RODA

Repositório de Objectos Digitais Autênticos

Relatório Final

Projecto 613/2006 POAP

Identificador	41012-011
Versão	Final
Autor	Luís Faria e Rui Castro
Data publicação	2007-03-18
Acesso	Público
Datas extremas	2006-04-01/2007-03-31
Início de projecto	2006-04-01
Equipa	Francisco Barbedo, José Carlos Ramalho, Luís Corujo, Luís Faria, Miguel Ferreira, Rui Castro

poap Programa Operacional da Administração Pública

UNIÃO EUROPEIA Fundo Social Europeu

Governo da República Portuguesa

Universidade do Minho

Conteúdo

1 Introdução

1.1 Contexto

> *"O sector público, na Europa, está actualmente perante uma encruzilhada, devendo fazer face a condições económicas e sociais difíceis, mudanças institucionais e impactos profundos das novas tecnologias."*
>
> *"No sector público, as administrações públicas defrontam-se com o desafio de melhorar a eficiência, a produtividade e a qualidade dos seus serviços."*
>
> *"As tecnologias da informação e das comunicações (TIC) podem ajudar as administrações públicas a fazer face aos numerosos desafios. No entanto, o centro das atenções deve ser, não as próprias TIC, mas a utilização das TIC em combinação com mudanças organizativas e novas qualificações com vista à melhoria dos serviços públicos, dos processos democráticos e das políticas públicas. É esta a vocação da administração em linha, aqui abreviadamente designada por* 'eGoverno'*."*
>
> Texto retirado de [Comissão ao Conselho et al., 2003]

A Direcção Geral de Arquivos (DGARQ, anteriormente Instituto dos Arquivos Nacionais/Torre do Tombo), tem na sua função de preservação histórica um grande desafio perante o crescimento da produção de documentos digitais pelas instituições da administração pública, devido à sua própria evolução no sentido do *eGoverno*. Não existem actualmente estruturas que suportem os processos de incorporação e gestão de informação de arquivo electrónica. É premente garantir a preservação dos documentos digitais e o seu valor evidencial, a autenticidade, para que os testemunhos das actividades das organizações públicas sejam guardados em memória social e patrimonial.

É neste contexto que se desenvolve o projecto RODA (Repositório de Objectos Digitais Autênticos), um projecto que visa desenvolver e promover uma solução tecnológica, ultimada na construção de um protótipo de repositório digital capaz de incorporar, descrever e dar acesso a todo o tipo de informação digital produzida no contexto da Administração Pública. Procura-se desta forma iniciar um processo sustentado e pró-activo que leve o IAN/TT a responder positivamente às solicitações governamentais e comunitárias no sentido do governo electrónico.

1.2 Fases

O projecto RODA está planeado em 3 macro-fases: Análise e Planeamento, Prototipagem e, finalmente, Teste e Disseminação. Cada uma destas fases é composta por várias tarefas que estão enumeradas na listagem a seguir.

As tarefas de disseminação planeadas para a 3ª fase (Teste e Disseminação) foram sendo executadas no decorrer do projecto, nomeadamente, no 4º Congresso Nacional da Administração Pública, na International Workshop for Database Preservation, no 9º Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas e nas I Jornadas de trabalho - Gestão da Informação na Era Digital.

1. Análise e Planeamento
   (a) Identificação e caracterização de requisitos funcionais para preservação digital
   (b) Análise e selecção do esquema de metainformação descritiva aplicável
   (c) Configuração do esquema de metainformação descritiva
   (d) Análise e mapeamento de funções do repositório de acordo com a norma Interpares
   (e) Construção de modelo conceptual (plano de arquitectura geral)
   (f) Desenvolvimento da arquitectura lógica (modelos: classes e sequências)
   (g) Especificação do modelo de dados
   (h) Produção de documentos de projecto
2. Prototipagem
   (a) Desenvolvimento de componentes funcionais
   (b) Desenvolvimento de interfaces gráficas
   (c) Definição de taxionomias significativas e de propriedades diplomáticas
   (d) Produção de documentos de projecto
3. Teste e Disseminação
   (a) Teste e avaliação do protótipo
   (b) Reprogramação e correcção de disfunções observadas
   (c) Divulgação interna
   (d) Divulgação externa
   (e) Produção de documentos de projecto

1.3 Objectivos e Requisitos Funcionais

Neste projecto consideram-se como objectivos primários o desenvolvimento e definição de:

- Requisitos funcionais para um arquivo digital, clientes e aplicações a integrarem;
- Modelos conceptual, lógico e de um modelo de dados para um arquivo digital;
- Identificação e selecção dos esquemas de metainformação:
  - Metainformação descritiva (e.g. Dublin core, EAD, etc.)
  - Metainformação técnica (depende da classe de objectos);
  - Metainformação estrutural (e.g. METS)
  - Metainformação de preservação (e.g. PREMIS)
- Requisitos técnicos e organizacionais;
- Protótipo arquivo digital para preservar objectos digitais susceptíveis de conservação definitiva;
- A elaboração de uma ferramenta, enquanto módulo da anterior, capaz de se "acoplar"com sistemas de gestão documental existentes na AP e assegurar funções de preservação digital numa perspectiva de gestão administrativa.

O protótipo de arquivo digital será planeado na perspectiva de obter um sistema capaz de assegurar todas as funcionalidades de um arquivo digital constantes do OAIS: integração (ingestão), armazenamento, gestão e acesso, detalhadas nos modelos Interpares. A limitação deste protótipo residirá na restrição de formatos a integrar.

Foram considerados para este projecto três classes de objectos digitais:

- Texto estruturado (e.g. documentos Word, PDF, OpenOffice, etc.)
- Imagens (jpeg, tiff, png, gif, etc.)
- Bases de dados relacionais (Access, Oracle, SQL Server, etc.)

O projecto contempla ainda alguns objectivos secundários, nomeadamente:

- A definição de uma política de arquivo para os objectos digitais produzidos pela AP (avaliação e selecção);
- Definição de uma política de preservação para o Arquivo Digital;
- Modelo(s) de financiamento que poderia(m) suportar o Arquivo Digital;
- Definição de uma taxionomia de propriedades significativas para cada uma das classes de objectos a considerar, i.e. imagens, documentos de texto e bases de dados relacionais;

1.4 Estrutura do Relatório

Este relatório é constituido pelas seguintes secções: Introdução, Taxionomias de Objectos Digitais, Normas e conceitos, Selecção da plataforma de desenvolvimento, Arquitectura interna, Descrição técnica, Avaliação e Trabalho futuro.

Em Taxionomias de Objectos Digitais apresentam-se os três tipos de objectos digitais que o repositório pretende preservar e as suas propriedades significativas.

De seguida são introduzidos conceitos e normas usados como base no RODA, nomeadamente, o modelo OAIS, os esquemas de metainformação e a autenticidade num ambiente digital.

Segue-se o procedimento da selecção da plataforma de desenvolvimento, desde a selecção de candidatos, comparação de caraterísticas dos mesmos e a conclusão sobre a escolha mais indicada.

Posteriormente apresenta-se a arquitectura interna do repositório tendo por base a plataforma escolhida. A arquitectura compreende a organização da informação e metainformação na arquitectura da mesma plataforma.

Na descrição técnica, descreve-se todo o desenvolvimento efectuado interna e externamente à plataforma de modo a implementar as funcionalidades pretendidas para o nosso repositório.

Este relatório termina com uma avaliação do trabalho elaborado ao longo do projecto e uma proposta para o trabalho futuro tendo em consideração tudo o que é necessário para o repositório entrar em exploração.

2 Taxionomias de Objectos Digitais

Os tipos de documentos contemplados neste protótipo são o texto estruturado, que poderá conter tabelas e imagens, as imagens e as bases de dados relacionais.

As conclusões da equipa de desenvolvimento relativamente aos formatos de preservação mais adequados para os objectos que se pretendem preservar são apresentadas nas secções 2.1, 2.2 e 2.3.

Para mais detalhes sobre as resoluções relativas às taxionomias de objectos pode ser consultado [Barbedo, 2006b].

2.1 Texto estruturado

O formato de preservação do texto estruturado é o PDF/A [pdf, 2007]. Actualmente existem poucas ferramentas desenvolvidas para serem usadas na manipulação deste formato, nenhuma delas de acesso livre, portanto será usado o formato PDF 1.4 (do qual o PDF/A é um subtipo) como formato de preservação até que as ferramentas adequadas para lidar com ficheiros em formato PDF/A estejam disponíveis para uso generalizado.

A escolha do PDF/A como formato de preservação para texto estruturado é sustentado pelo facto de permitir a persistência da aparência do layout original do objecto, factor relevante para a intelegibilidade do mesmo e por o formato PDF, do qual PDF/A é um subtipo, ser bastante disseminado por toda a comunidade cibernética garantir a persistencia do mesmo. Além disto, o subtipo PDF/A foi especialmente criado para motivos de preservação e arquivo.

Este formato é capaz de conter documentos de texto estruturado (com tabelas e imagens) mantendo o aspecto e paginação do documento original.

Segundo o relatório [Barbedo, 2006b] para um documento manter a sua autenticidade os seguintes elementos diplomáticos deverão estar explicitamente presentes:

- Autor
- Destinatário
- Originador
- Produtor
- Data de criação
- Data de recepção

- Descrição da acção
- Relações
- anexos (que inclui elementos de validação como assinatura digital)

2.2 Imagens fixas bidimensionais

O formato de preservação para imagens fixas é o TIFF [Adobe, 2002] sem qualquer tipo de compressão. Este formato está especificado de uma forma aberta e é muito bem suportado por inúmeras ferramentas de código-aberto e aceite pela comunidade como um bom formato de preservação para imagens fixas bidimensionais[1].

Segundo o relatório [Barbedo, 2006b] os elementos diplomáticos que devem estar presentes na imagem para assegurar a sua autenticidade são idênticos aos dos documentos textuais. A maior parte destes, como a data de produção, autor, originador, produtor podem ser encontradas no cabeçalho (particularmente nos headers dos ficheiros TIFF). Esta informação está presente no cabeçalho do documento-imagem. O assunto corresponde à representação iconográfica suportada pela imagem.

2.3 Bases de dados relacionais

O formato de preservação de bases de dados relacionais é o DBML [Henriques et al., 2002] (um formato XML). Este formato é uma proposta deste projecto na tentativa de solucionar a falta de opções adequadas para um formato de preservação para bases de dados relacionais. Os alvos de preservação serão a estrutura da base de dados (tabelas e relações entre tabelas) e os dados.

Segundo o relatório [Barbedo, 2006b] os atributos diplomáticos propostos pelo Projecto UBC tornam-se dificeis de identificar directamente numa BD:

- Autor
- Destinatário
- Originador
- Produtor
- Data de criação

[1] Veja http://www.digitalpreservation.gov/formats/fdd/fdd000022.shtml

- Data de recepção
- Descrição da acção
- Relações
- anexos (que inclui elementos de validação como assinatura digital)

A preservação de bases de dados implica a separação lógica e física dos dados relativamente ao sistema que os gere - SGBD. Para assegurar a propriedade de totalidade é essencial preservar todos os dados assim como informação sobre eles (tipos de dados) e a estrutura em que são conservados.

Alguns elementos diplomáticos requeridos para assegurar a autenticidade e confiabilidade documentais existem na BD e no SGBD.

Neste último caso destacam-se as queries que descrevem informação e as rotinas de auditoria (*audit trails*) que permitem demonstrar a validação dos dados.

O autor e a data são elementos diplomáticos que podem existir na BD.

Por estes motivos é importante sob o ponto de vista arquivístico preservar a BD e algumas funcionalidades do SGBD.

Relativamente a BD o problema de preservação é complexo.

