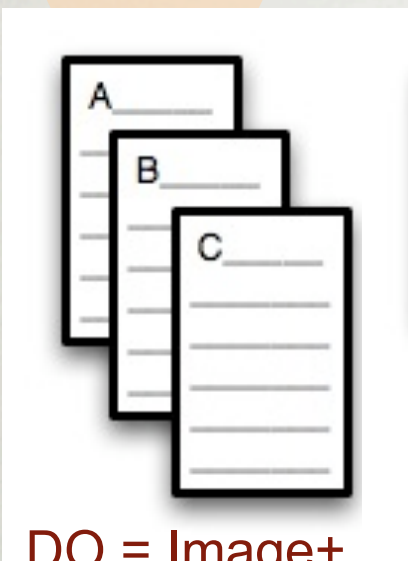


DO = Image+

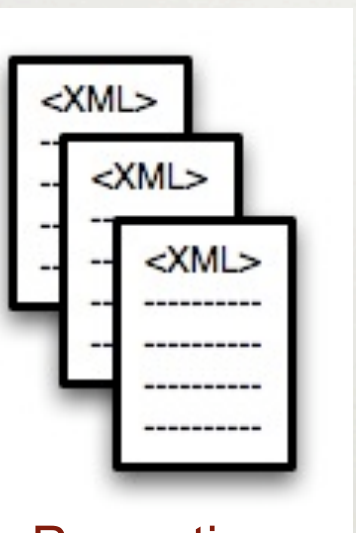
Properties

SIP Structure (+complex)





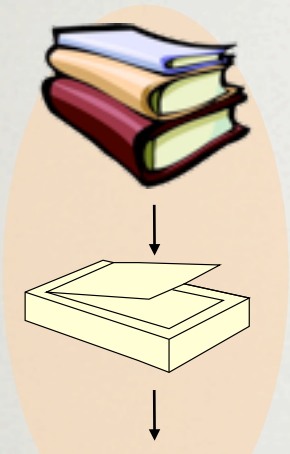
DO = Image+

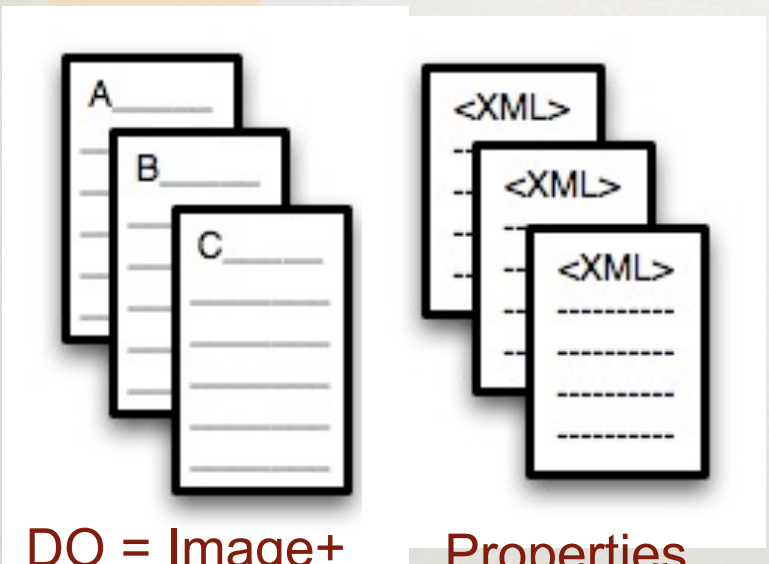


Properties

Technical Metadata

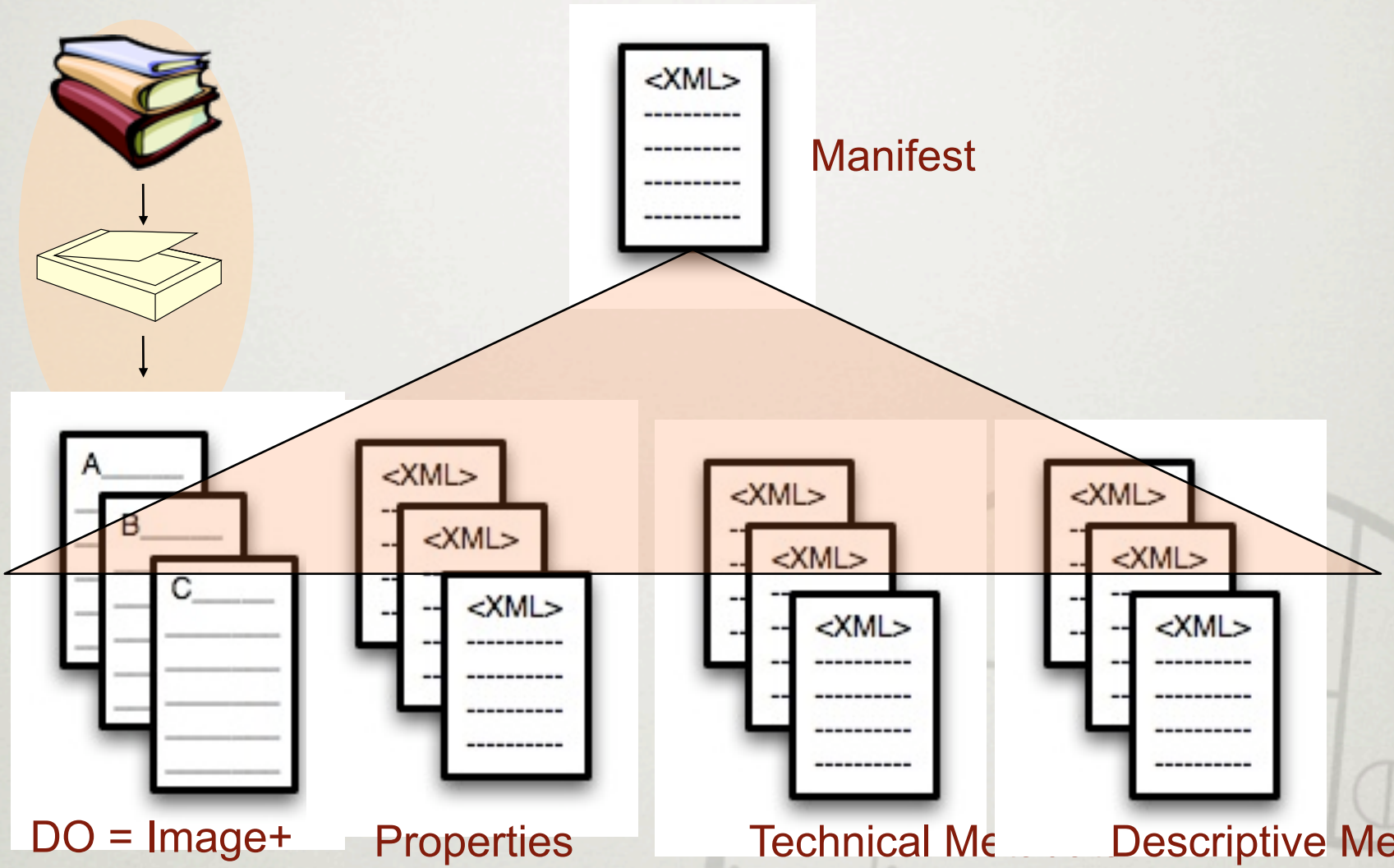
SIP Structure (+complex)



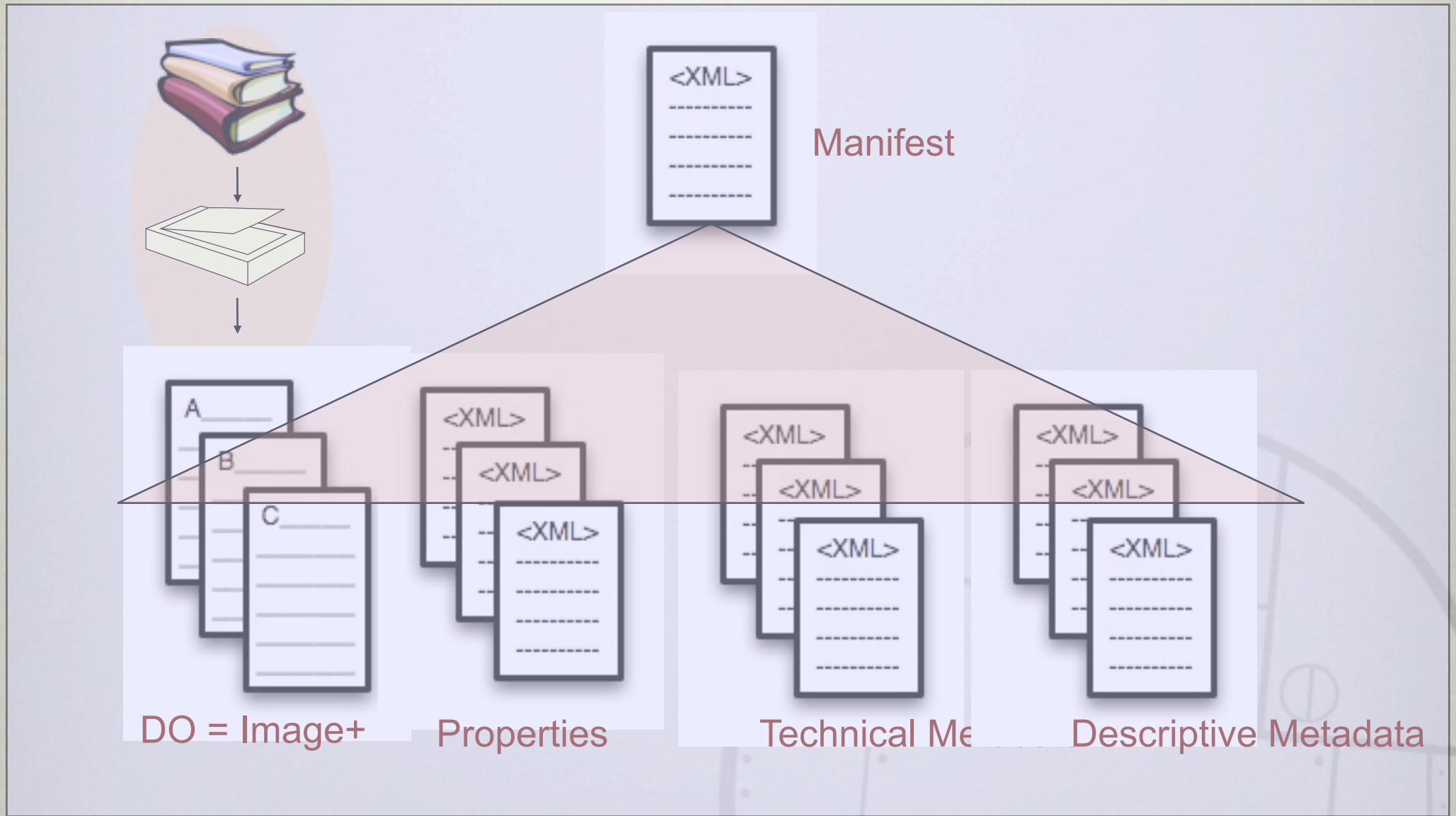


DO = Image+ Properties Technical Me Descriptive Metadata

SIP Structure (+complex)



SIP Structure (+complex)