Decrementar pata texto simples ou RTF é inviável. A perda de estrutura é total num caso e no outro é claramente comprometida.

Decrementar para csv (uma variante de texto) não permite guardar a estrutura.

A utilização de xml permite guardar o conteúdo e a estrutura, havendo duas possibilidades: guardar o conteúdo juntamente com a estrutura ou guardar estas componentes separadamente.

Para além da questão do que guardar e como o fazer (ou seja, a produção do AIP) há ainda o problema da constituição do DIP, ou seja a apresentação ao utilizador final. Neste caso colocam-se os problemas de fornecer não apenas a informação mas as funcionalidades inerentes ao SGBD que tenham sido objecto de preservação.

3 Normas e conceitos

3.1 O modelo de referência OAIS

Em 1990, o Consultative Comitee for Space Data Systems (CCSDS) iniciou um esforço conjunto com a International Organization for Standardization (ISO) a fim de desenvolver um conjunto de normas capazes de regular o armazenamento a longo prazo de informação digital produzida no âmbito de missões espaciais.

Deste esforço nasceu o modelo de referência OAIS (Open Archival Information System), um modelo conceptual que visa identificar os componentes funcionais que deverão fazer parte de um sistema de informação dedicado à preservação digital [Lavoie, 2004, CCSDS, 2002, Ferreira, 2006]. O modelo descreve ainda as interfaces internas e externas do sistema e os objectos de informação que serão manipulados no seu interior [Lavoie, 2004, Ferreira, 2006].

O modelo de referência OAIS foi aprovado como uma norma internacional em 2003 - ISO Standard 14721:2003 [CCSDS, 2002, Ferreira, 2006].

Um dos contributos mais notáveis desta iniciativa tem que ver com a definição de uma terminologia própria que viria a facilitar a comunicação entre os diversos intervenientes envolvidos na preservação de objectos digitais [Saramago, 2004, Ferreira, 2006].

A figura 1 ilustra os diferentes componentes funcionais, assim como os pacotes de informação trocados no interior de um repositório digital compatível com o modelo de referência OAIS.

O Produtor deverá ser entendido como a entidade externa ao repositório que se responsabiliza pela submissão de material. O material submetido a arquivo está aqui representado pelo SIP (*Submission Information Package*) [2].

Durante o processo de submissão ou incorporação, designado neste contexto por Ingestão, o repositório é responsável por garantir a integridade da informação recebida. Ainda nesta fase, é produzida toda a Informação Descritiva que irá suportar a descoberta e localização do material depositado. Essa informação descritiva (ou metainformação) é armazenada e gerida pelo componente Gestão de Dados[3]. O material a preservar (i.e. AIP (*Archival Information Package*) [4] será conservado no Repositório de Dados[5]. O componente de ingestão constitui, assim, a interface entre o arquivo OAIS e os

[2]Em português poderia chamar-se Pacote de Informação de Submissão.
[3]Do inglês *Data Management*.
[4]Em português poderia chamar-se Pacote de Informação de Arquivo.
[5]Do inglês *Archival Storage*.

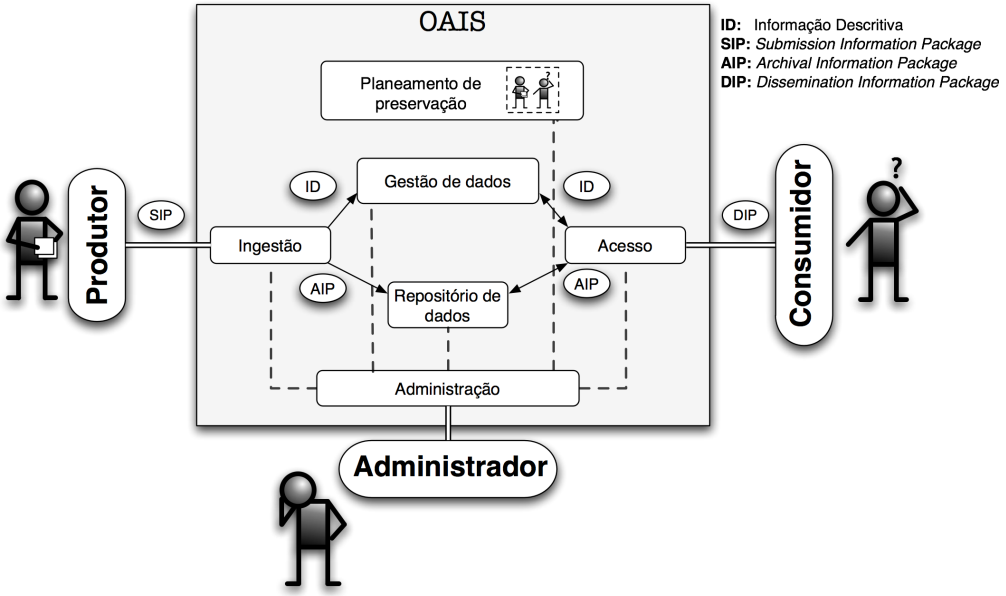


Figura 1: Modelo de referência Open Archival Information System (OAIS).

respectivos produtores de informação [Lavoie, 2004, Ferreira, 2006].

O componente Planeamento de Preservação encarrega-se da definição de políticas de preservação. Este serviço é responsável pela monitorização do ambiente externo ao repositório e definir ou actualizar os termos em que a Administração actua na forma de políticas e procedimentos. É, por exemplo, da responsabilidade deste componente definir as estratégias de preservação a utilizar no interior do repositório, monitorizar as tendências comportamentais da sua comunidade de interesse ou identificar formatos na iminência de se tornarem obsoletos [Lavoie, 2004, Ferreira, 2006].

O componente Acesso estabelece a ponte entre o repositório e a sua comunidade de interesse, i.e. o conjunto de Consumidores de material custodiado. Este componente é responsável facilitar a descoberta e localização dos objectos digitais, bem como preparar os mesmos para entrega ao consumidor.

Os pacotes que são entregues ao consumidor assumem a forma de DIPs *Dissemination Information Package*[6] [Lavoie, 2004]. É de realçar o facto de os DIPs poderem ser diferentes dos AIPs. A informação que é entregue ao consumidor poderá ser apenas um subconjunto da informação arquivada, ou até, uma versão transformada da mesma [Ferreira, 2006].

Por último, o componente Administração é responsável pelas operações

[6]Em português poderia chamar-se Pacote de Informação de Disseminação.

diárias de manutenção e, sobretudo, pela parametrização e monitorização dos processos que se desenrolam no interior do repositório. A sua acção é controlada através das políticas e procedimentos definidos pelo Planeamento de Preservação. A Administração interage com todos os restantes componentes de forma a assegurar o correcto funcionamento do repositório em geral [Lavoie, 2004, Ferreira, 2006].

3.2 Esquemas de metainformação

Esta secção apresenta uma breve descrição dos vários esquemas de metainformação mencionados ao longo deste relatório (figura 2).

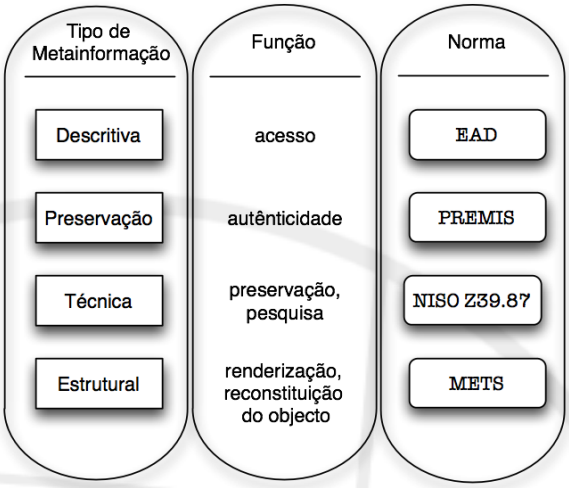


Figura 2: Esquemas utilizados no esboço

3.2.1 Esquema EAD

O EAD (*Encoded Archival Description*) define metainformação descritiva. A última versão deste esquema é a de 2002. Este esquema descreve a informação de forma contextual, ajudando a categorizar e localizar a mesma (i.e. a metainformação descritiva é utilizada por motores de busca para localizar informação).

Uma instância EAD contém três partes:

<eadheader> - contém informação sobre a metainformação em si.

<frontmatter> - contém informação conveniente para a apresentação ou publicação da metainformação.

<archdesc> - contém informação descritiva sobre um fundo documental e seus constituintes, informação contextual e administrativa associada.

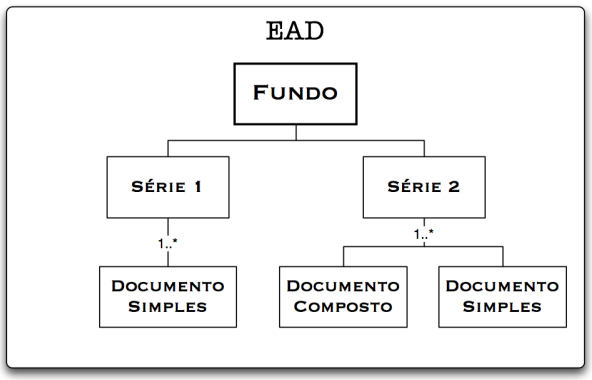


Figura 3: Esquema de um EAD exemplo

Cada instância contêm um ou mais elementos <c>. Estes elementos podem ser múltiplos e estar aninhados, criando uma estrutura hierárquica. Cada elemento tem um identificador único e um nível (figura 3) que pode assumir os seguintes valores[7]:

fundo e subfundo - O fundo é a mais ampla unidade de descrição arquivística. É constituído pelo conjunto de todos os documentos, independentemente da sua forma ou formato, produzidos/acumulados por uma entidade singular ou colectiva no exercício das suas funções e actividades. Em casos de entidades produtoras especialmente complexas, o fundo pode desdobrar-se num ou mais subfundos, normalmente expressão de macro-entidades orgânicas ou funcionais.

classe e subclasse - A classe é uma unidade de descrição arquivística de nível intermédio, que corresponde a uma especificação funcional/orgânica da entidade produtora (quando subordinada ao fundo) ou de uma das suas macro-entidades (quando subordinada a um subfundo). Um fundo e/ou um subfundo pode comportar uma ou várias classes, e cada uma delas pode desenvolver-se em subclasses, nos termos do plano de classificação utilizado pela entidade produtora.

série e subsérie - A série é uma unidade de descrição arquivística constituída por um conjunto de documentos simples ou compostos que par-

[7] estes valores são os considerados pela nossa implementação, usando o atributo *other-level* do EAD, na definição oficial do EAD estes valores diferem um pouco

tilham alguma propriedade particular e cuja identidade decorre da estrutura do plano de classificação utilizado pela entidade produtora - a situação mais usual é reportarem-se ao exercício de uma mesma actividade específica. A série pode estar directamente subordinada a qualquer entidade superior (fundo, subfundo, classe ou subclasse) e pode subdividir-se ou não em subséries, de acordo com as regras definidas pela entidade produtora.

documento composto - unidade de descrição arquivística constituída por um conjunto dos documentos agregados pela entidade produtora, cuja identidade normalmente decorre do facto de se reportarem todos a um mesmo caso, procedimento ou assunto (por exemplo, um processo administrativo relativo a um concurso determinado, um processo judicial, um processo clínico, uma base de dados com o recenseamento da população numa data determinada). Na situação mais frequente, mas não imperativa, o documento composto está directamente subordinado a uma série ou subsérie.

documento simples - O documento simples é a mais pequena e intelectualmente indivisível unidade de descrição arquivística. É a expressão documental de um acto ou de uma ocorrência (um parecer, uma acta, um relatório, uma fotografia, um registo numa base de dados).

Cada nível de descrição contém informação descritiva, seguindo o modelo da *General International Standard of Archival* (ISAD(G), [International Council on Archives, 1999]). Como exemplos de campos existem: título, datas extremas, história biográfica, história arquivística, âmbito e conteúdo, existência e localização dos originais e cópias, etc.

Para mais informação sobre este esquema de metainformação consulte:

- Official EAD Version 2002 Web Site [The Library of Congress, 2002a]
- Society of American Archivists [Society of American Archivists, 2003]
- RLG Best Practices Guidelines for Encoded Archival Description [RLG EAD Advisory Grou
- EAD Tools Survey [Society of American Archivists, 2006]
- RLG EAD Report Card [RLG, 2002]

Um exemplo de um ficheiro EAD pode ser consultado no anexo F.

3.2.2 Esquema PREMIS

Em 2003 a OCLC (*Online Computer Library Center*) e a RLG (*Research Libraries Group*) estabeleceram o grupo de investigação *PREservation Metadata: Implementation Strategies* (PREMIS). Em Maio de 2005 este grupo apresentou o seu relatório final, o *Data Dictionary for Preservation Metadata* - [OCLC and RLG, 2005], que define o esquema que é apresentado.

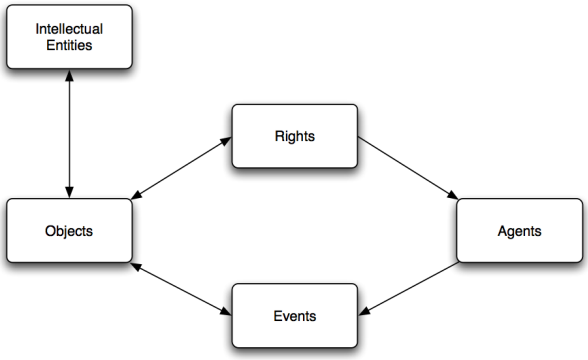


Figura 4: PREMIS Data Model

O esquema está organizado segundo um modelo simples (figura 4) com cinco tipos de entidades envolvidas nas actividades de preservação digital:

Object - ou Objecto Digital, é a unidade discreta de informação no formato digital. Um *Object* pode ser uma *Representation*, *File*, *Bitstream* ou *Filestream*.

Intellectual Entity - é um conjunto coerente de conteúdos, que pode ser razoavelmente descrito como uma unidade (ex. um livro, uma imagem, uma base de dados). Uma *Intellectual Entity* pode conter outras *Intellectual Entities*, por exemplo um livro pode conter uma imagem.

Event - é uma acção que envolve pelo menos um *Object* ou *Agent* conhecidos pelo repositório de preservação.

Agent - é uma pessoa, organização ou programa que realiza eventos de preservação (*Events*) no tempo de vida de um *Object*.

Rights - é um conjunto de um ou mais direitos ou permissões relativos a um *Object* e/ou *Agent*

O *PREMIS Data Dictionary* inclui unidades semânticas para *Objects*, *Events*, *Agents* e *Rights*. O quinto elemento no modelo, *Intellectual Entity*,

foi considerado fora do contexto deste *Data Dictionary* pois é bem servida pelos esquemas de metainformação descritiva existentes (e.g. EAD, MARC [The Library of Congress, 2005], MODS [The Library of Congress, 2006a], Dublin Core [OCLC, 1995], etc.) e porque é demasiado específica do domínio em consideração.

No *PREMIS Data Dictionary* a entidade *Object* tem três subtipos: *file*, *bitstream* e *representation*.

Um *file* é uma sequência de *bytes* com ordem e nome, reconhecida por um sistema operativo. Um ficheiro tem propriedades como permissões, tamanho e data da última modificação.

Um *bitstream* é um conjunto de dados dentro de um ficheiro (*file*) que tem algumas propriedades comuns significativas para efeitos da preservação digital.

Uma *representation* é um conjunto de ficheiros e metainformação estrutural, necessários para interpretação completa e razoável de uma Entidade Intelectual (*Intellectual Entity*).

A entidade *Event* agrega metainformação sobre acções. Um repositório de preservação irá criar *Events* por variadas razões. Documentação sobre acções que modificam (e.g. criam uma nova versão) de um objecto digital são fundamentais para manter um registo das intervenções realizadas sobre o objecto, elemento chave para a autenticidade. Acções que criam *Objects* ou que modificam *Objects* existentes são importantes para explicar os mesmos. Até acções que não alteram nada, como validações e análises à integridade nos objectos, podem ser importantes registar para efeitos de gestão.

Para mais informação sobre o PREMIS veja [The Library of Congress, 2006d].

Um exemplo de um ficheiro PREMIS contendo metainformação técnica NISO Z39.87 pode ser consultado no anexo G.

3.2.3 Esquema NISO Z39.87

Este esquema define um conjunto normalizado de elementos de metainformação para imagens digitais. O esquema utilizado data de 2002, no entanto está neste momento em período de revisão a versão de 2005 que vem trazer uma nova organização mais compatível com o PREMIS.

A versão de 2002 divide a metainformação técnica em quatro secções:

1. **Basic Image Parameters** - que agrupa elementos fundamentais para a reconstrução do ficheiro digital como uma imagem renderizável em interfaces electrónicas. Exemplos de elementos:

- **MIMEType** - o formato da imagem;
- **ByteOrder** - a ordem dos bits em que os números estão representados;
- **Compression** - o esquema de compressão e o nível de compressão utilizado;
- **ColorSpace** - o espaço de cores utilizada;
- **DisplayOrientation** - a orientação em que a imagem deve ser apresentada num monitor convencional;

2. **Image Creation** - algo como metainformação técnica descritiva, dá informação sobre aspectos logísticos e condições administrativas relativas à captura da imagem digital. Exemplos de elementos :

   - **SourceType** - o tipo de material analógico de foi digitalizado (e.g. microfilme);
   - **ImageProducer** - o produtor a nível organizacional da imagem;
   - **HostComputer** - o computador e/ou sistema operativo usado na criação da imagem;
   - **ScanningSystemCapture** - todas as propriedades relevantes do scanner usado na captura, caso este seja o caso;
   - **DigitalCameraCapture** - todas as propriedades relevantes da câmera digital usado na captura, caso este seja o caso;

3. **Imaging performance assessment** - o princípio operativo desta secção é manter os atributos da imagem inerentes à sua qualidade. Estes elementos servem como métricas para medir a fidelidade da imagem corrente e dos resultados de técnicas de preservação, especialmente a migração. Exemplos de elementos:

   - **XSamplingFrequency e YSamplingFrequency** - A resolução da imagem nos dois eixos;
   - **ImageWidth e ImageLength** - O tamanho da imagem nos dois eixos;

4. **Change history** - esta secção tem a função de documentar os processos aplicados aos dados da imagem no ciclo de vida desta. Elementos:

   - **Image Processing** - um sumário dos processos efectuados na imagem;

- **Previous Image Metadata** - metainformação técnica de versões anteriores da imagem, se dos processos efectuados na imagem resulta uma nova versão;

Para mais informação sobre este esquema de metainformação consulte:

- NISO Metadata for Images in XML Schema Official Web Site [The Library of Congress, 2004
- NISO Z39.87 -200x Development page [NISO, 2006]

3.2.4 Esquema METS

Metadata Encoding & Transmition Standard (METS) é uma norma que permite agrupar metainformação descritiva, administrativa e estrutural sobre objectos guardados num repositório digital[8].

Um documento METS consiste em sete secções principais:

1. **Cabeçalho METS** - O cabeçalho METS contém metadados descrevendo o documento METS em si, incluindo informação como o criador, editor, etc.

2. **Metadados Descritivos** - A secção de metadados descritivos pode apontar para metadados descritivos externos ao documento METS (por exemplo, um registo MARC num OPAC ou um registo EAD mantido num servidor Web), ou conter metadados descritivos embebidos, ou ambos. Múltiplas instancias de metadados descritivos, tanto internas como externas, podem ser incluídos na secção de metadados descritivos.

3. **Metadados Administrativos** - A secção de metadados administrativos oferece informação sobre como os ficheiros foram criados e armazenados, direitos de propriedade intelectual, metadados sobre o objecto original a partir do qual o objecto digital foi derivado, e informação sobre a proveniência dos ficheiros que compõem o objecto digital (isto é, relações de ficheiros originais/derivados, e informação de migração/transformação). Tal como os metadados descritivos, os metadados administrativos podem ser tanto externos ao documento METS, como codificados internamente.

4. **Secção de Ficheiros** - A secção de ficheiros lista todos os ficheiros que contêm as versões electrónicas do objecto digital. Elementos <file>podem ser agrupados em elementos <fileGrp>, para permitir a subdivisão de ficheiros por versão do objecto.

[8]No RODA apenas a função estrutural é utilizada

5. **Mapa Estrutural** - O Mapa Estrutural é o coração do documento METS. Este esboça uma estrutura hierárquica para o objecto digital e liga os elementos dessa estrutura a ficheiros com conteúdos e metadados referentes a cada elemento.

6. **Ligações Estruturais** - A secção de Ligações Estruturais do METS permite aos criadores METS registar a existência de hiperligações entre nós, na hierarquia esboçada no Mapa Estrutural. Esta secção tem um valor particular na utilização do METS para arquivar sítios Web.

7. **Comportamento** - Uma secção de comportamento pode ser usada para associar comportamentos exequíveis com o conteúdo do objecto METS. Cada comportamento do conjunto tem um interface e um mecanismo. A interface representa uma definição abstracta do conjunto de comportamentos. O mecanismo identifica um módulo de código executável, o qual implementa e executa os comportamentos definidos de forma abstracta na interface.

Para mais informação visite o METS Official Web Site [The Library of Congress, 2006b].

Vários exemplos de ficheiros METS podem ser consultados nos anexos C, D e E.

3.3 Autenticidade num ambiente digital

O termo "Autenticidade"em informação documental conota um significado preciso, e no entanto dispare, em diferentes contextos e comunidades. Esta, pode significar ser original, mas também ser fiel a um original; pode significar a não corrupção, mas também ser de clara e conhecida proveniência, existindo, ou não, "corrupção".

No entanto, em qualquer dos contextos, o conceito de autenticidade tem profundas implicações na tarefa de catalogar e descrever um item de informação. Este conceito tem igualmente profundas ramificações na preservação, definindo os parâmetros do que é preservado e, consequentemente, por que técnica ou conjunto de técnicas.

No projecto Interpares[The InterPARES Project, 2007] são definidos guias para a preservação de documentos digitais autênticos. No RODA usar-se-á a migração como estratégia fundamental de preservação. Esta, utiliza conversores para transformar o formato dos objectos digitais que estejam em perigo de obsolescência. Deste modo, a autenticidade de um documento digital que será migrado ao longo dos anos não se pode basear na sua originalidade, mas sim na sua fidelidade em relação ao original. Contudo, esta fidelidade tem de

ser medida em cada migração, pois as conversões normalmente incluem perdas de informação. Logo, todas as migrações têm de ser bem documentadas de forma a manter a autenticidade dos objectos digitais.

O registo de todas as transformações e validações dos objectos digitais preenchem a metainformação de preservação que será guardada no RODA segundo o esquema PREMIS.

4 Selecção da plataforma de desenvolvimento

Implementar um repositório de raiz é um trabalho bastante extenso e fora dos objectivos deste projecto. Existem várias iniciativas *open-source* nos quais um repositório deste tipo se poderia basear, mas há dois candidatos que se destacaram: DSpace e Fedora.