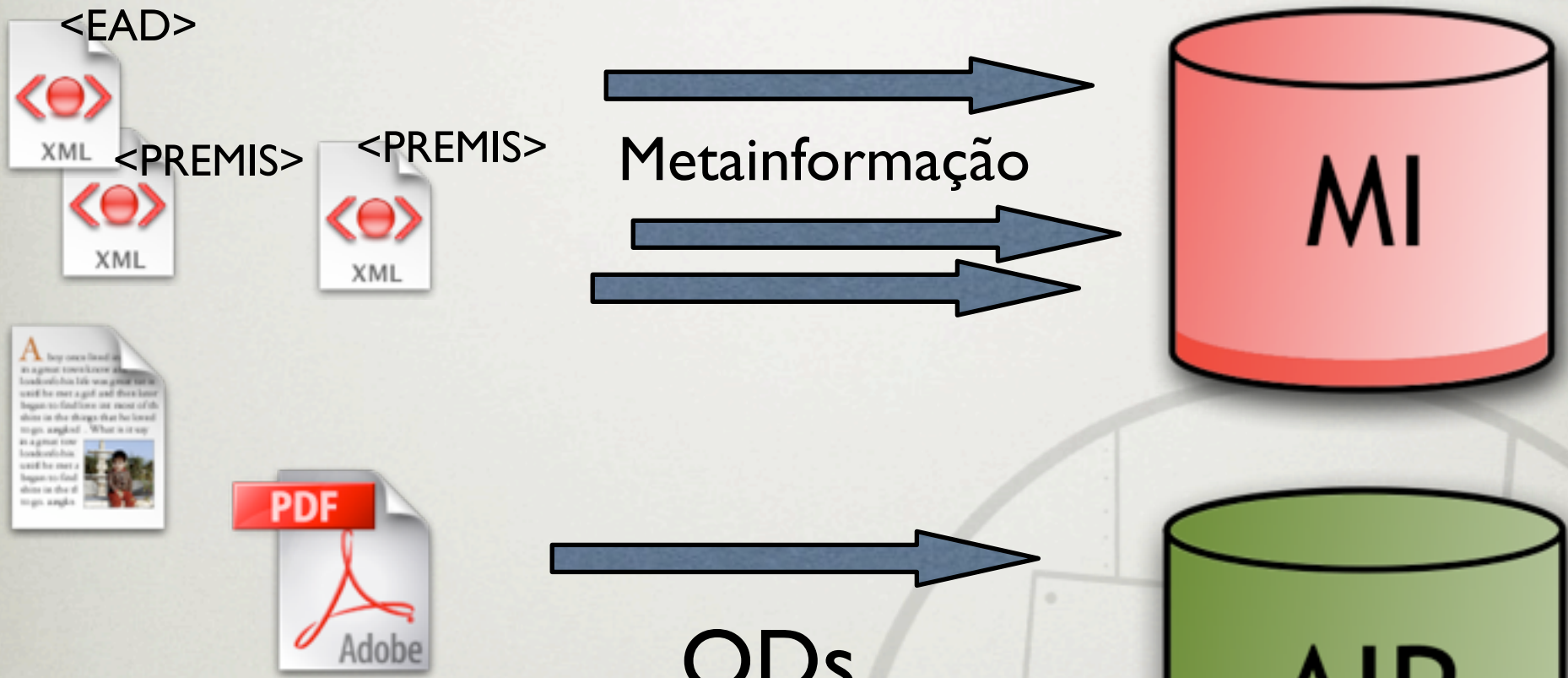
Compressed File

Workflow de Ingestão

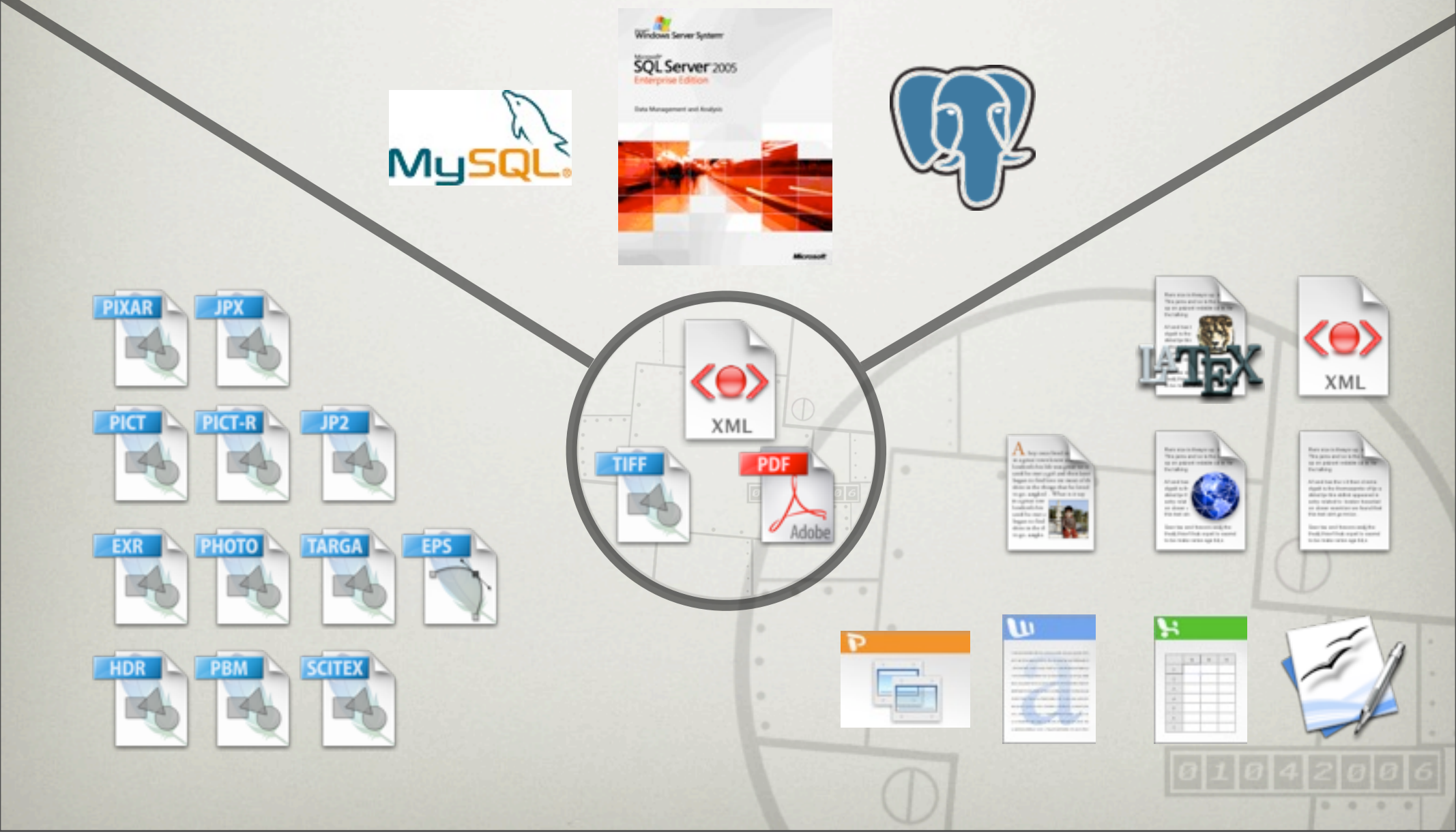
SIP → **AIP**

- verificação da integridade do pacote;
- verificação de vírus;
- geração da metainformação de preservação inicial;
- conversão num formato normalizado;
- geração de metainformação técnica;
- geração de metainformação de preservação (após normalização).

AIP (Armazenamento)

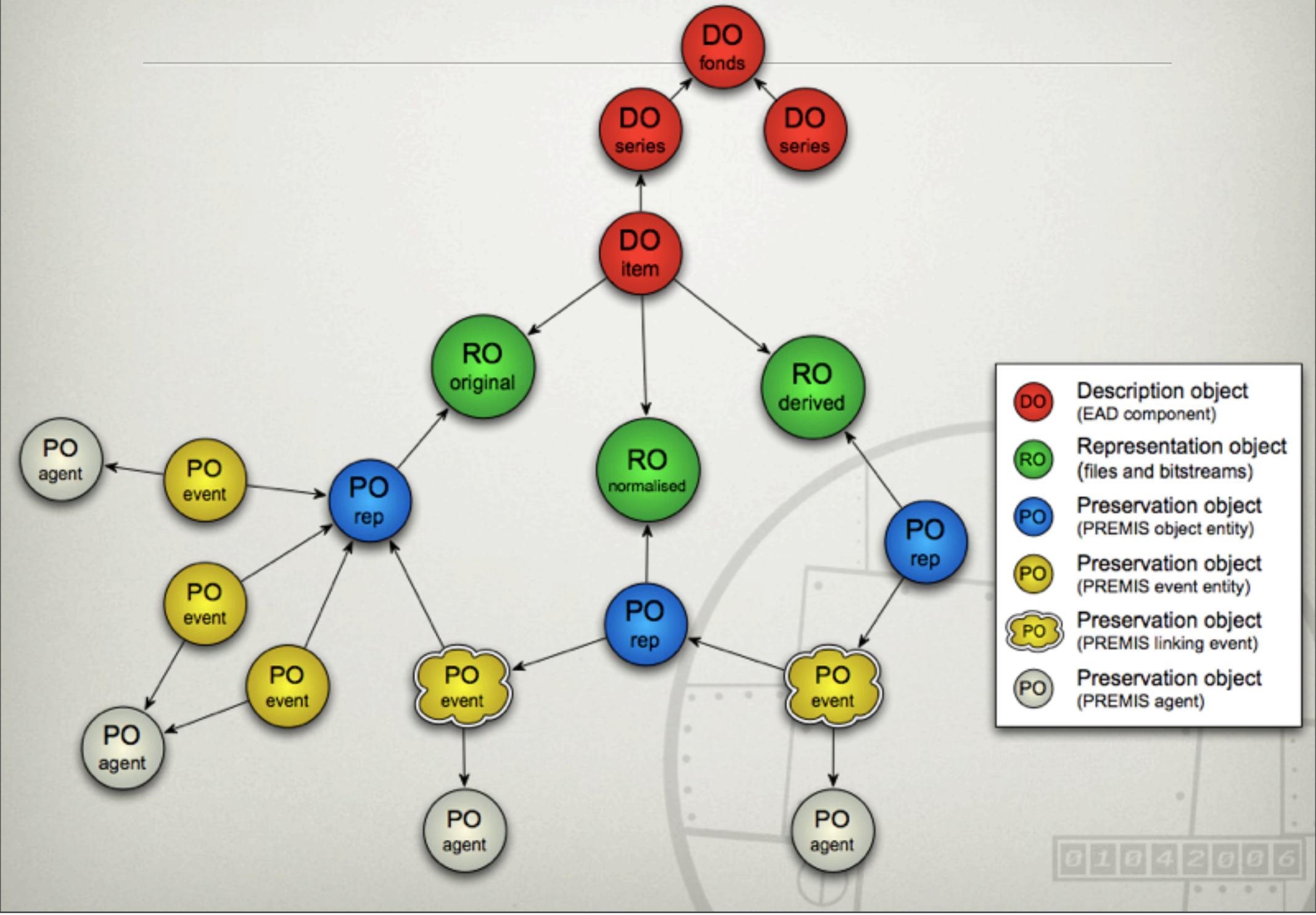


Normalização

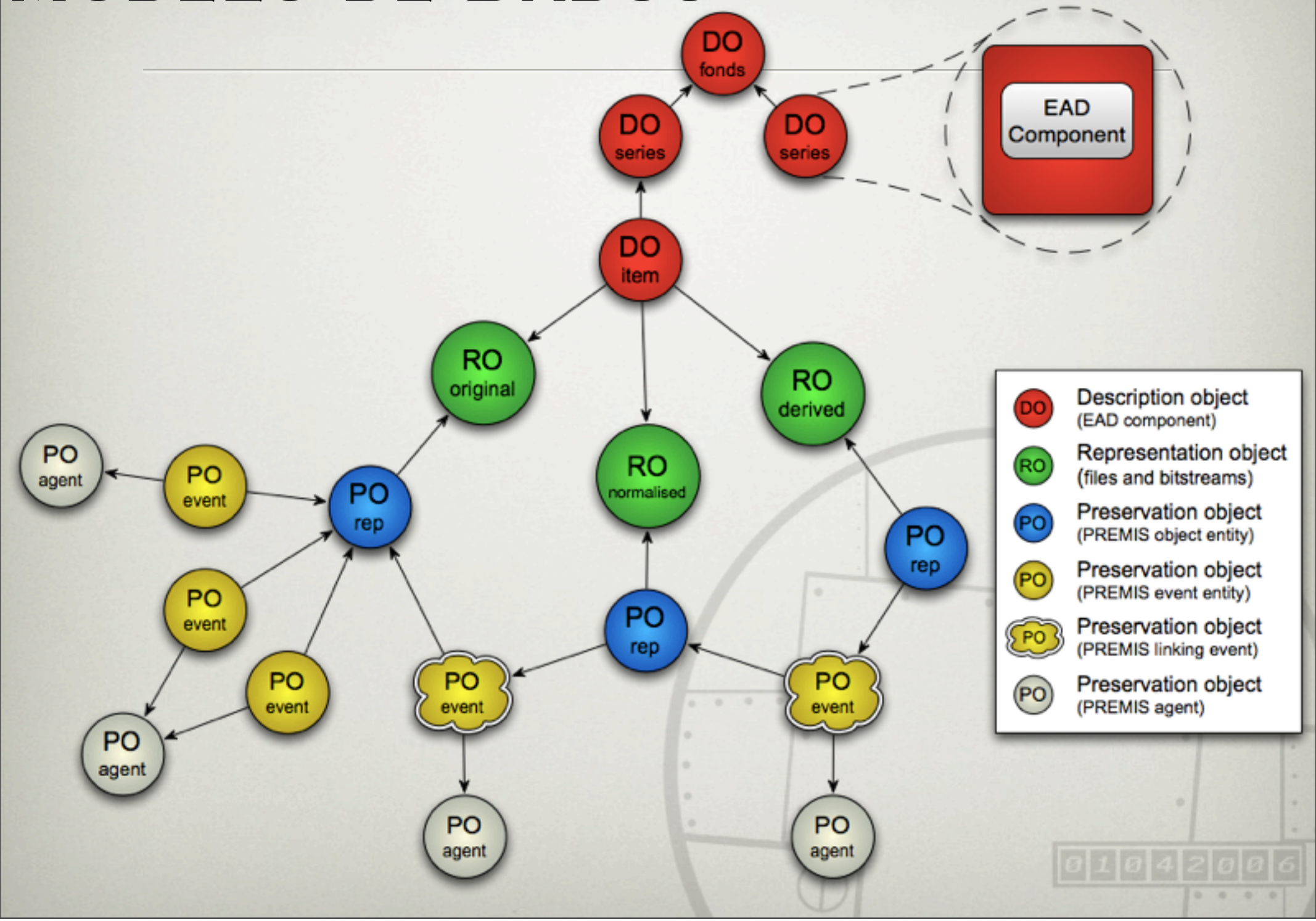


Normalização

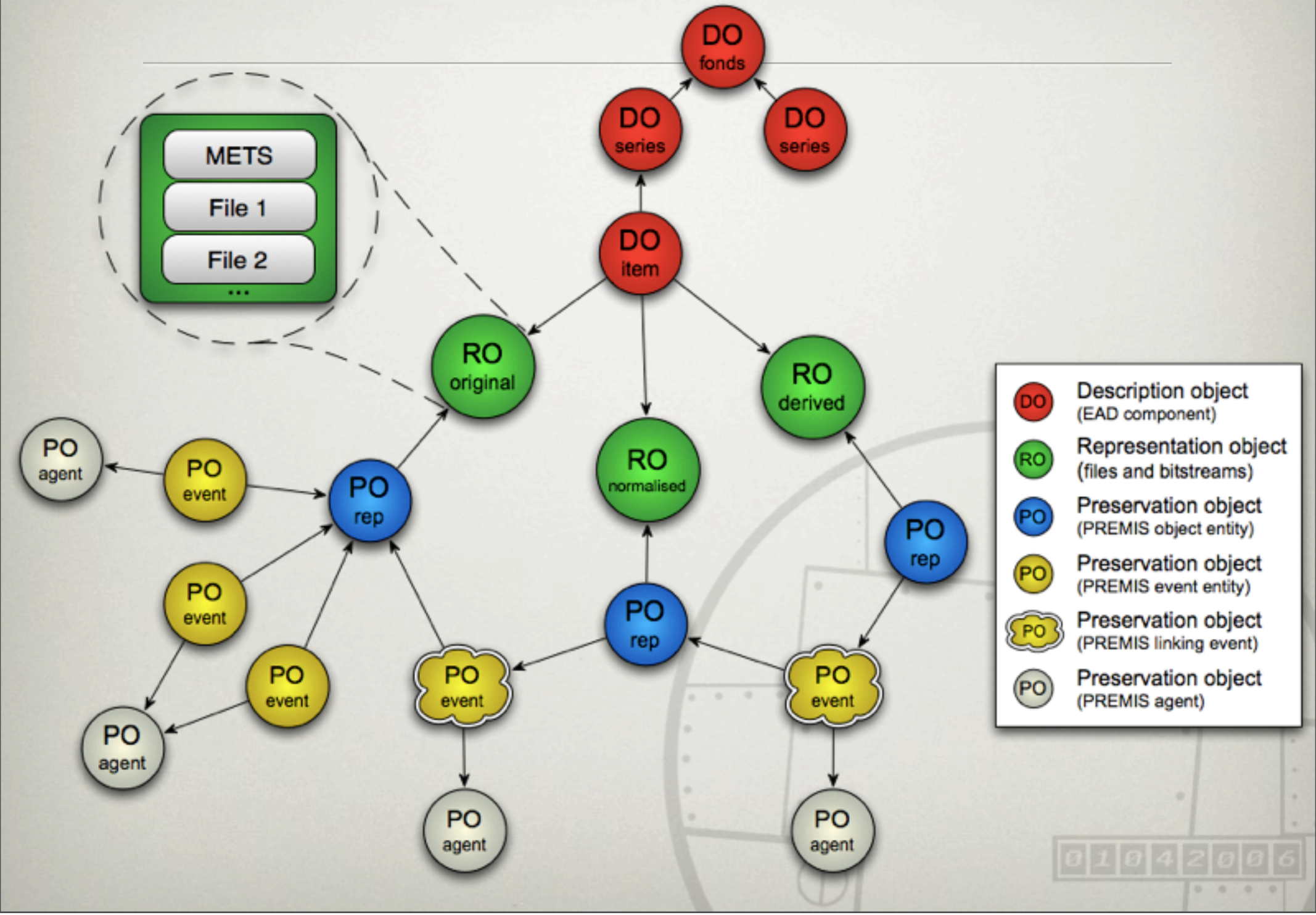
Modelo de Dados



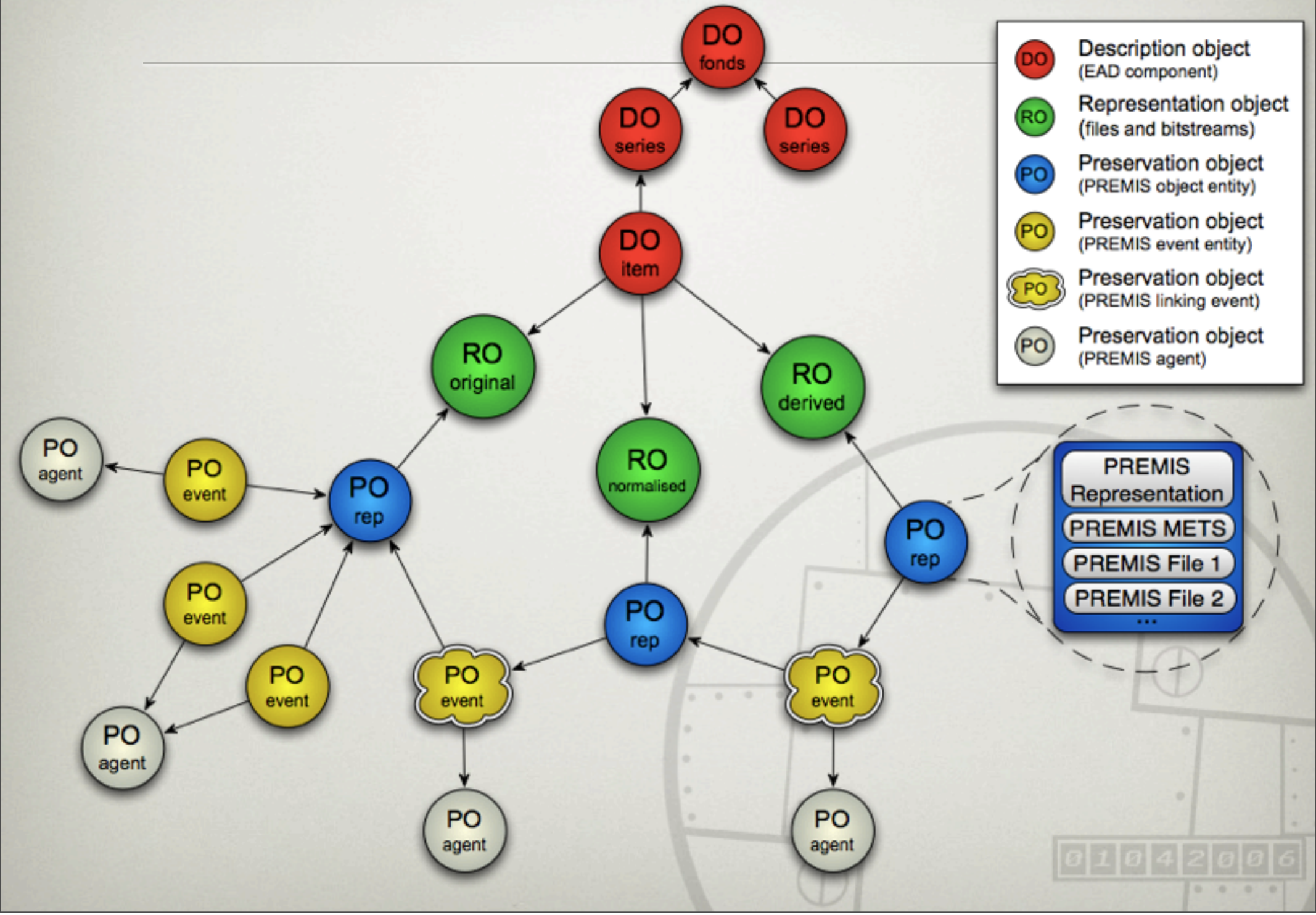
Modelo de Dados



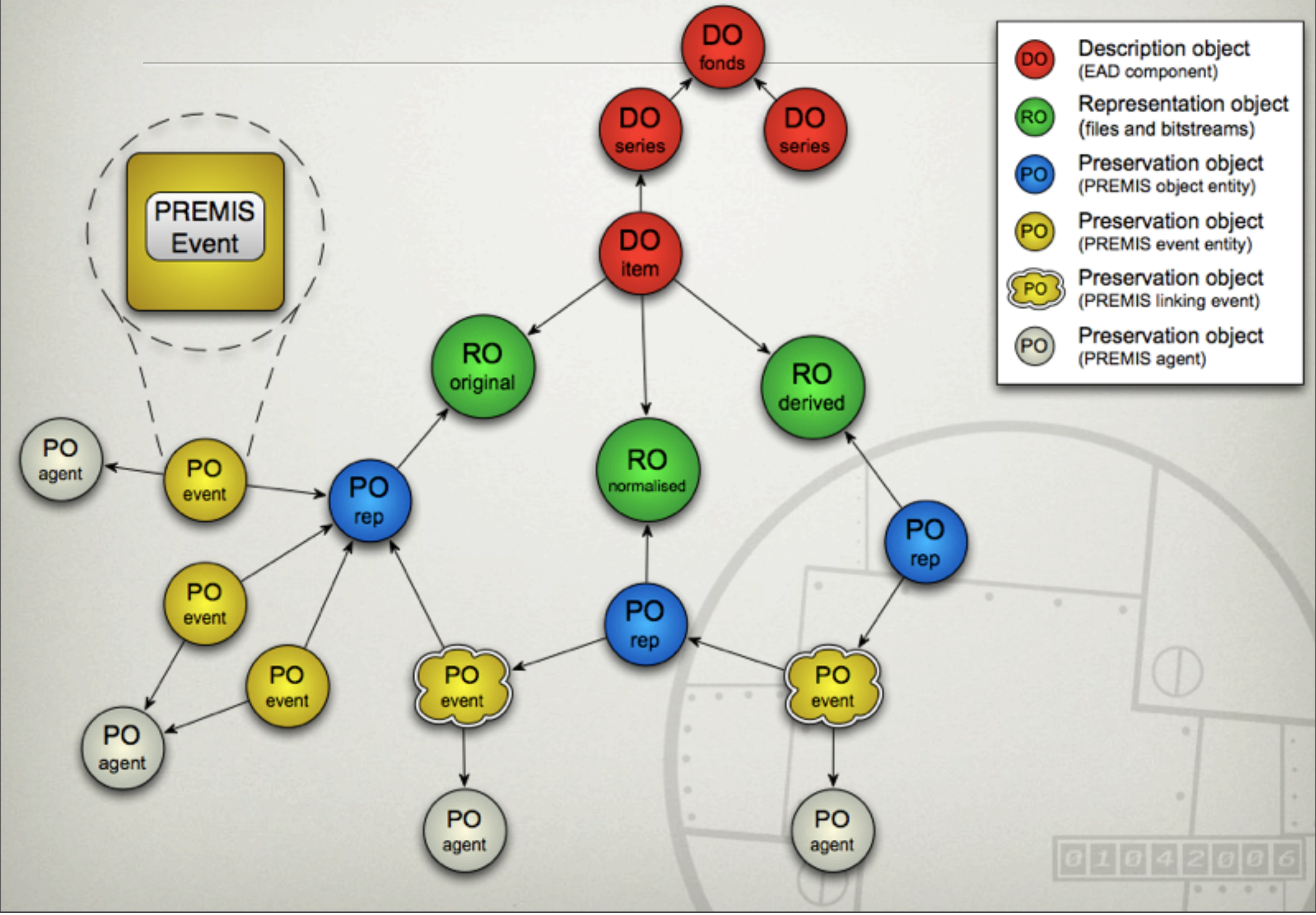
Modelo de Dados



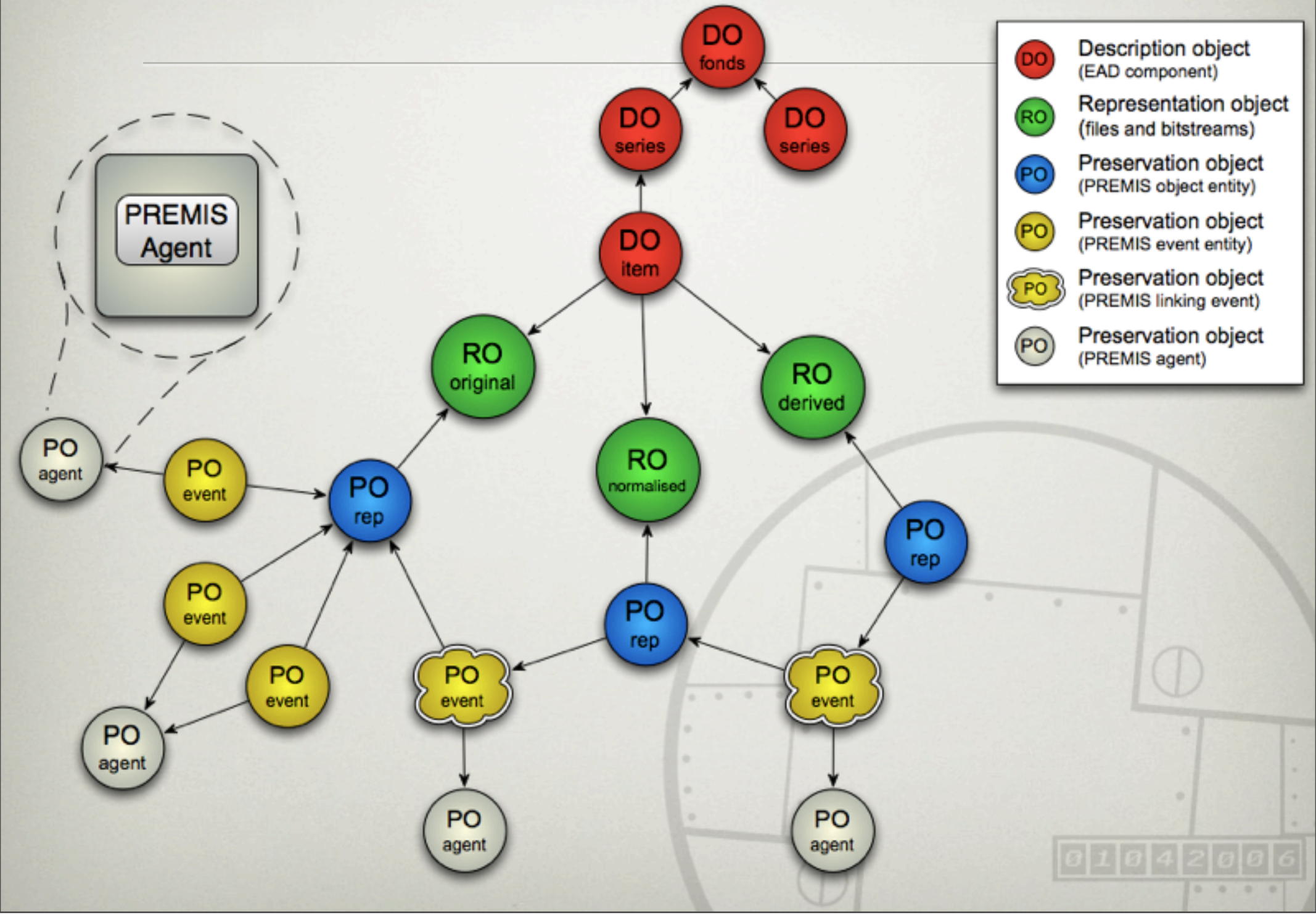
Modelo de Dados



Modelo de Dados

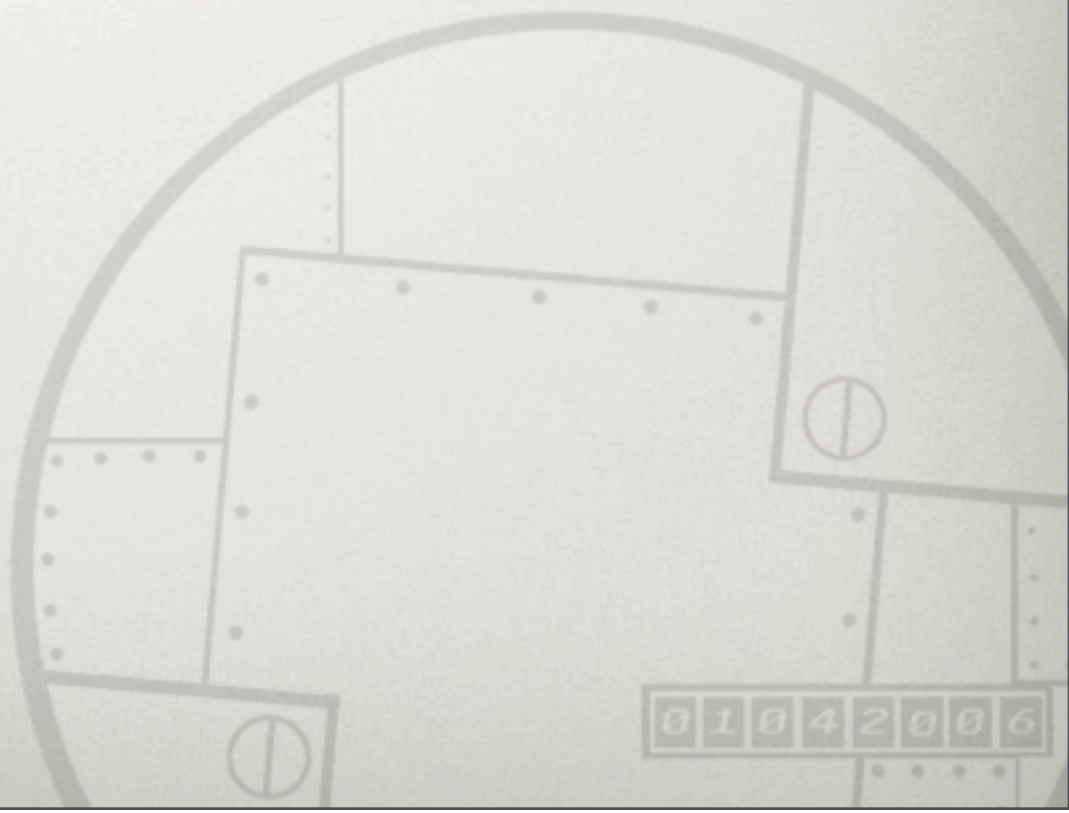


Modelo de Dados



Fases do Projecto

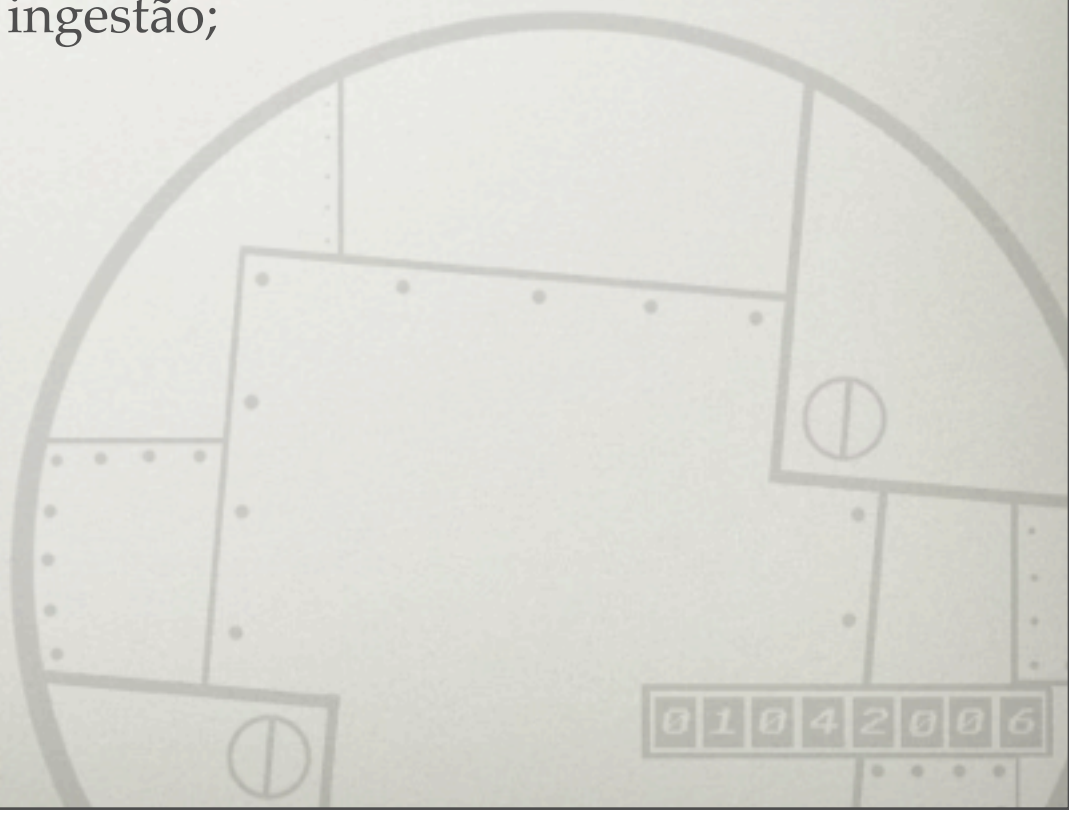
- Análise e planeamento;
- Prototipagem;
- Testes e disseminação.