4.1 DSpace

Figura 5: Logótipo do DSpace

O DSpace[9] (logótipo na figura 5) é um repositório digital *open-source* para instituições de investigação. Desenvolvido numa cooperação entre a biblioteca do MIT (*Massachusetts Institute of Technology*) e os Laboratórios da Hewlett-Packard, o DSpace está disponível sob uma licença *open-source* BSD[10] para instituições de investigação utilizarem na sua forma original ou modificarem e estenderem conforme as suas próprias necessidades. Muitas instituições de investigação por todo mundo utilizam o DSpace como solução para os mais variados tipos de arquivos digitais, entre eles:

- Repositórios Digitais,
- Repositórios de material pedagógico (*Learning Object Repositories*),
- Teses electrónicas (*eTheses*),
- Gestão de Arquivos Electrónicos (*Electronic Records Management*),
- Preservação Digital,
- Publicação Electrónica

4.2 Fedora

Fedora[11] (logótipo na figura 6) é um software *open-source* que oferece uma arquitectura flexível de serviços para gestão e disseminação de conteúdos. Tem no seu núcleo um modelo de dados totalmente flexível que suporta múltiplas vistas/disseminações de cada representação digital e das relações entre elas. Estas representações podem encapsular conteúdos geridos localmente ou fazer referência a conteúdos remotos. Vistas/disseminações dinâmicas são possíveis associando *web services* às representações. As representações existem dentro de uma arquitectura de repositório que suporta uma variedade de funções de gestão. Todas as funções do Fedora, tanto ao nível da representação como a nível do repositório, são expostas como *web services*. Estas funções podem ser protegidas com políticas de controlo de acessos de granularidade fina.

Esta combinação de características faz do Fedora uma solução atractiva em vários domínios. Alguns exemplos de aplicações que foram construídas sobre o Fedora incluem: gestão de bibliotecas, sistemas de produção de multimédia, repositórios de arquivo, repositórios institucionais, bibliotecas digitais para educação.

[9] http://dspace.org

[10] Licença BSD é uma licença open-source em que a redistribuição do código terá de ser sob o mesmo tipo de licença. Só os detentores dos direitos de todo o código, podem mudar a licença ou transferir o copyright.

[11] http://fedora.info

Figura 6: Logótipo do Fedora

4.3 DSpace vs. Fedora

De seguida serão analisados os requisitos funcionais do projecto RODA descritos em [Barbedo, 2006a] e é feita uma verificação das funcionalidades das bases de desenvolvimento comparativamente aos requisitos. Os requisitos estão divididos em três processos: Ingestão, Gestão e Disseminação, listados no Anexo B. O resultado desta análise está ilustrado no gráfico da figura 7.

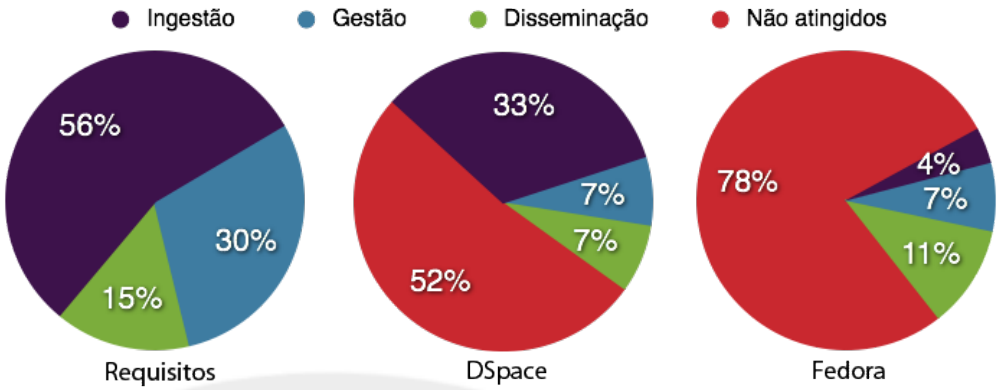


Figura 7: Percentagem de requisitos satisfeitos

Como se pode observar no gráfico, o DSpace cumpre, em média, mais requisitos do que o Fedora. Este facto acentua-se no processo de ingestão, que é o mais complexo, e no qual o Fedora apenas cumpre 6% dos requisitos. Passamos seguidamente à tentativa de prototipar uma solução em ambas as plataformas, de modo a verificar até que ponto estas cumprem os requisitos.

4.4 Prototipagem da Solução

O RODA deve assegurar o suporte, no seu modelo de dados, dos requisitos de metainformação definidos nos objectivos do projectos. Na secção 2 são introduzidos os vários esquemas de metainformação, dos quais os mais predominantes são o PREMIS e o EAD. O suporte para estes esquemas traduz-se, de uma maneira simplificada, na capacidade do repositório descrever entidades intelectuais de uma forma hierárquica, como definido pelo EAD, e guardar metainformação de preservação relativa a cada representação destas entidades intelectuais, segundo o esquema do PREMIS.

4.4.1 Solução no DSpace

Existe um protótipo de um sistema de eventos para o DSpace em que cada entidade com relevância arquivística (dentro da estrutura do DSpace) é um

Criador de eventos. Estes eventos são passados a Processadores que filtram os tipos de eventos e as entidades que os originaram. Os Processadores passam os eventos a Consumidores segundo o filtro configurado para este. É possível configurar um Processador JMS (Java Messaging Service) possibilitando receber eventos de forma assíncrona e remota. Utilizando este protótipo é possível criar uma classe local, ou um serviço remoto que utilize o JMS, para registar informação relevante dos eventos criados no esquema PREMIS.

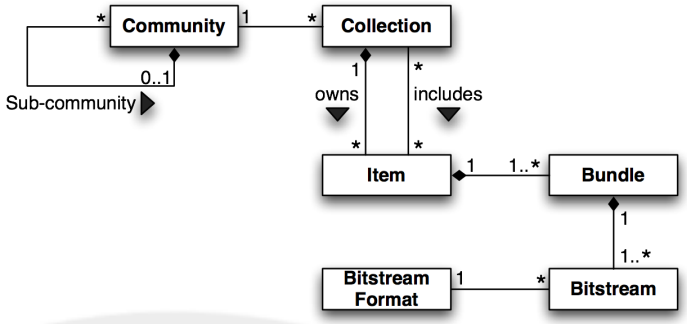


Figura 8: Modelo de dados do DSpace

Embora a implementação da funcionalidade de suporte ao registo de eventos PREMIS seja clara e simples o mesmo não acontece com o suporte a esquemas de metainformação descritiva hierárquica. O DSpace baseia-se num modelo de dados próprio (figura 8) que não reflecte nenhum esquema internacional de metainformação descritiva hierárquica mas antes um suporte interno a uma hierarquia oca (basicamente sem informação associada) que serve apenas para organizar conjuntos que são descritos com esquemas de metainformação planos (i.e. não hierárquicos), como é o caso particular do Dublin Core [OCLC, 1995].

Como podemos observar na figura 8, as comunidades podem aninhar-se infinitamente, formando uma hierarquia. Estas podem ter várias colecções e só as colecções podem conter itens. Esta estrutura traz várias limitações como, por exemplo, um item não pode existir em qualquer nodo da hierarquia (e.g. directamente por baixo da raiz) porque este tem que estar contido numa colecção e a colecção tem de estar contida numa comunidade. Outro ponto que vai completamente contra as noções arquivísticas é o de uma colecção poder pertencer simultaneamente a várias comunidades.

Existem, no entanto, abstracções que podem ser assumidas de modo a tornar o modelo de dados do DSpace mais permissivo. Podemos mapear uma colecção debaixo de uma comunidade como um nodo da árvore da hi-

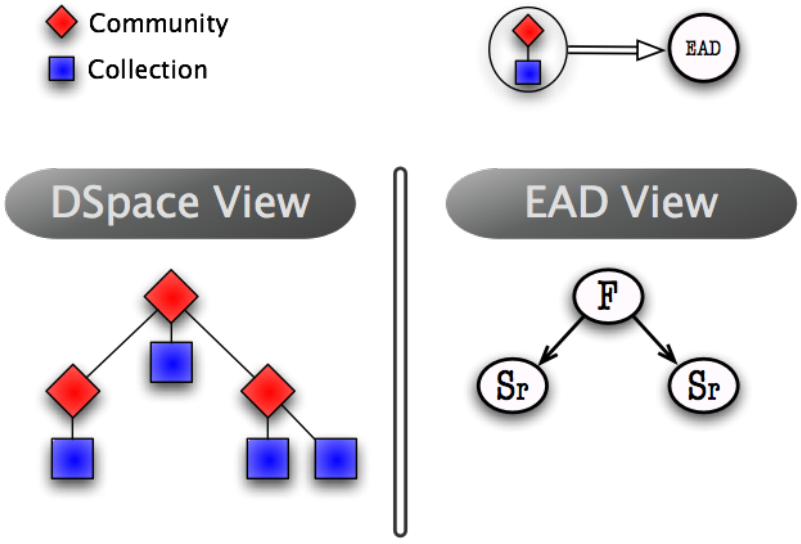


Figura 9: Abstracção do modelo de dados do DSpace para EAD

erarquia (figura 9). A comunidade dá-nos a possibilidade de criar outros níveis hierárquicos debaixo deste nodo e a colecção deixa-nos inserir itens directamente neste nodo. Contudo este mapeamento pode trazer problemas de incoerência se não forem antecipados todos os casos possíveis do modelo de dados do DSpace e decisões estranhas podem ter de ser tomadas, como decidir se mais que uma colecção deve ser mapeada dentro de uma comunidade (caso que se pode observar na figura 9).

4.4.2 Solução no Fedora

O Fedora é uma plataforma muito flexível a nível da informação e/ou metainformação que é possível armazenar e a nível da organização dessa informação no repositório. Do ponto de vista do repositório existem objectos fedora (OF) e relações (RF) entre esses objectos. O conteúdo e as relações de um objecto são da inteira responsabilidade dos criadores dos mesmos.

Os nodos de descrição (<c>) podem ser guardados em objectos separados (um objecto por unidade de descrição) e podem ser usadas relações entre estes para estabelecer a hierarquia desejada.

A metainformação de preservação e técnica relativa a uma entidade intelectual pode ser guardada num objecto fedora e ligada a uma entidade intelectual através de uma relação.

Por sua vez as representações e a metainformação estrutural podem ser

também guardadas num objecto (uma representação por objecto) e relacionadas com a metainformação de preservação através de relações.

Esta organização é apenas uma das diversas possibilidades para mapear os esquemas de metainformação dentro do Fedora. A flexibilidade em termos de estrutura e organização de objectos é praticamente ilimitada. A evolução de uma estrutura para outra é também relativamente simples do ponto de vista do repositório.

O Fedora é também bastante flexível relativamente às funcionalidades que é possível implementar sobre a informação e/ou metainformação. Cada objecto pode ter serviços associados, tantos quantos os necessários para implementar um determinado comportamento. Por exemplo, um disseminador que exporta uma base de dados para o formato *Access* pode ser associado a todas as representações de bases de dados presentes no repositório.

Outros serviços podem ser criados sem estar associados a qualquer objecto em particular. Estes serviços podem, por exemplo, disponibilizar funcionalidades de ingestão de SIPs no RODA, funcionalidades de gestão, etc.

As funcionalidades de preservação, requisito essencial do RODA, não existem no Fedora. No entanto, a possibilidade de guardar metainformação de preservação (PREMIS) existe intrinsecamente no Fedora. Não obstante, os mecanismos para a actualização automática dessa metainformação ainda não existem. Há, contudo, um grupo de trabalho associado à equipa de desenvolvimento do Fedora que tem como objectivo desenvolver as funcionalidades de preservação no Fedora. A proposta deste grupo é criar um sistema de eventos que permita a notificação de qualquer tipo de eventos (ingestão, validação, disseminação, etc) a todos os componentes "interessados". Isto possibilitará a criação de serviços que registam todos os eventos de preservação (no caso do RODA, em formato PREMIS).