Análise e
Planeamento
0 1 0 4 2 0 0 6

Requisitos

- Interface gráfica para suportar a ingestão;
- Registo de produtores;
- Ferramenta offline para produção de SIPs;
- Relatório de ingestão;
- Ingestão parcial;
- "Quarentena": cache, buffer de ingestão;
- validação do SIP;
- Identificação dos erros;
- Identificadores persistentes;
- Geração de eventos PREMIS;
- Assinatura digital de DIPs;
- etc.



FRAMEWORK


fedora

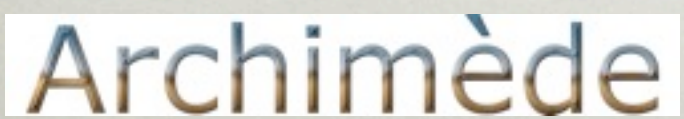

arno
academic publications online


Tools and technology
for Open Repositories

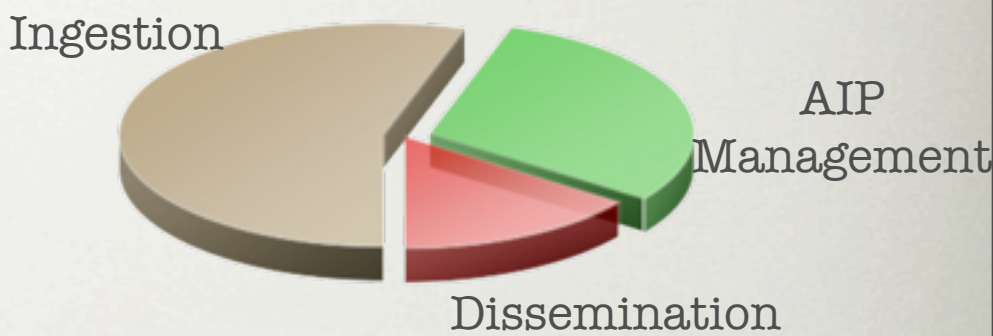

eprints

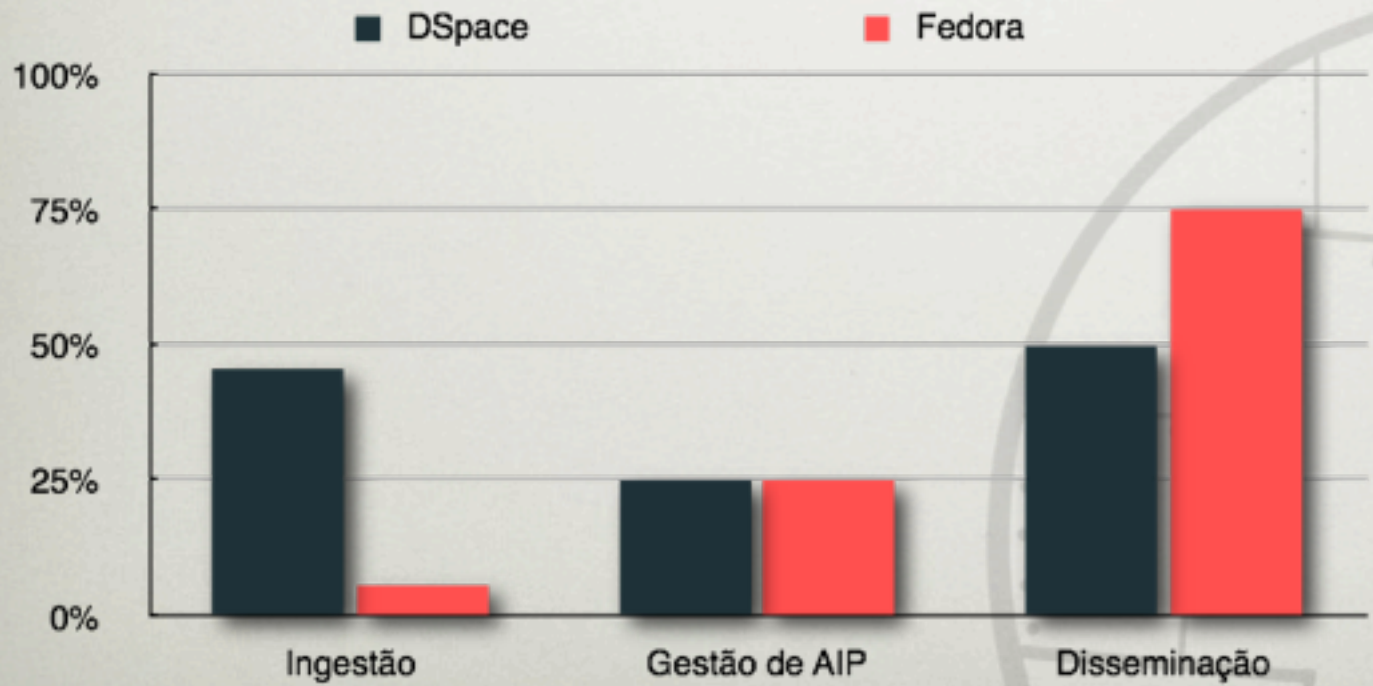

MyCoRe


DSPACE


Archimède

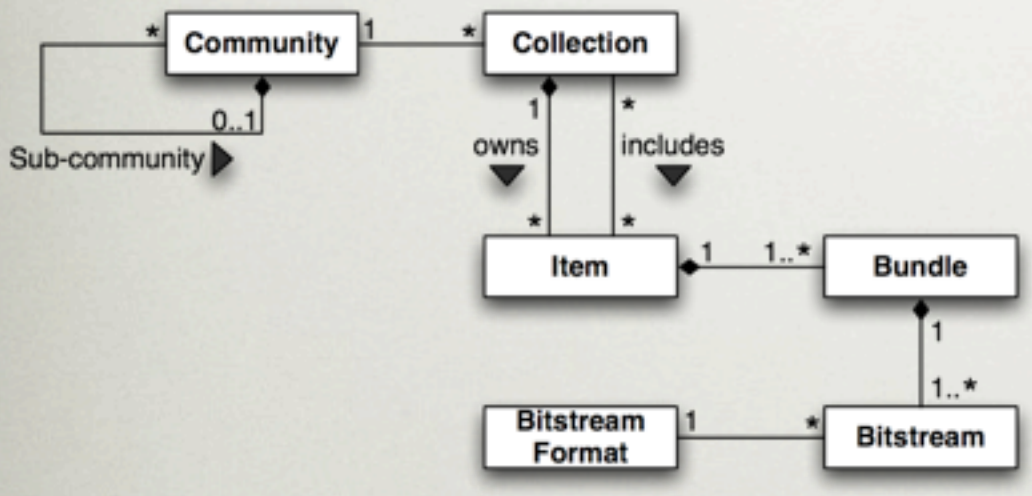
Comparação baseada nos requisitos

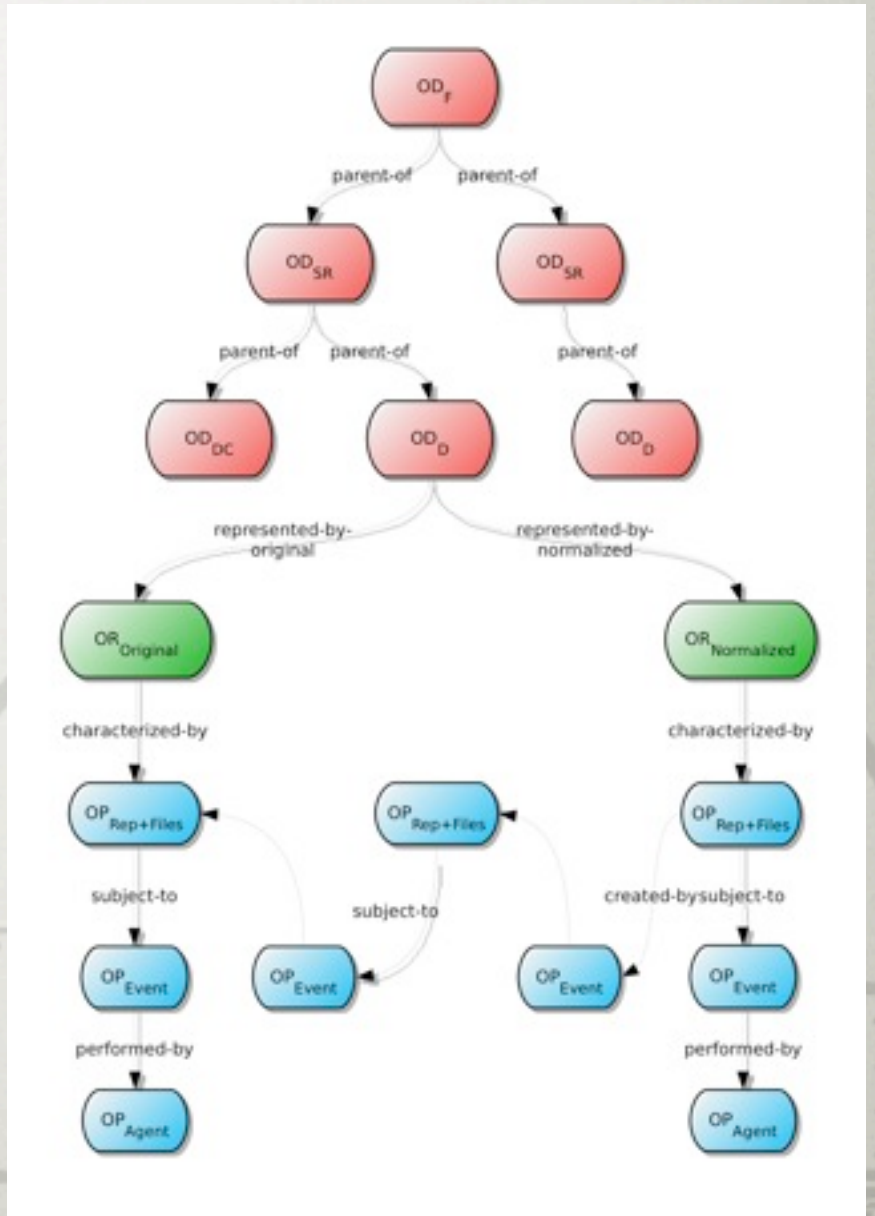




Comparação dos modelos

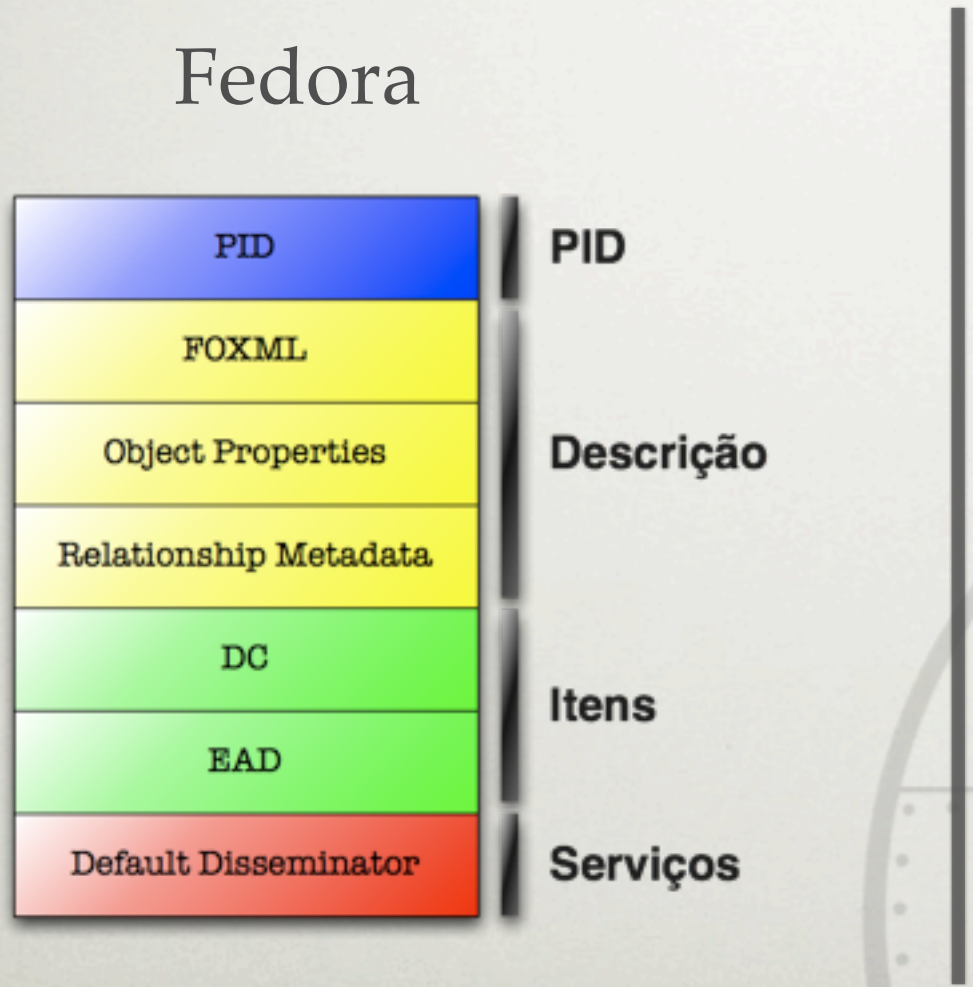
DSpace

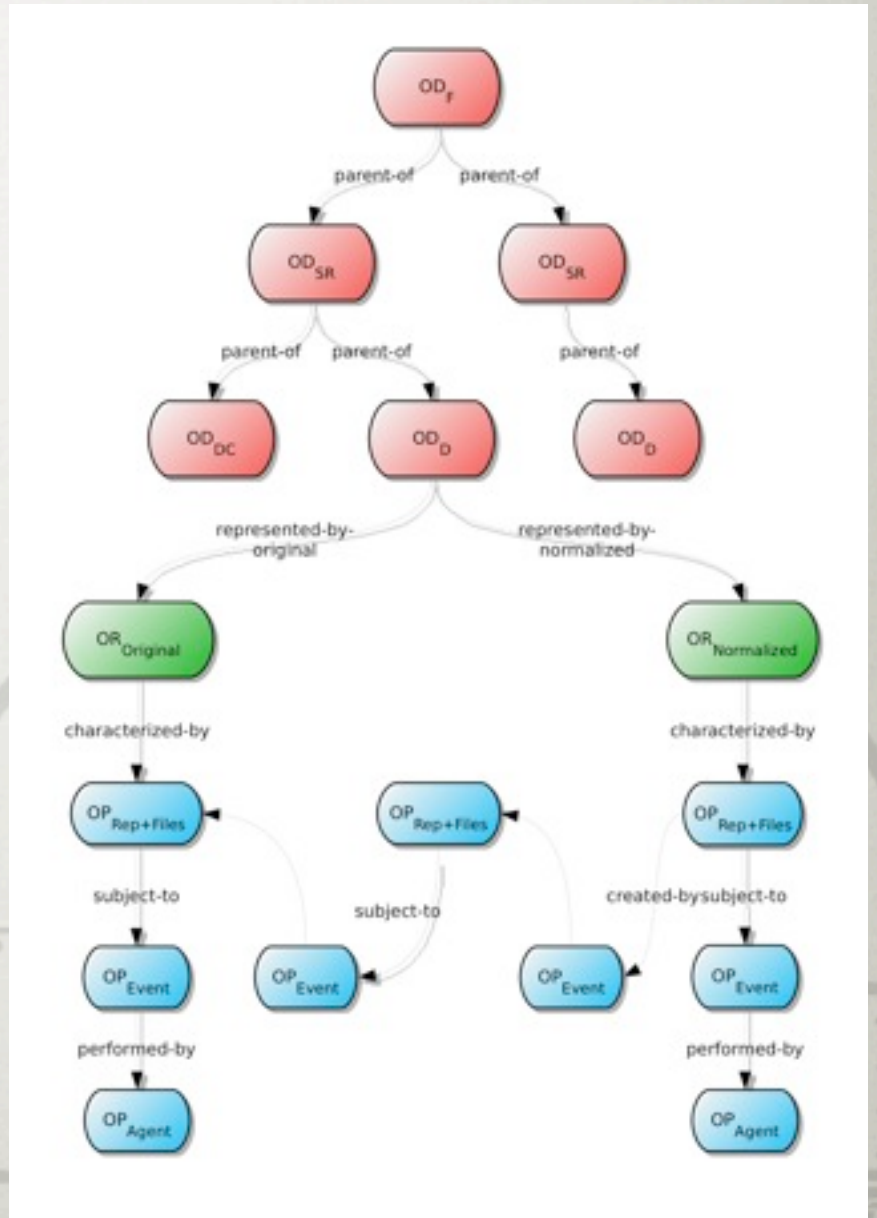




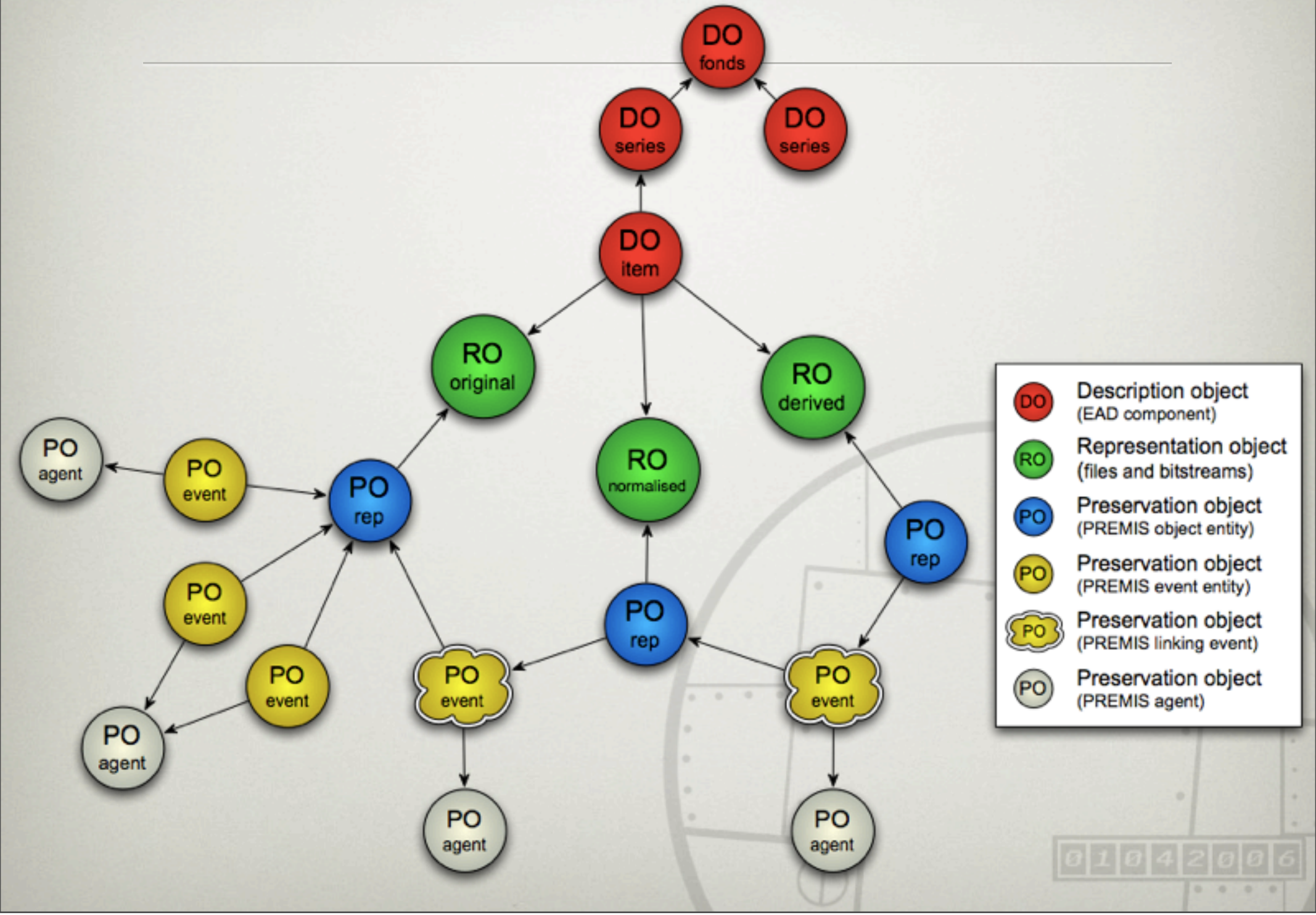
Comparação dos modelos

Fedora