O Fedora é uma plataforma bastante genérica e flexível que pode ser usada para implementar basicamente qualquer solução no âmbito dos repositórios digitais. Esta flexibilidade vem com o custo de distanciar o Fedora de uma solução final, em termos de trabalho de implementação, relativamente a outras soluções menos genéricas.

4.5 Discussão

A plataforma escolhida para o desenvolvimento do RODA foi o Fedora. As razões para esta escolha são basicamente as expostas em 4.4.1 e 4.4.2 e assentam essencialmente na flexibilidade e potencial desta plataforma em relação aos outros repositórios analisados.

O DSpace em termos de funcionalidades para o utilizador está mais completo, mas impõe uma estrutura de dados interna que é desadequada aos objectivos do RODA o que obrigaria o uso de "remendos"de modo a ser possível utilizar um esquema de metainformação descritiva hierárquico (EAD). Além disso, a extensibilidade deste é limitada porque não implementa um sistema de extensões (como os serviços do Fedora).

O Fedora é a solução mais adequada para o RODA porque não traz qualquer tipo de restrições em termos de esquemas de metainformação que se queiram usar e possui uma arquitectura de serviços que possibilita que funcionalidades sejam adicionadas ao repositório de forma elegante e independente da implementação do próprio repositório.

5 Arquitectura Interna

A arquitectura do RODA é baseada na arquitectura do Fedora visto que este é a peça central do repositório. O Fedora gere a informação que é inserida, permite que serviços sejam associados a essa informação e disponibiliza uma interface sobre a qual outros serviços podem ser desenvolvidos. Nesta secção e na seguinte são descritas as arquitecturas interna e externa, respectivamente. A arquitectura interna refere-se à estrutura da informação dentro do Fedora e à forma como esta é inserida e mantida. A arquitectura externa refere-se aos serviços acoplados ao Fedora que providenciam funcionalidades necessárias ao uso do repositório.

5.1 Esboço do Repositório

O RODA usa dois esquemas de metainformação primários, o EAD (Encoded Archival Description) para guardar a metainformação descritiva e o PREMIS (PREservation Metadata: Implementation Strategies) para guardar metainformação de preservação. Para além destes são usados vários esquemas secundários para guardar metainformação técnica que não têm lugar no PREMIS, como o NISO Z39.87 para imagens fixas digitais. Para cada tipo de documentos que o repositório armazena pode haver, caso seja necessário, um esquema de metainformação técnica para refinar o PREMIS. É usado ainda outro tipo de metainformação estrutural para objectos digitais constituidos por vários ficheiros que devem estar organizados por uma determinada ordem e/ou agrupamento[12].

Na Secção 3.2.2 vimos que o *PREMIS Data Model* [OCLC and RLG, 2005] faz referência a 5 entidades: Objectos, Eventos, Agentes, Direitos e Entidades Intelectuais. No entanto esta última não é descrita pelo PREMIS, é apenas referenciada porque está fora do domínio de aplicação deste esquema que se prende apenas com questões de preservação. Para descrever as Entidades Intelectuais e as relações hierárquicas entre elas é usado o EAD (Secção 3.2.1).

O nível mais baixo de descrição do EAD é o do documento composto ou documento simples, os quais correspondem ao nível da entidade intelectual, para o qual o esquema PREMIS faz referência através da propriedade *linkingIntellectualEntityIdentifier* da representação. Consideramos que a representação é a concretização "física"da entidade intelectual e que uma entidade intelectual pode ter várias representações. No entanto, como a partir do PREMIS da representação mais recente conseguimos obter todas as outras e como só à última versão interessa aceder directamente a partir do EAD,

[12]No caso de estudo da AACC (Anexo A.1) foi usado o METS para guardar a ordem das páginas e a sua organização em volumes.

consideramos que o EAD apenas aponta para o PREMIS de uma única representação (figura 10).

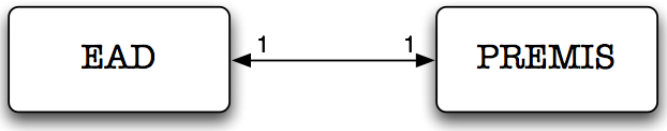


Figura 10: Relação do EAD com o PREMIS

O PREMIS guarda informação necessária à preservação de objectos digitais ao longo do tempo. Este esquema descreve representações, agregado de ficheiros necessários para renderizar a entidade intelectual, e descreve também os ficheiros que a constituem, que serão o verdadeiro alvo dos processos de preservação. No entanto o PREMIS é um esquema generalista e não guarda metainformação técnica específica sobre qualquer tipo de ficheiros. Logo, para completar a metainformação técnica dos vários tipos de ficheiros, é usado um esquema de metainformação técnica específico para cada tipo de ficheiros a ser preservado. No caso de estudo da AACC (anexo A) existem apenas imagens TIFF, por isso foi escolhido o esquema NISO Z39.87 (versão de 2002) para completar a metainformação guardada no PREMIS (figura 11). Este esquema é guardado dentro do campo *ObjectCharacteristics* do PREMIS.

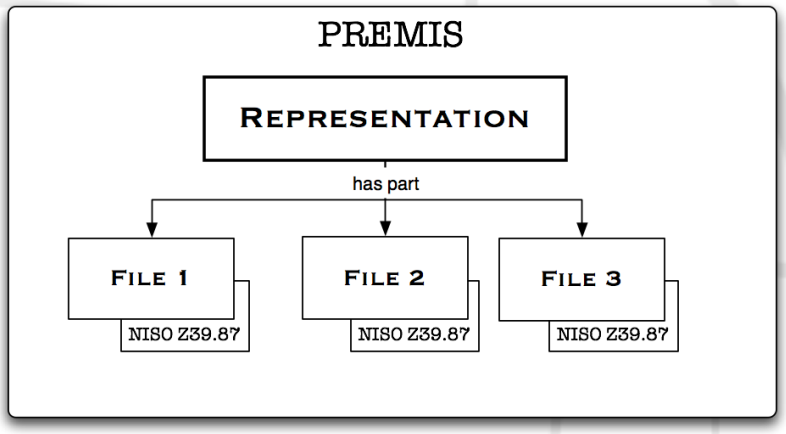


Figura 11: Esquema exemplo PREMIS com extensão do NISO Z39.87

Uma representação é descrita no esquema PREMIS por um elemento do tipo *Representation*. Este elemento poderá conter elementos do tipo *File*

ou do tipo *Representation*. Cada elemento *File* descreve um ficheiro especificando propriedades como o seu tamanho, formato, data de criação, localização no sistema de ficheiros (figura 12), etc.

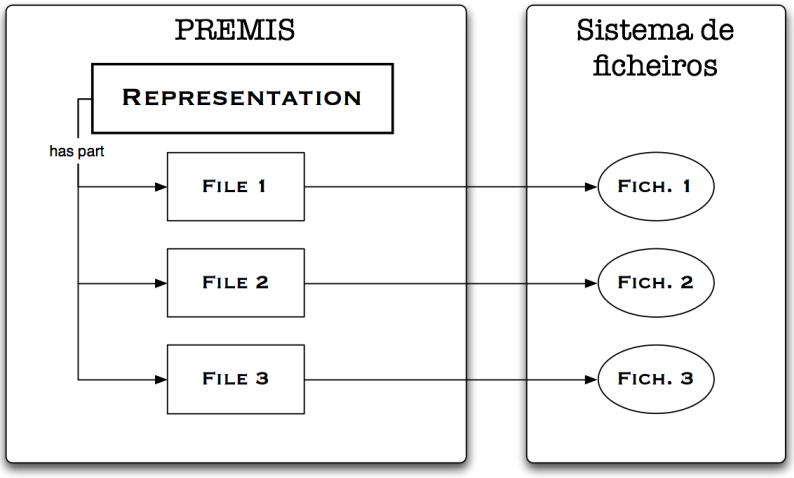


Figura 12: Relação entre o PREMIS e o sistema de ficheiros

Embora todos os ficheiros de uma representação estejam referenciados no elemento *Representation*, é necessária ainda metainformação estrutural que sirva a função de ponto de acesso a esta representação e que permita a navegação pelos vários ficheiros da representação de uma forma ordenada. Apesar do PREMIS ter algumas potencialidades para guardar metainformação estrutural, esta não seria suficiente para todos os casos possíveis. Foi, portanto, decidido utilizar outro esquema de metainformação, guardando no PREMIS uma referência para o mesmo com a propriedade *has root*. No nosso caso de estudo utilizamos o METS como esquema de metainformação estrutural. Assim, o METS guardará informação sobre a ordem das imagens (páginas) e também sobre a divisão em subgrupos (organização interna) dos ficheiros de uma representação (figura 13).

Com este esquema, os ficheiros que compõem uma representação são suficientes para assegurar a renderização completa da entidade intelectual, coadunando-se com a nossa definição de representação. O PREMIS aponta para estes ficheiros, formando uma representação e descreve a sua metainformação técnica e de preservação. O PREMIS aponta também para o EAD que contêm a metainformação descritiva. O EAD, no topo da pirâmide, guarda toda a informação descritiva e aponta para o PREMIS e também para o ficheiro que é o ponto de entrada da representação (neste caso o METS). Assim conseguimos uma estrutura flexível e segura, pois qualquer que seja a

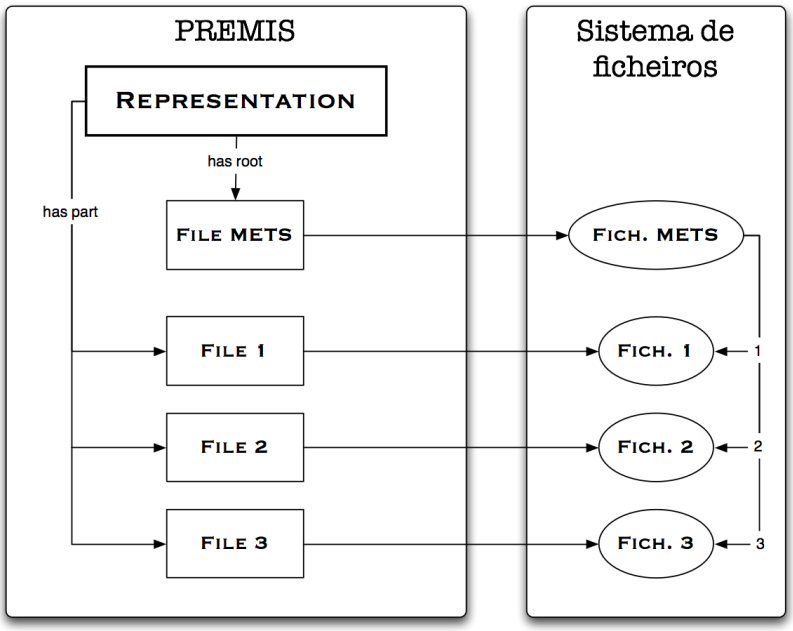


Figura 13: PREMIS referenciando o ponto de entrada (METS)

perspectiva que utilizarmos (descritiva ou de preservação) podemos sempre chegar à outra perspectiva e aos ficheiros (figura 14).

Note-se que existem mais dados nos esquemas que não estão representados neste esboço como os Eventos e os Agentes. Os Eventos registam as acções que ocorrem sobre os objectos e o resultado dessas mesmas acções (i.e. validação, conversão do formato, etc) e os Agentes são os responsáveis pela execução das acções que produzem os referidos Eventos. Os Eventos e Agentes são apenas referênciados pelos Objectos (de representação e de ficheiro).

5.2 Sobre o Fedora

No Fedora a unidade de informação é o objecto. Toda a informação (e metainformação) terá de fazer parte de um ou mais objectos. Um objecto está estruturado em 4 partes distintas (figura 15):

PID - identificador único persistente.