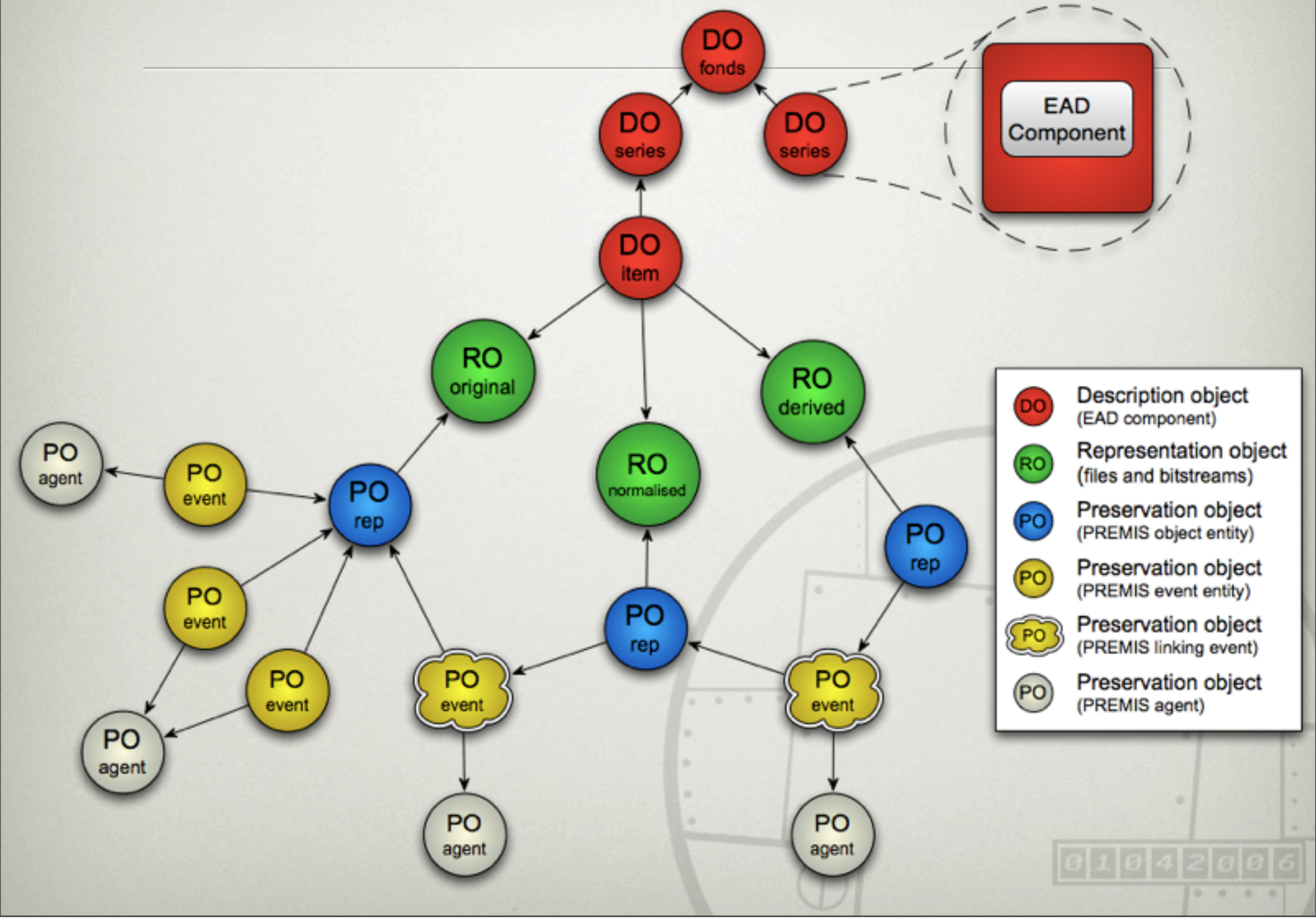




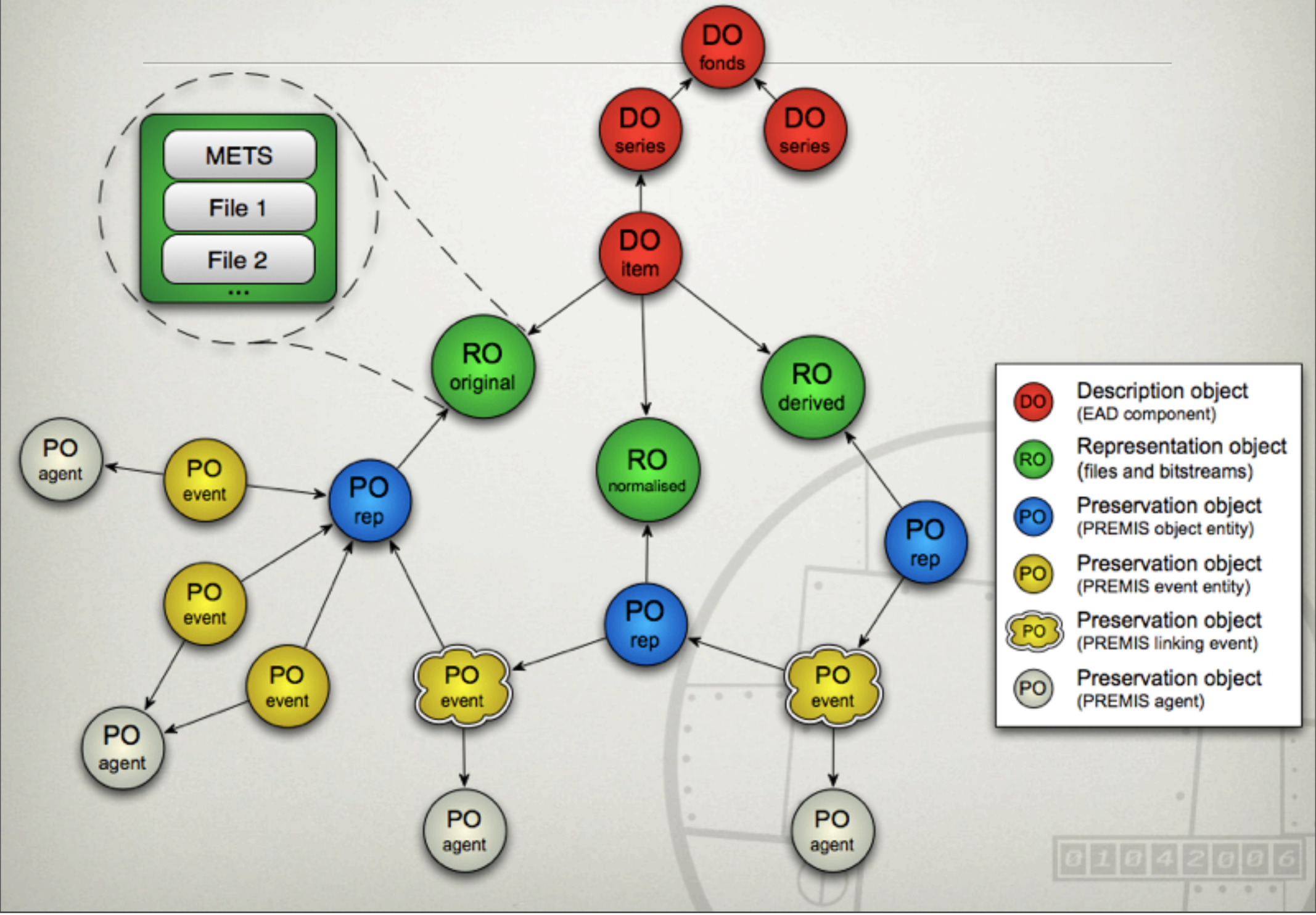
Modelo de Dados do RODA



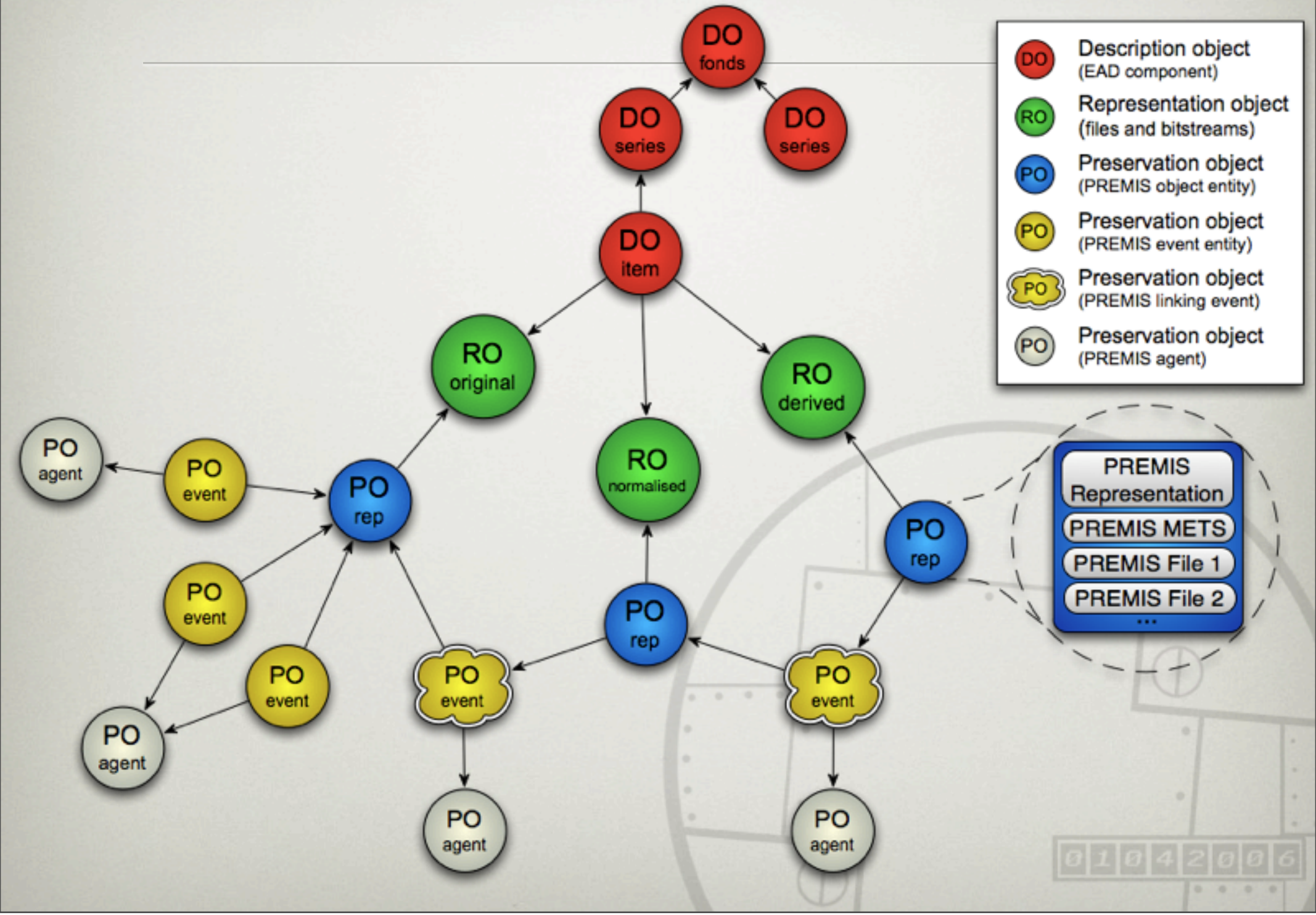
Modelo de Dados do RODA



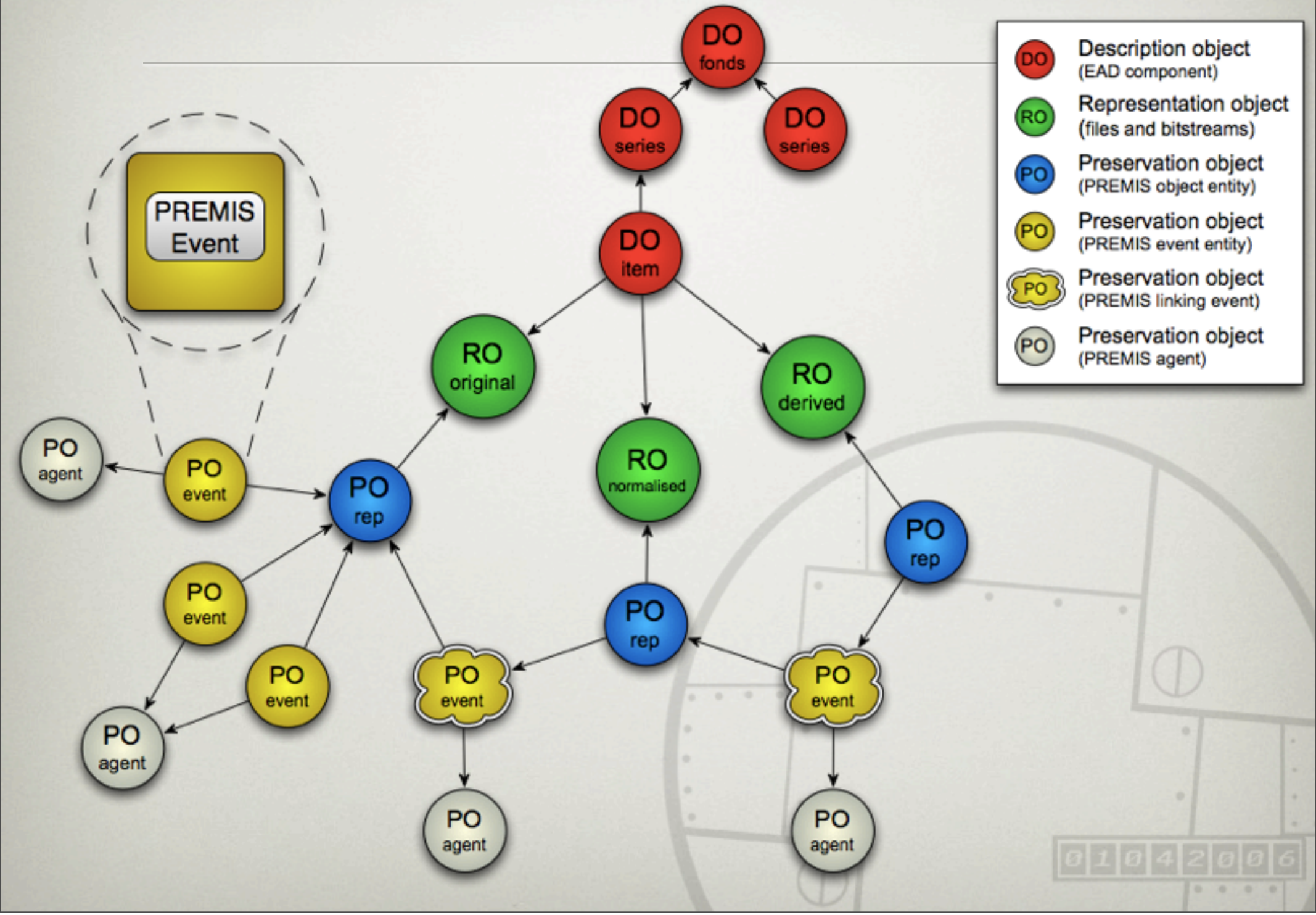
Modelo de Dados do RODA



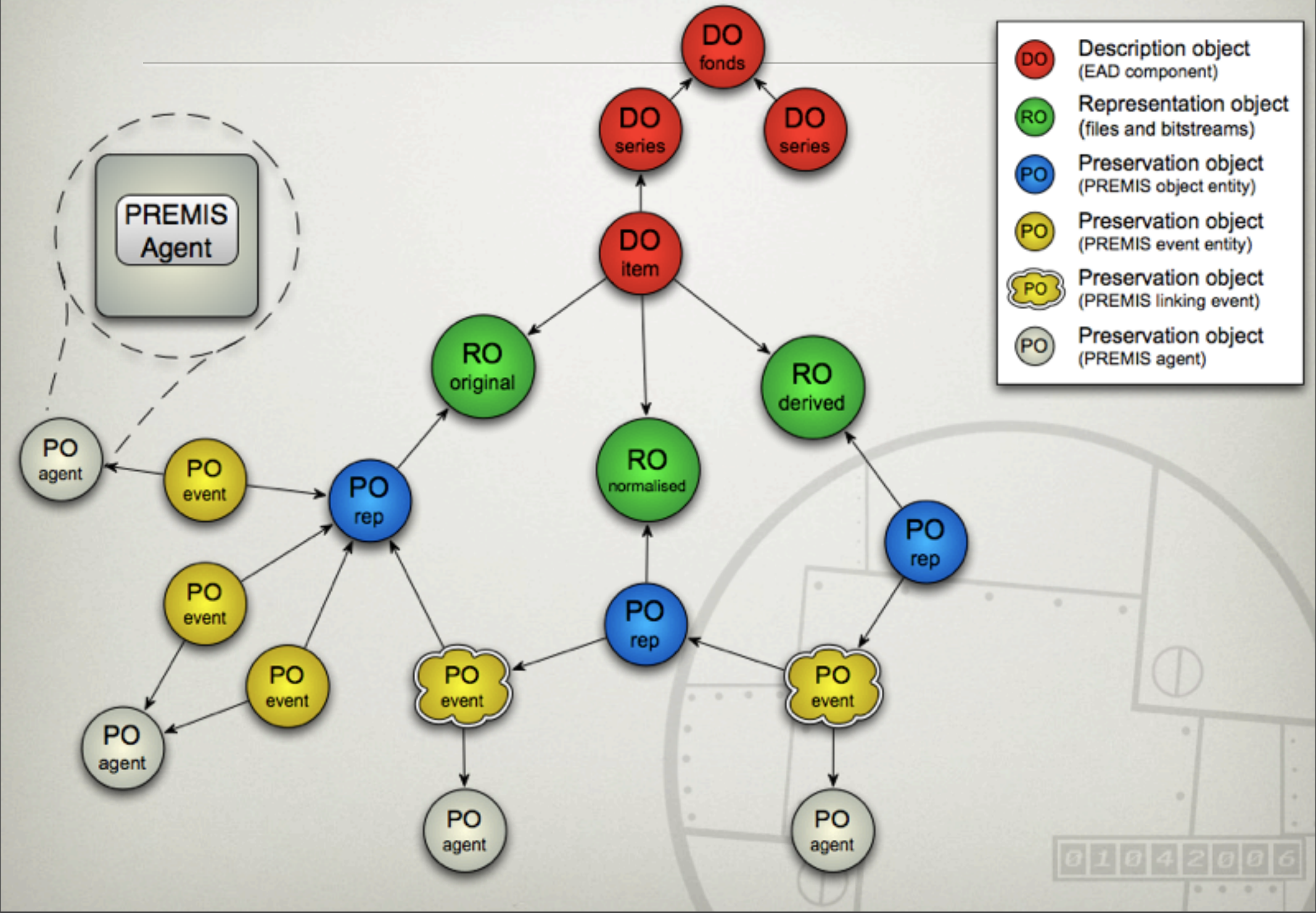
Modelo de Dados do RODA



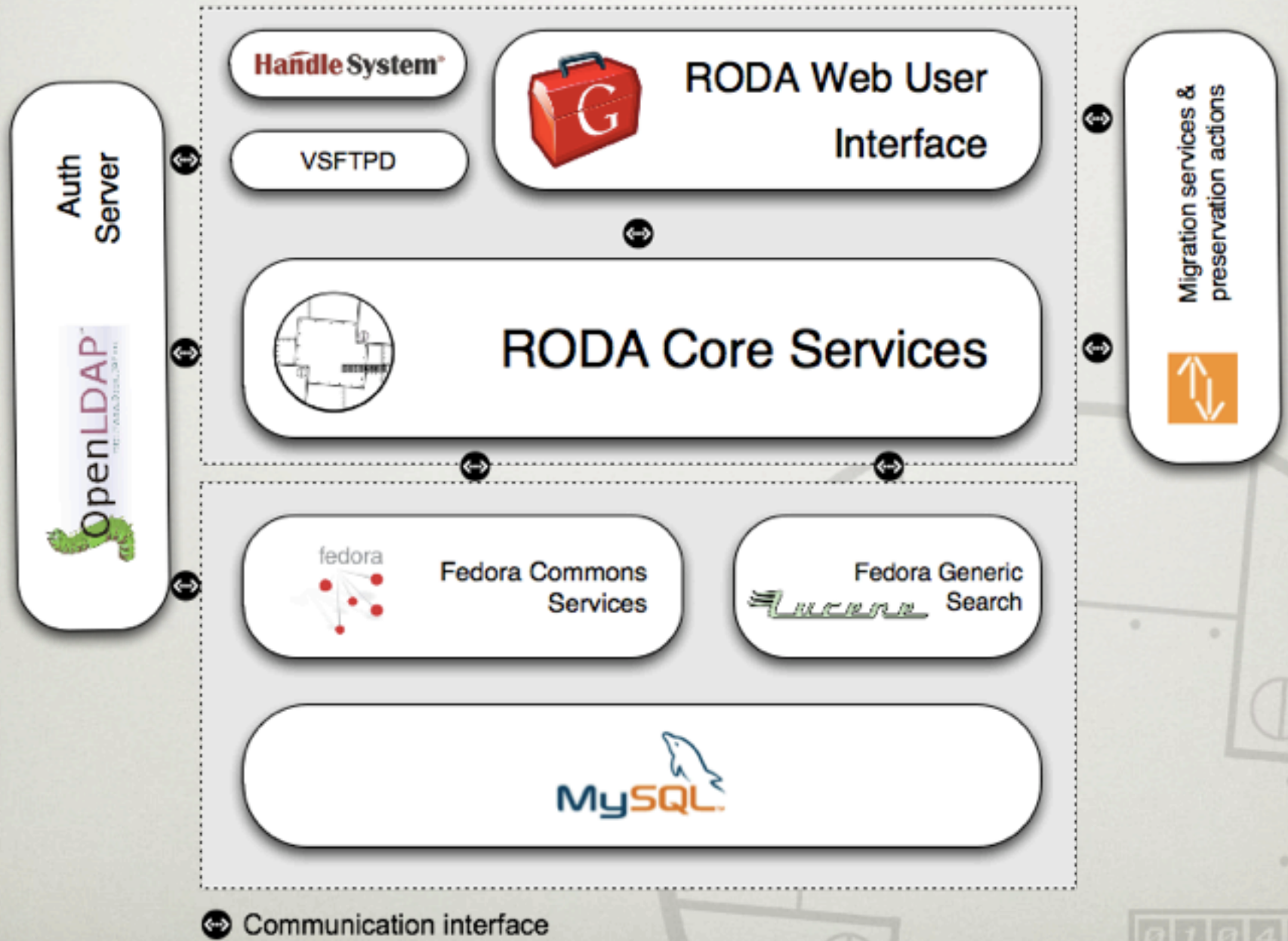
Modelo de Dados do RODA



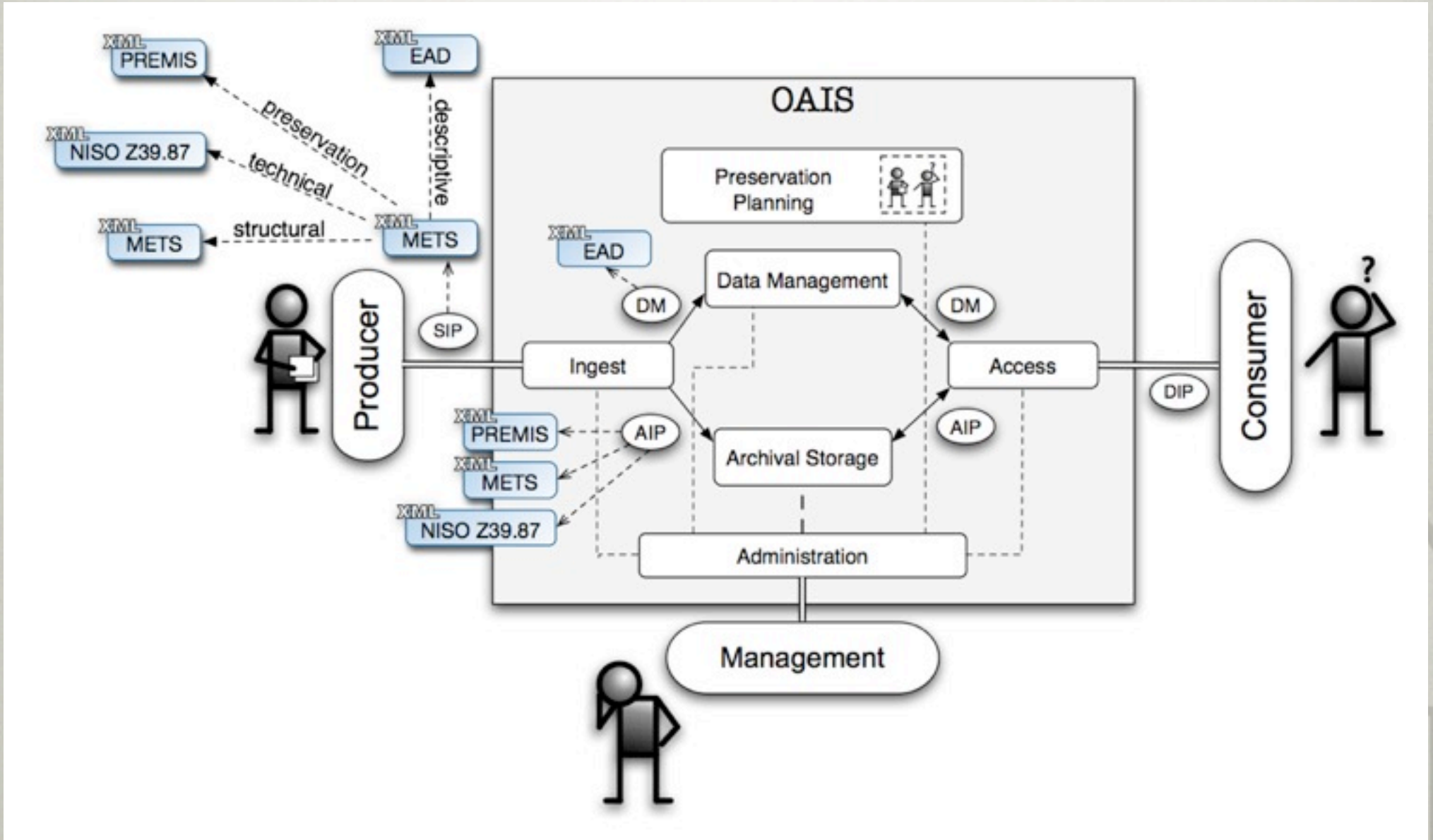
Modelo de Dados do RODA

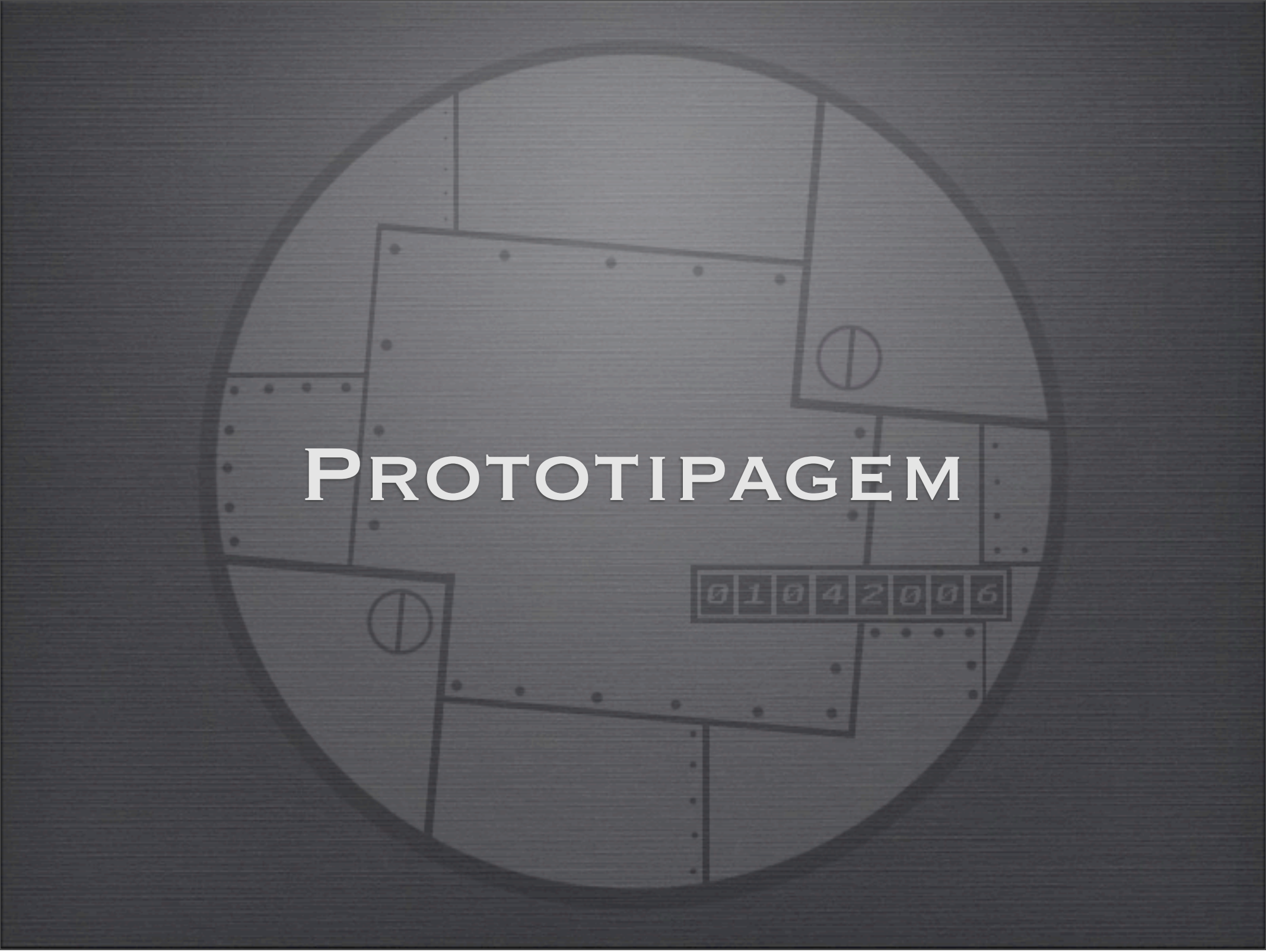


Arquitectura

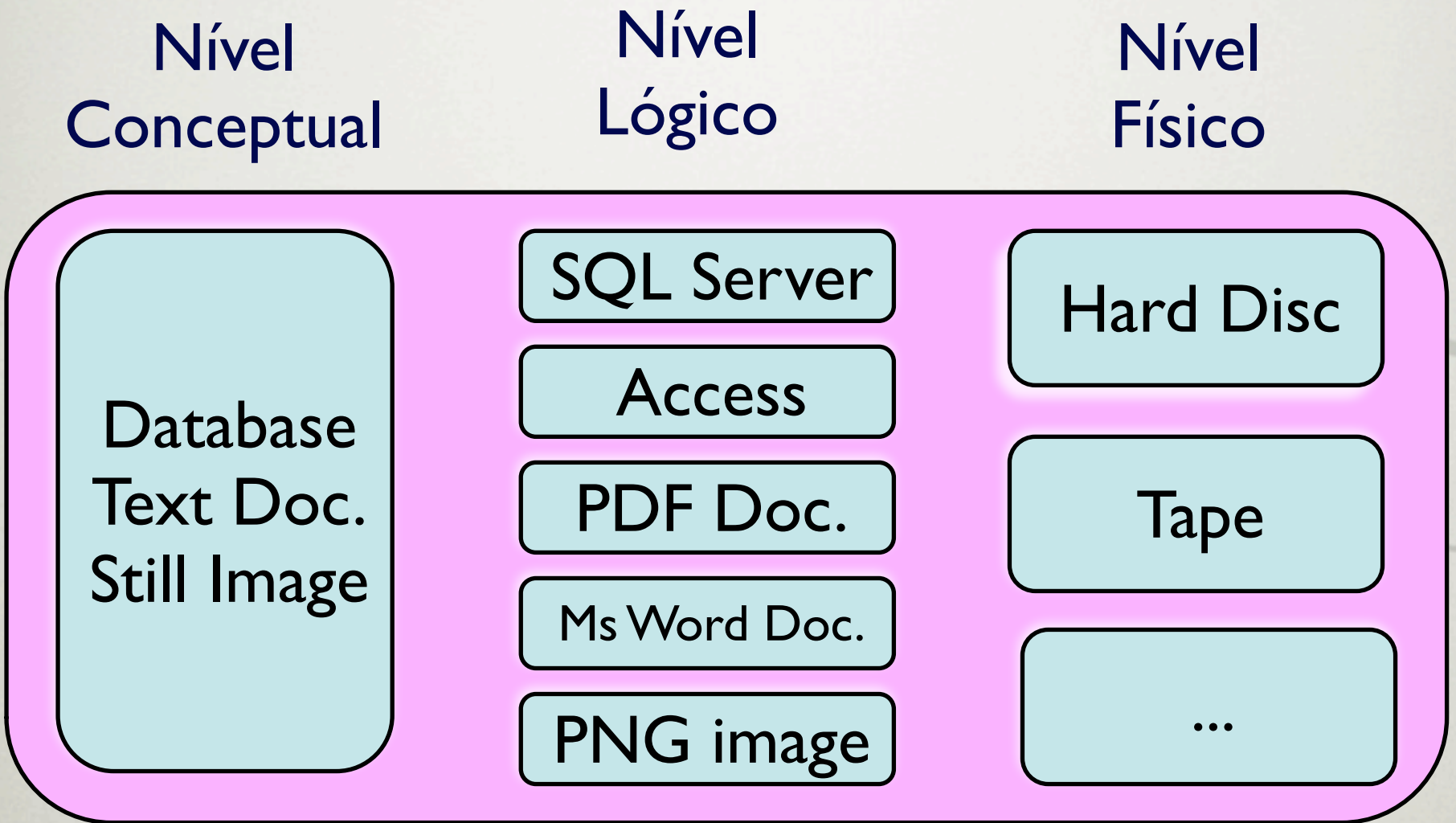


Schemas do RODA

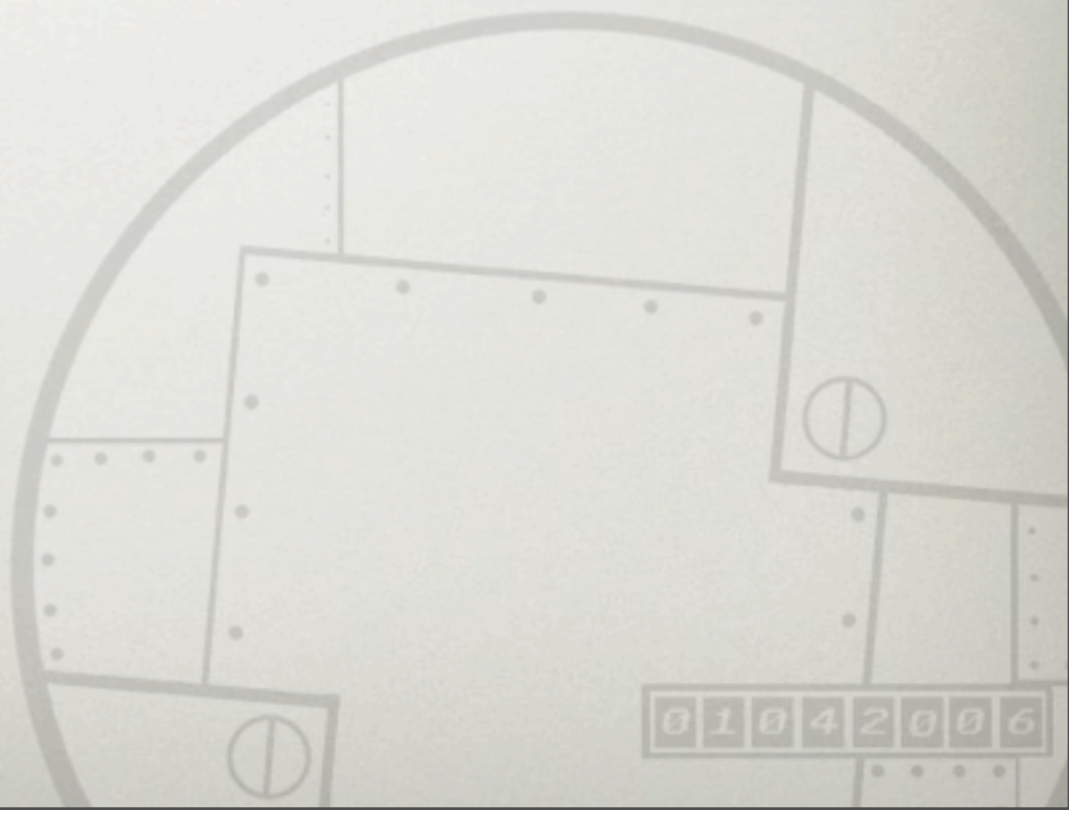



Prototipagem
0 1 0 4 2 0 0 6

Preservaçãodo Objecto Conceptual



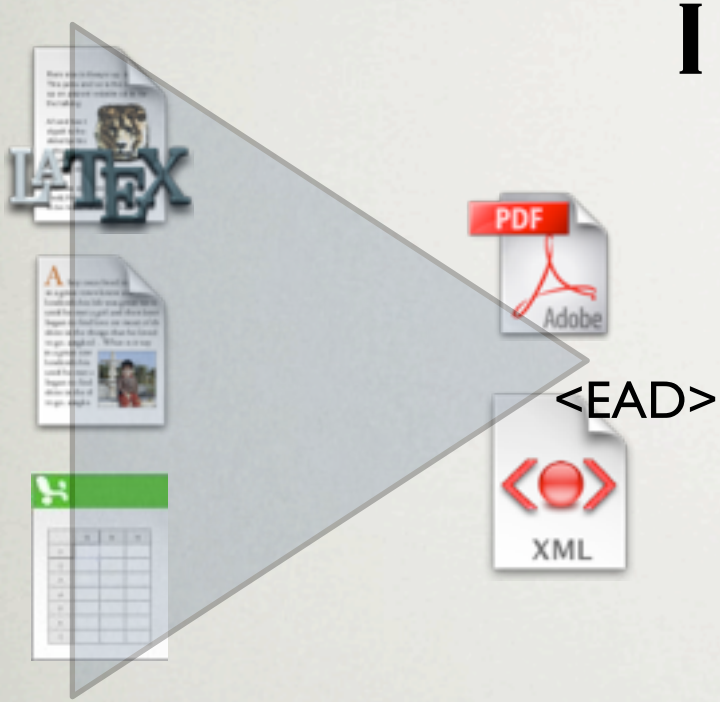
Documentos de texto e Imagens



Documentos de texto e Imagens

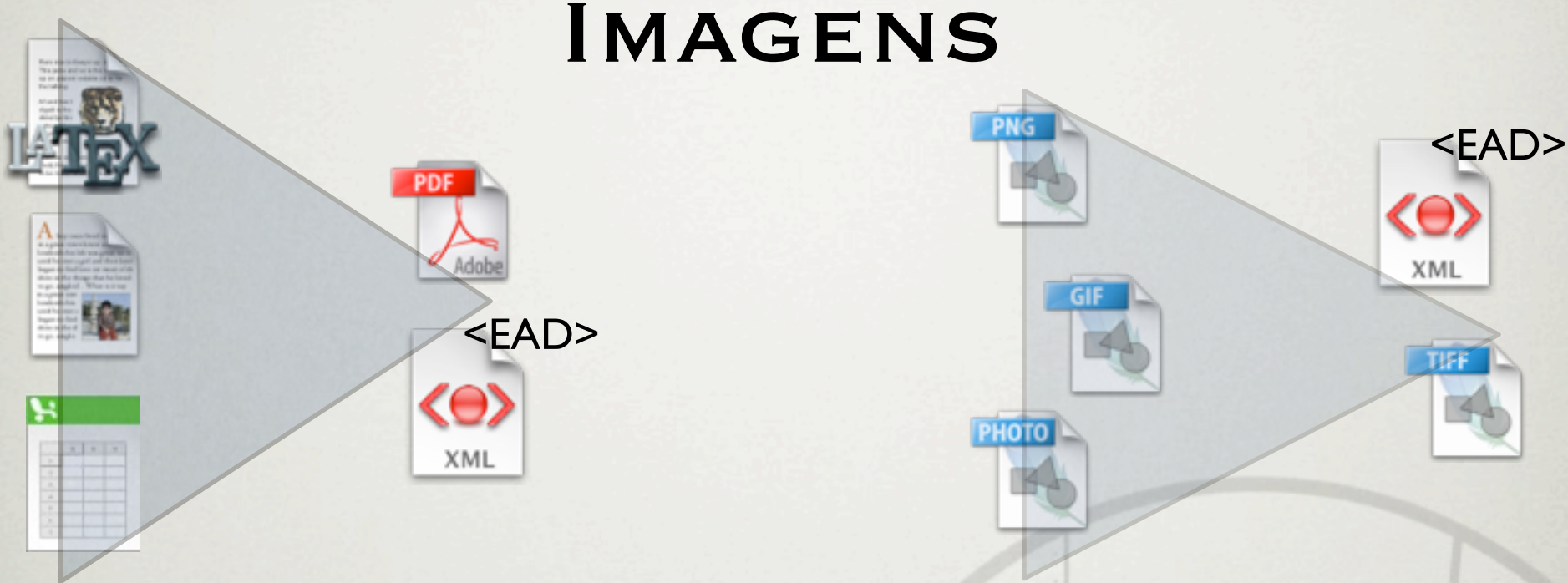
- O EAD retem a maior parte das propriedades significativas: proveniência, história, contexto, etc;
- O conteúdo é mantido num formato normalizado: PDF e TIFF não comprimido.