Descrição - propriedades e relações. Este componente do objecto é sempre necessário para a gestão interna dos objectos por parte do sistema. As propriedades são obrigatórias, as relações são opcionais.

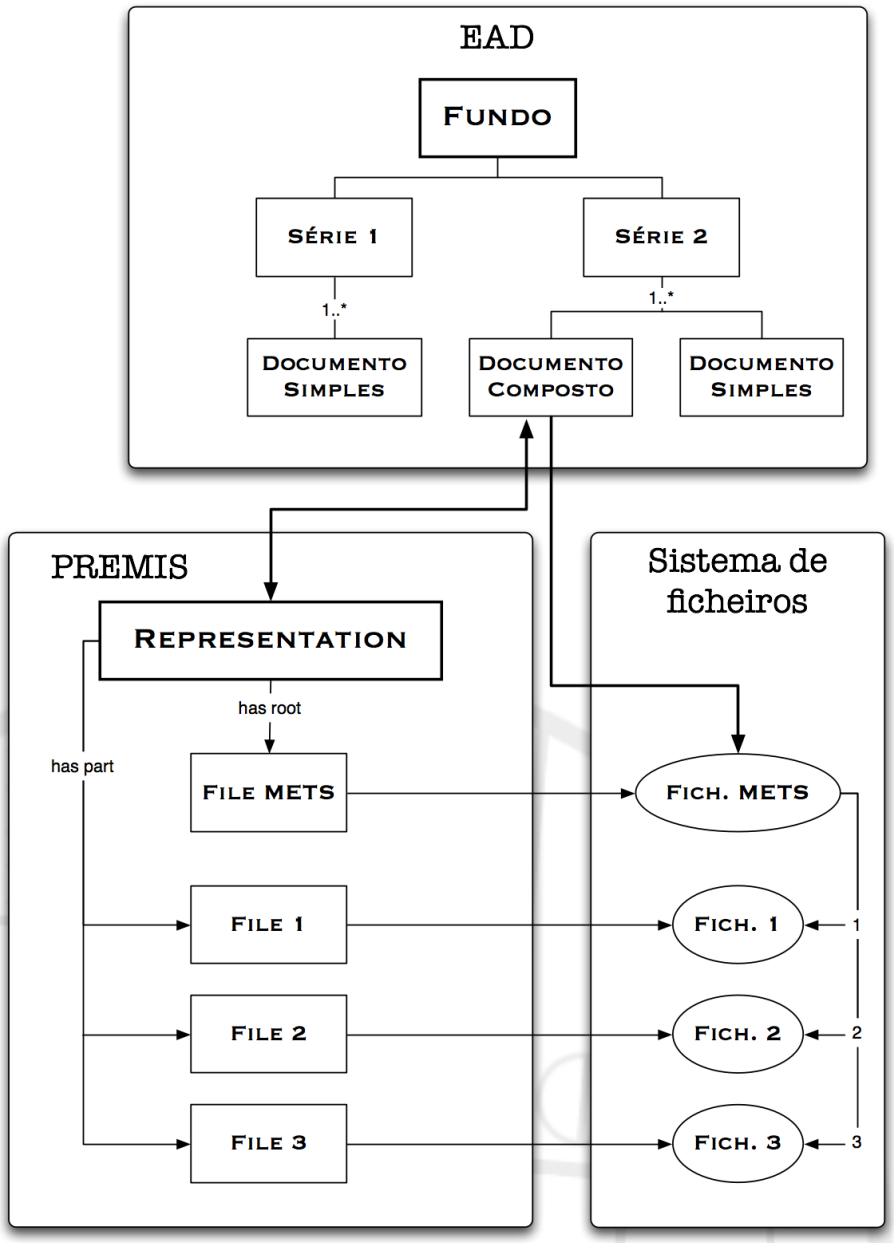


Figura 14: Esquema completo (excepto NISO Z39.87)

Itens - conjunto dos ficheiros de informação e/ou metainformação contidos no objecto (*datastreams*, na terminologia do Fedora). Um objecto tem no mínimo um ficheiro com metainformação no esquema Dublin Core (DC). Este ficheiro é incluído por omissão em cada objecto e contem

obrigatoriamente os campos `identifier` e `title`.

Serviços - conjunto de funcionalidades associadas ao objecto. Por omissão, a cada objecto criado é associado o serviço *Default Disseminator*. Este serviço permite que o objecto seja disseminado na sua forma mais básica, disponibilizando o acesso às propriedades e aos ficheiros. Outros serviços podem (e devem) ser associados aos objectos conforme as necessidades do próprio repositório ou da comunidade de interesse.

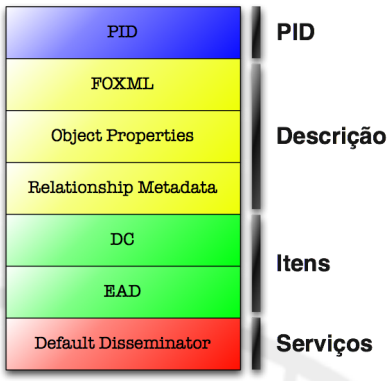


Figura 15: Estrutura interna do FoxML para um objecto de descrição

Propriedades básicas de um objecto

Status <enumeração>, `Active`, `Inactive` ou `Deleted`.

Label <texto livre>, ex: Descrição do Fundo AACC, PT-TT-AACC, PT/TT/A-ACC, etc.

Content Model <texto livre>, ex: `roda:d:f`, `roda:d:sr`, `roda:r`. Este campo é muito importante pois será usado para distinguir entre os vários tipos de objectos que o RODA irá lidar.

Created <automático> (Data de criação).

Modified <automático> (Data de alteração).

Owner <automático> (Utilizador a quem pertençe o objecto).

5.3 Tipos de Objectos

O RODA distingue 3 tipos de objectos: *Objectos de Descrição (OD)*, *Objectos de Representação (OR)* e *Objectos de Preservação (OP)*. Os *Objectos de Descrição* guardam informação descritiva (EAD), os *Objectos de Representação* contêm a representação de um objecto descrito num *Objecto de Descrição* e os *Objectos de Preservação* são usados para guardar metainformação de preservação (PREMIS).

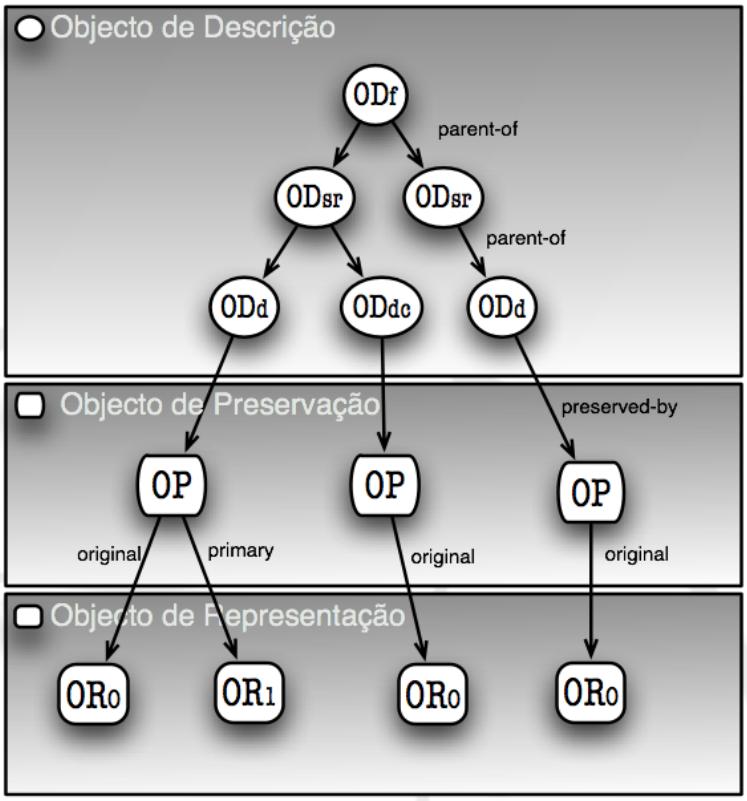


Figura 16: Objectos do RODA divididos por tipos

Todos os objectos relativos a um determinado fundo presente no repositório estão relacionados através do mecanismo de relações do Fedora (usando RDF[13]) formando uma árvore de descrição arquivística. Na raiz da árvore está um *OD* que descreve o fundo, conectados a este estão as descrições dos sub-fundos ou séries e estes por sua vez estão conectados às unidades de

[13] Resource Description Framework - http://www.w3.org/RDF

descrição necessárias para descrever adequadamente os documentos do fundo. Os *OP* estão ligados às folhas dos *OD* e guardam toda a metainformação de preservação relativa às representações da entidade intelectual descrita no *OD*. Por sua vez, os *OR* estão associados aos *OP* e contém as representações da entidade intelectual.

Para distinguir entre vários tipos de objectos é usada a propriedade *Content Model*. O valor desta propriedade é do tipo texto livre, portanto é possível introduzir um valor no formato que for mais conveniente. É então necessário criar uma sintaxe para os valores deste campo de forma a tornar fácil a identificação dos objectos com um algoritmo simples. Uma possível solução é usar um esquema análogo ao sistema de PIDs do Fedora que é da forma `namespace:id`. Na práctica tem-se `roda:d` para identificar *OD*, `roda:r` para identificar *OR* e `roda:p` para identificar *OP*. No caso dos *OD* pode-se ainda especificar o nível descritivo (Fundo, Série, Documento, etc.) com mais um nome à frente da identificação de *OD*; ex. `roda:d:f` para um fundo, `roda:d:sr` para uma série, `roda:d:dc` para um documento composto, etc. Este método simplifica tarefas comuns, como por exemplo identificar todos os fundos presentes no repositório; para tal basta procurar todos os objectos com a propriedade *Content Model* igual a `roda:d:f`. A tabela 1 resume todos os valores usados para esta propriedade nos diversos objectos existentes.

Tipo de Objecto	***Content Model***
Objectos de Descrição	`roda:d`
Fundo	`roda:d:f`
Subfundo	`roda:d:sf`
Classe	`roda:d:c`
Subclasse	`roda:d:sc`
Série	`roda:d:sr`
Subsérie	`roda:d:ssr`
Documento Composto	`roda:d:dc`
Documento Simples	`roda:d:d`
Objectos de Preservação	`roda:p`
Objectos de Representação	`roda:r`
Imagens com METS de estrutura	`roda:r:digitalized_work`
Texto estruturado	`roda:r:structured_text`
Base de dados relacionais	`roda:r:relational_database`

Tabela 1: Tipos de objectos e respectivos valores para a propriedade *Content Model*

5.4 Ajustes à Metainformação

5.4.1 EADPART

O esquema de metainformação descritiva usado no RODA é o EAD [The Library of Congress, 2002b]. Um ficheiro EAD descreve um fundo na sua totalidade, no entanto para a estrutura interna do repositório foi decido usar um modelo de dados que consiste numa árvore de *Objectos de Descrição (OD)* em que cada um dos nodos da árvore contém um componente do EAD, algo que designamos por EADPART (ver figura 16). Um ficheiro EADPART descreve um nível de descrição, i.e. um fundo, uma série ou um documento simples. Os ficheiros EADPART são ficheiros XML com uma gramática derivada da gramática de um EAD. Um ficheiro EADPART para além de descrever apenas um nível de descrição, não pode conter todos os campos de um ficheiro EAD. Os campos permitidos são apenas os necessários tendo em consideração a natureza dos materiais a descrever (objectos digitais). Apresenta-se na tabela 2 uma listagem dos campos permitidos num documento EADPART e as limitações em relação à gramática oficial do EAD.