Documentos de texto e Imagens



- O EAD retem a maior parte das propriedades significativas: proveniência, história, contexto, etc;
- O conteúdo é mantido num formato normalizado: PDF e TIFF não comprimido.

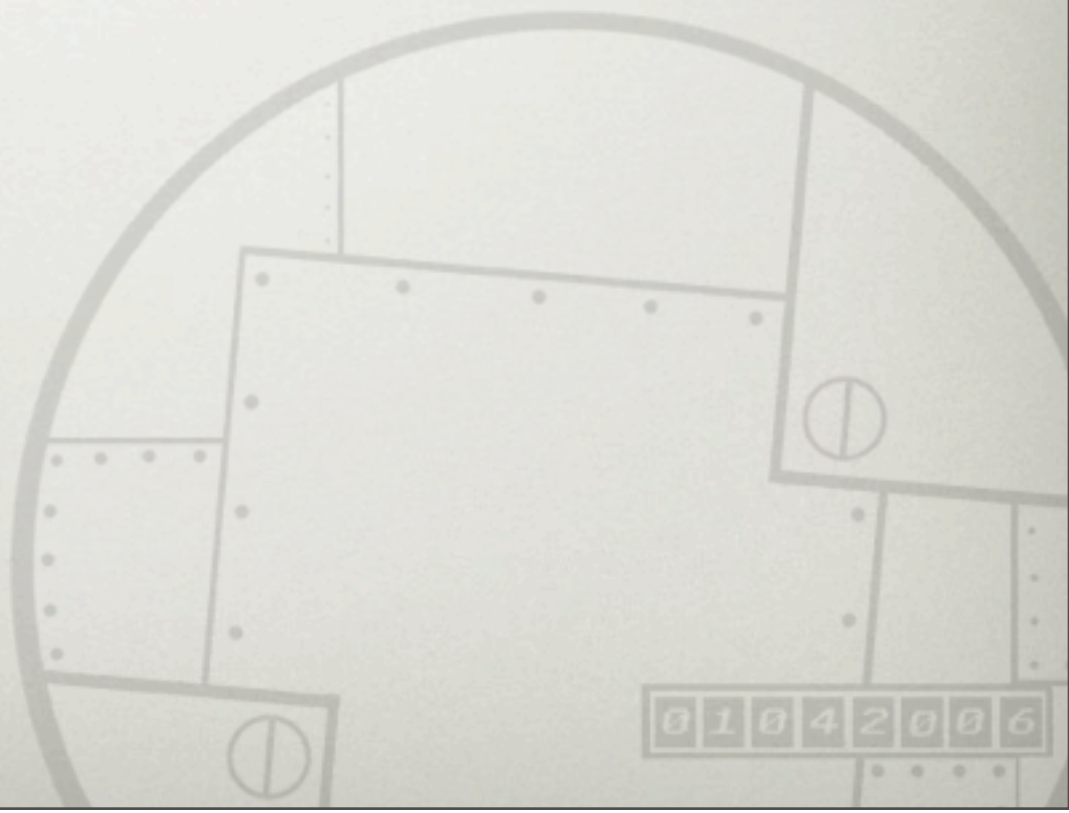
Documentos de texto e Imagens



- O EAD retem a maior parte das propriedades significativas: proveniência, história, contexto, etc;
- O conteúdo é mantido num formato normalizado: PDF e TIFF não comprimido.

Databases

- Dados?
- Estrutura?
- Views?
- Reports?
- Stored Procedures?
- ...



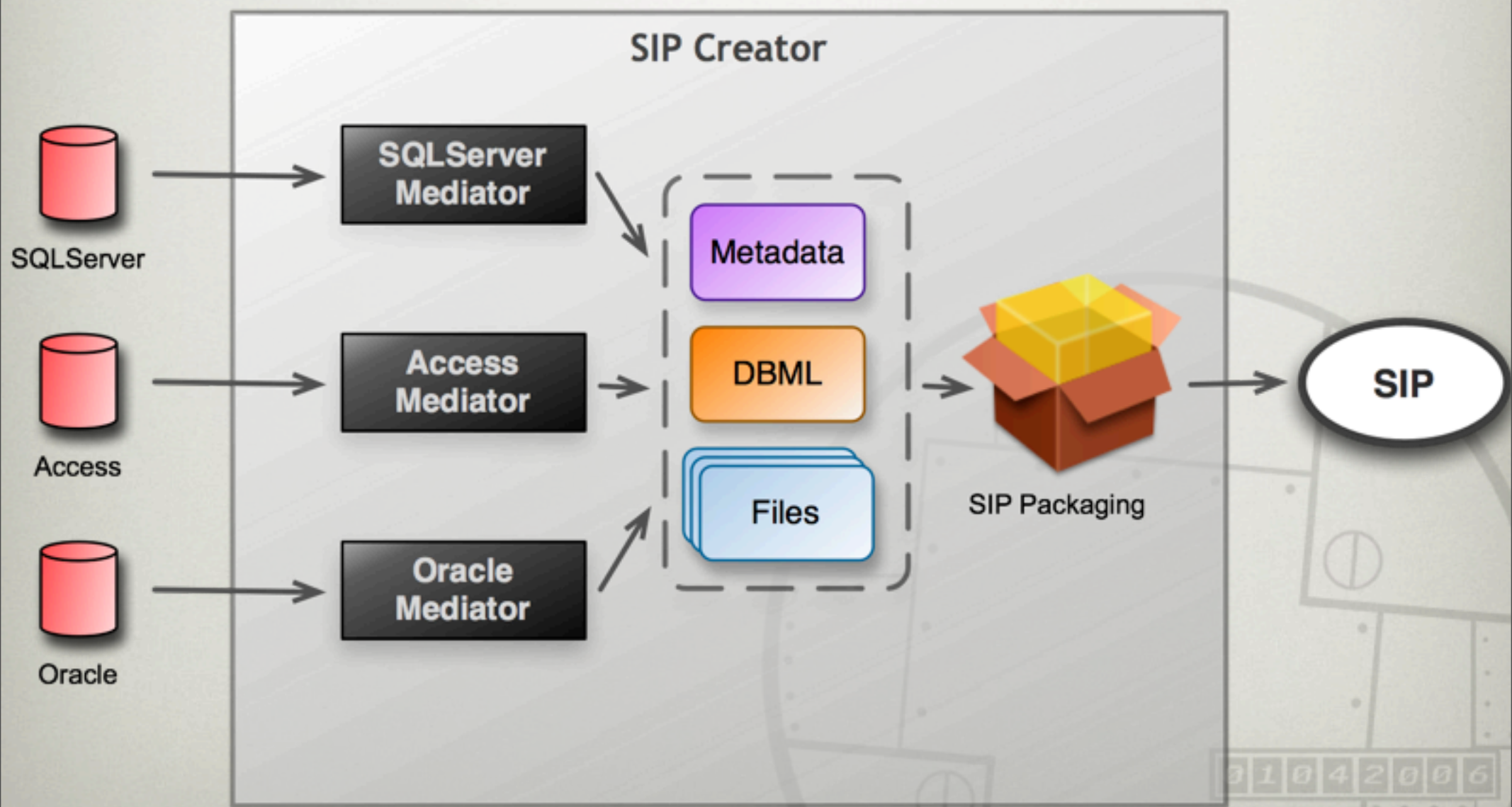
Databases

- Dados?
- Estrutura?
- Views?
- Reports?
- Stored Procedures?
- ...

First prototype:

- Data
- Structure
- Modelo Conceptual

Construtor de SIPs



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>
```

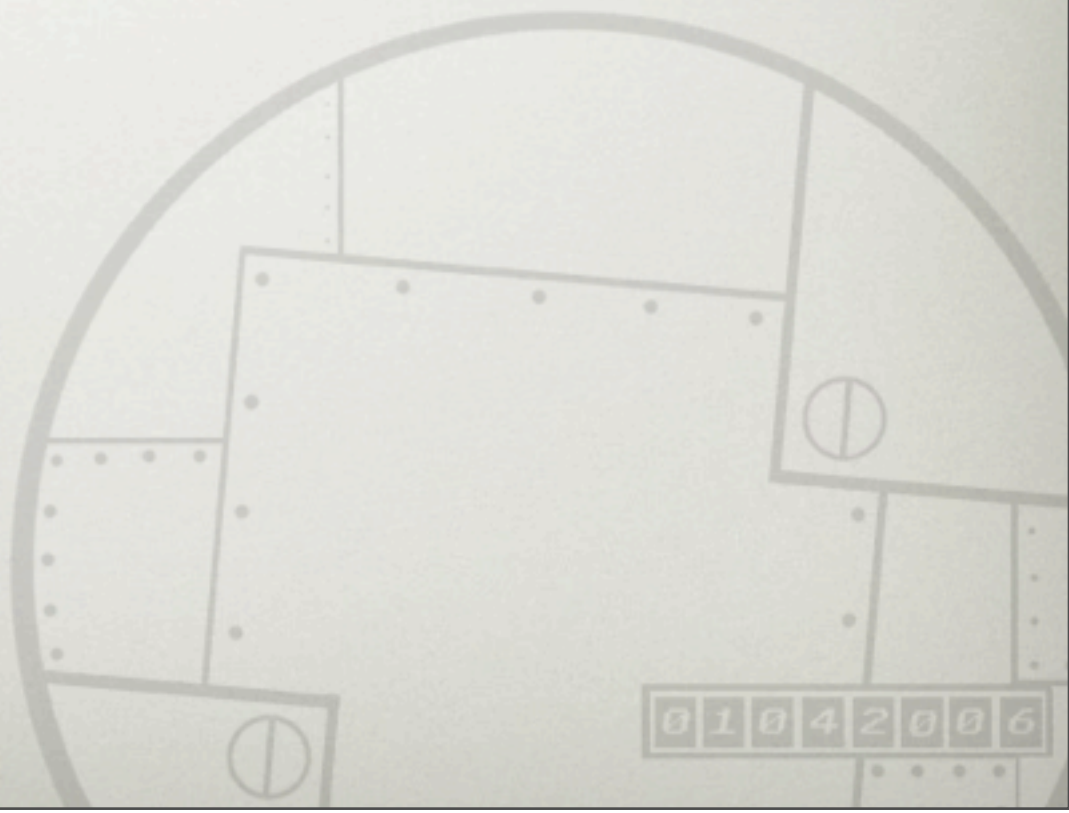
DBML

Composição do SIP:

- METS como manifesto;
- EAD na descrição das propriedades intelectuais;
- ficheiro(s) DBML
- um OD para cada BLOB
- ficheiro METS + MIX para cada OD

Final thoughts

"Data Preservation is a people problem"
Michael Lesk



Final thoughts

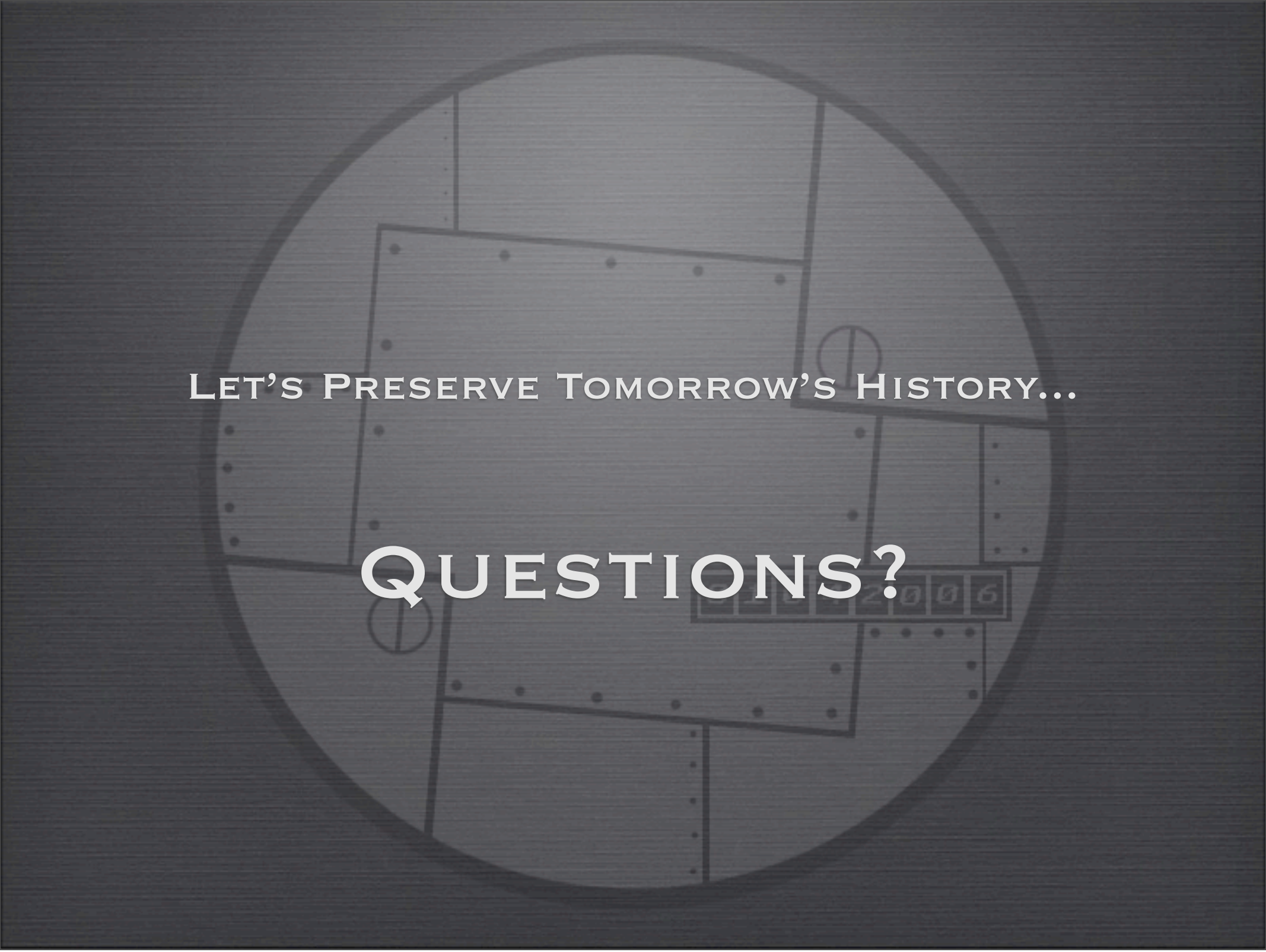
"Data Preservation is a people problem"
Michael Lesk

- As pessoas precisam de ser treinadas para guardar os dados de modo apropriado;
- Preservar o quê? Dados, Estrutura, Semântica...
- A preservação é para utilizadores futuros mas são os utilizadores de hoje que votam o orçamento;
- Temos de fazer com que as pessoas responsáveis por colectar dados tenham preocupações de preservação;
- A preservação tem falhas. Todos os sistemas são imperfeitos.

Trabalho Futuro

- 2 teses de doutoramento;
- 1 tese de mestrado;
- 1 projecto europeu:
  - ✳ www.scape-project.eu

- Criar uma comunidade;
- Promover o desenvolvimento Open Source;


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