Identificação		
`<eadpart>`	`@otherlevel`	Este atributo é obrigatório e o seu valor é um dos seguintes `F`, `SF`, `C`, `SC`, `SR`, `SSR`, `DC`, `D`
`<unitid>`	`@countrycode` `@repositorycode`	Este elemento é único e obrigatório.
`<unititle>`		Pode conter apenas texto.
`<unitdate>`	`@normal`	O elemento não pode conter texto. O atributo `@normal` é obrigatório.
`<physdesc>` `<dimensions>`	`@unit`	Pode conter apenas texto e o atributo `@unit`.
`<physdesc>` `<physfacet>`	`@unit`	Pode conter apenas texto e o atributo `@unit`.
`<physdesc>` `<date>`	`@normal`	O atributo `@normal` é obrigatório e não pode conter texto, tal como o elemento `<unitdate>`.
`<physdesc>` `<extent>`	`@unit`	Pode conter apenas texto e o atributo `@unit`.
`<physdesc> <p>`		Pode conter apenas texto.
Contexto		
`<origination>`		Pode conter apenas texto.

<bioghist> <p>		Pode conter apenas texto.
<bioghist> <chronlist>		Pode conter vários <chronitem>.
<chronitem>		Tem que conter um elemento <date> e um <event> ou <eventgrp>. O <eventgrp> pode conter vários elementos <event>.
<custodhist> <p>		Pode conter apenas texto.
<acqinfo> <p>		Pode conter apenas texto.
Conteúdo Estrutura		
<scopecontent> <p>		Pode conter apenas texto.
<appraisal> <p>		Pode conter apenas texto.
<accruals> <p>		Pode conter apenas texto.
<arrangement> <p>		Pode conter apenas texto.
<arrangement> <table>		
Condições Acesso e Utilização		
<accessrestrict> <p>		Pode conter apenas texto.
<userestrict> <p>		Pode conter apenas texto.
<langmaterial>		Pode conter vários elementos <language>. Os elementos <language> podem conter apenas texto.
<phystech> <p>		Pode conter apenas texto.
<materialspec>		Pode conter apenas texto.
<physfacet>		Pode conter apenas texto (ver em cima).
<physloc>		Pode conter apenas texto.
<daogrp>		Pode conter 1 ou mais sub-elementos <daoloc>.
<otherfindaid> <p>		Pode conter apenas texto.
Documentação Associada		

<relatedmaterial> <p>		Pode conter apenas texto.
<bibliography> <p>		Pode conter apenas texto.
Notas		
<notes> <p>		Pode conter apenas texto.
Controlo de Descrição		
<processinfo> <p>		Pode conter apenas texto.
<descrules>		Este elemento aparece dentro do elemento <profiledesc> que por sua vez aparece dentro do elemento <eadheader> e este dentro do elemento <ead> que não existe nos nosso <eadpart>. É preciso ver se é possível passar esta informação para outro elemento. Neste momento este elemento não é suportado.
<processinfo> <date>		Pelo schema actual do EAD o elemento <processinfo> não pode conter como "filho"um elemento <date>. Podemos ter <date> como filho de <p> e este como filho de <processinfo>. Neste momento é possível ter <processinfo> <p> <date>.
<prefercite> <p>		Pode conter apenas texto.

Tabela 2: Elementos de um documento EADPART

5.4.2 PREMIS

O PREMIS é bastante permissivo em relação aos valores possíveis para vários campos chave, como `objectCategory`, `relationshipType`, `relationshipSubType`, etc. Como estes valores são essenciais para que seja possível validar o PREMIS é necessário definir vocabulários controlados para estes campos.

objectCategory o PREMIS Data Dictionary sugere `representation`, `file`

e `bitstream` para este campo. Esta sugestão será uma obrigatoriedade para o PREMIS do RODA.

relationshipType o PREMIS Data Dictionary sugere `structural` para referências a partes de um objecto e `derivation` para referências a objectos do qual este objecto derivou. Mais uma vez estas sugestões serão obrigatoriedades no RODA.

relationshipSubType PREMIS Data Dictionary sugere vários valores, entre eles `has root`, `has part`. Tal como nos casos anteriores iremos seguir estas sugestões como obrigatórias. `has root` será usado para referenciar o ficheiro que é o ponto de entrada de uma representação, caso esta tenha mais do que um ficheiro. `has part` será usado para referenciar um ficheiro que faz parte de uma representação, mas não é o ponto de entrada.

eventType o PREMIS Data Dictionary sugere `capture`, `compression`, `deaccession`, `decompression`, `decryption`, `deletion`, `digital signature validation`, `dissemination`, `fixity check`, `ingestion`, `message digest calculation`, `migration`, `normalization`, `replication`, `validation`, `virus check`. No RODA, e até à data, existem apenas dois eventos: o evento de ingestão e o de verificação de integridade; para estes são usados os termos `ingestion` e `fixity check`, respectivamente.

5.5 Conteúdos dos objectos

Para cada tipo de objecto do repositório, *OD*, *OR* e *OP*, é necessário definir também os ficheiros que cada um deverá conter.

Para o caso dos *OD* os ficheiros são:

RELS-EXT - ficheiro RDF com informação sobre as relações entre o objecto e outros objectos do repositório (ex. `<info:fedora/roda:1> <roda:parent-of> <info:fedora/roda:2>`).

DC - ficheiro com metainformação Dublin Core em formato XML. Este ficheiro está presente em todos os objectos Fedora por omissão. No nosso repositório apenas irá conter o titulo (title) e o identificador (identifier) que são os elementos mínimos exigidos pelo Fedora.

EADPART - ficheiro com metainformação descritiva EAD em formato EADPART (XML). Este formato será basicamente um elemento `<c>` do EAD renomeado para `<eadpart>` com os mesmos

atributos e elementos do elemento `<c>`. A árvore de descrição do RODA será uma árvore de objectos Fedora contendo ficheiros no formato EADPART. A hierarquia é obtida à custa de relações entre estes objectos.

Para o caso dos *OP* os ficheiros são:

RELS-EXT - ficheiro RDF com informação sobre as relações entre o objecto e outros objectos do repositório (ex. `<info:fedora/roda:1> <roda:parent-of> <info:fedora/roda:2>`).

DC - ficheiro com metainformação Dublin Core em formato XML. Este ficheiro está presente em todos os objectos Fedora por omissão. No nosso repositório apenas irá conter o titulo (title) e o identificador (identifier) que são os elementos mínimos exigidos pelo Fedora.

PREMIS+ - ficheiro(s) com metainformação de preservação PREMIS (figura 17).

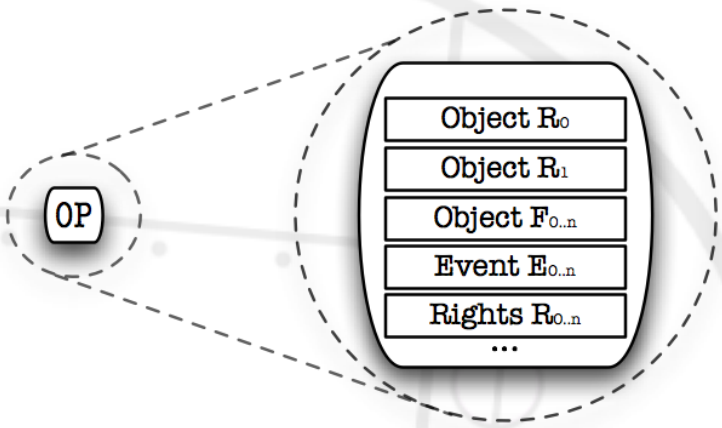


Figura 17: Conteúdo de um objecto de preservação

Para o caso dos *OR* os ficheiros são:

RELS-EXT - ficheiro RDF com informação sobre as relações entre o objecto e outros objectos do repositório (ex. `<info:fedora/roda:1> <roda:parent-of> <info:fedora/roda:2>`).

DC - ficheiro com metainformação Dublin Core em formato XML. Este ficheiro está presente em todos os objectos Fedora por omissão. No nosso repositório apenas irá conter o titulo (title) e o identificador (identifier) que são os elementos mínimos exigidos pelo Fedora.

FICHEIRO+ - ficheiro(s) que compõem a representação (figura 18).

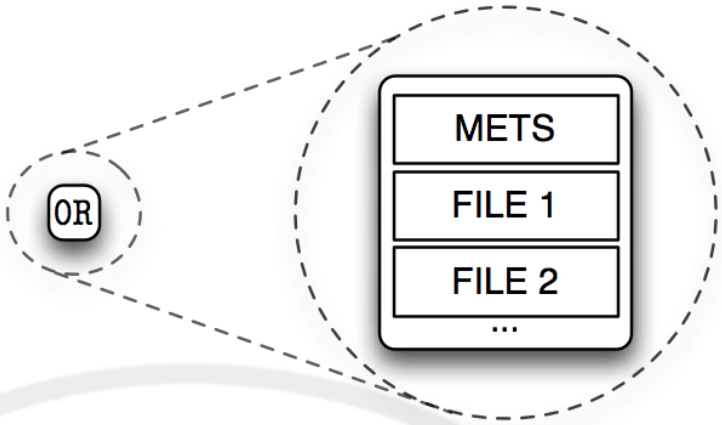


Figura 18: Conteúdo de um objecto de representação

5.6 Relações entre Objectos Fedora

As relações entre os vários objectos estão descritas no ficheiro RELS-EXT de cada um dos objectos que possui ligações a outros. Este ficheiro está no formato RDF e descreve as relações com outros objectos através de triplos (ex. `<info:fedora/roda:1> <roda:parent-of> <info:fedora/roda:2>`).

As relações entre os objectos descritos acima serão as seguintes:

parent-of - relação entre dois *OD*, permite estabelecer a relação hierárquica entre os componentes descritivos de um EAD.

preserved-by - relação entre um *OD* e um *OP*.

representation-(original/primary/alternative) - relação(ões) entre um *OP* e os respectivos *OR*. Esta relação pode ter três nomes diferentes porque é necessário distinguir entre a representação original, a representação principal (i.e. aquela que deverá ser disponibilizada ao consumidor por omissão) e eventuais representações alternativas.

5.7 Tipos de representações

Tal como foi mencionado na secção 2, o protótipo RODA faz preservação de três tipos de objectos digitais. Texto estruturado em formato PDF/A [pdf, 2007], imagens fixas bidimensionais em formato TIFF [Adobe, 2002] e bases de dados relacionais num formato BDML [Henriques et al., 2002] (XML).

Uma representação é uma "materialização" de um objecto digital. Dentro do repositório e do SIP RODA cada tipo de representação tem um identificador, para que o repositório e o ingestor possam saber de que tipo é a representação para "chamar"os procedimentos adequados para lidar com uma determinada representação.

Este identificador é o valor da propriedade *Content Model* mencionada na secção 5.3. Os valores para esta propriedade relativos aos três tipos de representações são apresentados na tabela 1.

Imagens estruturadas Este tipo de representação é composto por um ficheiro METS e vários ficheiros TIFF. O ficheiro METS contêm referências para todos os ficheiros TIFF e a estrutura desses mesmos ficheiros dentro da representação. Um exemplo de um ficheiro METS pode ser visto no anexo E.

Texto Uma representação deste tipo contém apenas um ficheiro PDF/A com o texto a ser preservado.

Base de dados Uma representação de uma base de dados relacional contém um ficheiro DBML [Henriques et al., 2002]. Este ficheiro contém a estrutura e os dados da base de dados. Campos da base de dados do tipo *blob*[14] serão armazenados em ficheiros separados referenciados no ficheiro DBML. Os dados binários guardados por cada campo do tipo blob serão copiados para um ficheiro distinto. Esses ficheiros fazem parte do AIP e serão alvo dos eventos de preservação apropriados ao seu tipo.

[14] *binary large object* é uma colecção de dados binários guardados como uma entidade única num SGDB

6 Descrição técnica

A arquitectura externa do projecto RODA refere-se a todo o desenvolvimento feito à volta do Fedora (ao invés do desenvolvimento feito na estrutura interna do Fedora). O diagrama completo da arquitectura do RODA é apresentado na figura 19. O *Back Office* e o *Front Office* foram colapsados num só componente: o *Office*. Foi definida uma API em Java, *Fedora API*, que abstrai os pormenores da ligação com o Fedora e os seus serviços. Esta API é utilizada pelo *RODA Office* para uso na web.

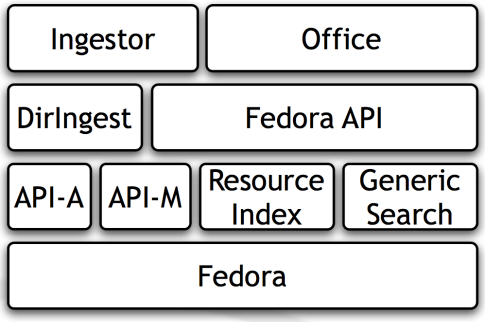


Figura 19: Arquitectura do RODA

O *Office* foi desenvolvido utilizando o *Google Web Toolkit*[15], sobre uma plataforma J2EE[16]. Na figura 20 é revelada a estrutura interna do *RODA Office*, dividida em *J2EE modules*[17]. As relações entre os módulos são relações

[15] O *Google Web Toolkit* (GWT) é uma projecto *open-source*, parte da iniciativa *Google Code*, para desenvolver aplicações *Ajax* na linguagem de programação *Java*. O GWT suporta um desenvolvimento cliente/servidor rápido e *debugging* em qualquer IDE de Java. Num passo seguinte de desenvolvimento, o compilador GWT traduz a aplicação Java corrente numa equivalente em JavaScript, que manipula programaticamente o HTML DOM do cliente web usando técnicas DHTML. GWT acentua soluções eficientes e reutilizáveis para desafios Ajax recorrentes, nomeadamente chamadas a procedimentos assíncronos, gestão da historia do cliente web, favoritos, e portabilidade nos vários clientes web[Wikipedia, 2007a].

[16] O J2EE (Java 2 Enterprise Edition) ou Java EE é uma plataforma de programação de computadores que faz parte da plataforma Java. Ela é voltada para aplicações multicamadas, baseadas em componentes que são executados em um servidor de aplicações. A plataforma Java EE é considerada um padrão de desenvolvimento já que o fornecedor de software nesta plataforma deve seguir determinadas regras se quiser declarar os seus produtos como compatíveis com Java EE. Ela contém bibliotecas desenvolvidas para o acesso a base de dados, RPC, CORBA, etc. Devido a essas características a plataforma é utilizada principalmente para o desenvolvimento de aplicações corporativas [Wikipedia, 2007b].

[17] Um *J2EE module* é uma unidade de *software* que consiste de um ou mais componentes do mesmo tipo e um descriptor desse tipo.

de herança, ou seja, se o módulo **A** tem uma relação com o módulo **B**, então **A** herda toda a funcionalidade do **B**. O módulo *Office* herda a funcionalidade de todos e expõe a mesma ao cliente. Assim é facil reutilizar funcionalidades já implementadas e disponibilizar novas funcionalidades ao cliente. Os módulos do *RODA Office* implementam as seguintes funcionalidades:

Edição de Metainformação implementado pelo módulo *Editor*, ver secção 6.2.1;

Pesquisa implementado pelo módulo *Search*, ver secção 6.3;

Navegação implementado pelo módulo *ListFonds*, ver secção 6.4;

Disseminação implementado no módulo *Browser*, ver secção 6.5.

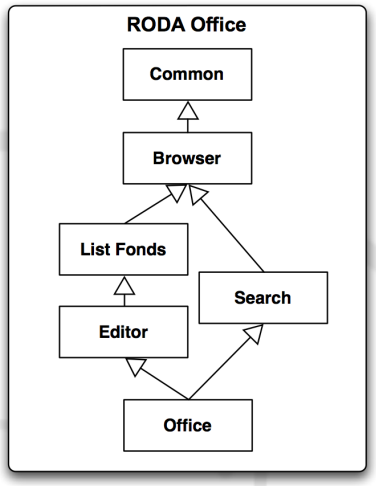


Figura 20: Hierarquia de Componentes do Office

O módulo de ingestão está fora do *RODA Office*, isto porque a funcionalidade de ingestão vai ser exposta programaticamente, de modo a poderem existir aplicações desktop que ajudam na criação de SIPs e na ingestão dos mesmos (ver a secção 6.1). No futuro esta funcionalidade deverá também existir no *RODA Office*.

O conjunto total destas funcionalidades completam o modelo OAIS. O Ingestor cumpre a função de Ingestão, o Editor de Metainformação é um dos mais importantes processos da Administração, a Pesquisa, Navegação e Disseminação cumprem em conjunto a função de Acesso. O Fedora ajuda em todas estas funcionalidades providenciando a base para todas e a funcionalidade de repositório de objectos digitais e metainformação.

6.1 Ingestão

A ferramenta de ingestão do RODA recebe como argumento um SIP e depois de o validar insere no repositório o objecto digital, com toda a sua metainformação. Um ficheiro SIP RODA é um ficheiro comprimido em formato `ZIP` que contém os ficheiros das representações, a metainformação e um ficheiro METS (envelope) que é descrito de seguida.

6.1.1 Diringest

O serviço Diringest[The Fedora Project, 2007] faz parte da *Fedora Service Framework*. Controi objectos Fedora a partir de SIPs `Diringest` e insere esses objectos no repositório. É exposto como um *webservice* REST que aceita ficheiros via HTTP POST e (caso a ingestão seja bem sucedida) devolve um documento XML com a lista de PIDs dos objectos ingeridos.
Um SIP `Diringest` é um ficheiro ZIP com um ficheiro METS e todos os ficheiros referenciados por este. Um exemplo de um ficheiro METS de um SIP `Diringest` pode ser consultado no anexo D.
O serviço Diringest é usado pelo Ingestor do RODA como irá ser descrito na secção 6.1.3.

6.1.2 SIP - METS

O esquema de metainformação METS [The Library of Congress, 2006c] é usado no RODA como metainformação estrutural para representações com vários ficheiros e como "envelope"para os SIP (*Submission Information Package*).

Nesta secção é descrita a estrutura de um ficheiro METS de um SIP RODA. O SIP RODA é um pacote que para além da representação a ser ingerida contém toda a metainformação referente à mesma. Contém metainformação descritiva, de preservação e técnica. A representação contida no SIP pode ainda conter metainformação estrutural nos casos em que esta é necessária.

A estrutura do ficheiro METS que deve constar de um SIP RODA é descrita de seguida. Na figura 25 pode-se ver a estrutura geral de um ficheiro METS. Para consultar um ficheiro METS de um SIP RODA de exemplo, ver anexo C.

<dmdSec> secção de metainformação descritiva (Figura 21). Nesta secção deve haver uma referência para o ficheiro EADPART que contém a descrição do objecto digital presente no SIP.

```
<dmdSec ID="PT-TT-AACC-1-1.EAD" ADMID="R2006.01.0001.premis.AMDSEC">
    <mdRef LOCTYPE="URL" MDTYPE="OTHER"
        xlink:href="PT-TT-AACC-1-1.ead" LABEL="roda:d:dc"/>
</dmdSec>
```

Figura 21: Excerto do METS de um SIP RODA - EADPART

<amdSec> secção de metainformação administrativa (figura 22). Nesta secção deve haver uma referência para o ficheiro PREMIS que contém metainformação de preservação e técnica.

```
<amdSec ID="R2006.01.0001.premis.AMDSEC">
    <techMD ID="PT-TT-AACC-1-1.PREMIS">
        <mdRef LOCTYPE="URL" MDTYPE="PREMIS"
            xlink:href="R2006.01.0001.premis"/>
    </techMD>
</amdSec>
```

Figura 22: Excerto do METS de um SIP RODA - PREMIS

<fileSec> secção de listagem de ficheiros (Figura 23). Nesta secção deve haver tantos elementos `<file>` como os ficheiros das contidos na representação.

```
<file ID="F2006.01.0001" MIMETYPE="text/xml"
    CHECKSUMTYPE="MD5" CHECKSUM="d20627bc137dbcb616af020ec7d45ded">
    <FLocat LOCTYPE="URL" xlink:href="R2006.01.0001/F2006.01.0001.METS.xml"/>
</file>
```

Figura 23: Excerto do METS de um SIP RODA - Ficheiros

<structMap> secção de estrutura (Figura 24). Nesta secção deve haver um elemento `<div>` para cada representação presente no SIP. Em cada um desses elementos deve aparecer a lista de apontadores (`<fptr>`) para os ficheiros que compõem a respectiva representação.

6.1.3 Processo de Ingestão

O processo de ingestão consiste basicamente em receber o SIP, validar o mesmo e inserir no repositório o objecto digital com toda a sua metainformação.
De seguida são descritos os passos do processo de ingestão.

1. Validação - Antes que um SIP possa ser ingerido vários testes são feitos para garantir que o SIP está completo e bem formado. Depois de extrair os conteúdos do ficheiro `ZIP` o ingestor realiza os seguintes testes.

```
<div ID="R2006.01.0001" DMDID="PT-TT-AACC-1-1.EAD"
        ADMID="R2006.01.0001.premis.AMDSEC" TYPE="digitalized_work">
    <fptr FILEID="F2006.01.0001" />
    <fptr FILEID="F2006.01.0001.0000001" />
    ...
</div>
```

Figura 24: Excerto do METS de um SIP RODA - Estrutura

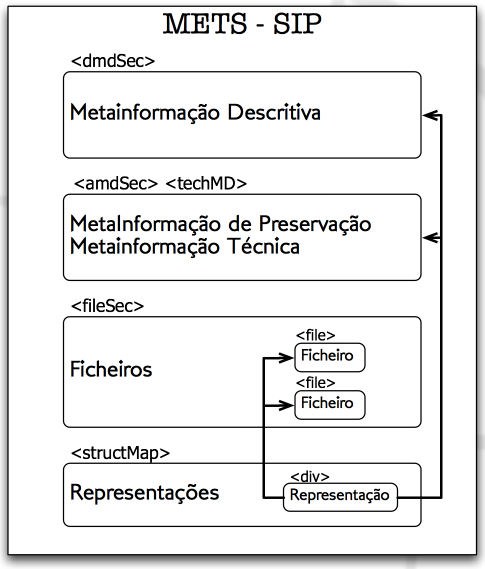


Figura 25: Estrutura de um ficheiro METS

(a) Verificação de vírus - Os ficheiros contidos no SIP são verificados para detectar a existência de vírus. O anti-vírus usado actualmente é o Grisoft AVG [Grisoft, 2007].

(b) Verificação da sintaxe do METS do SIP - A sintaxe é verificada automaticamente através de um parser gerado pelo XMLBeans [Apache, 2007] a partir do *schema* de um ficheiro METS.

(c) Verificação dos conteúdos do SIP - Depois de verificar que o ficheiro METS está sintaticamente correcto, verifica-se se todos os ficheiros referenciados existem e se todos os ficheiros que existem estão referenciados.

(d) Verificação da integridade dos ficheiros - Os ficheiros referênciados no ficheiro METS veem acompanhados dos atributos `CHECKSUMTYPE` e `CHECKSUM`. Os valor do *checksum* dos ficheiros é calculado e comparado com os valores no ficheiro METS garantindo assim que os ficheiros ingeridos são os mesmos que foram inseridos no SIP.

(e) Verificação da referência para o ficheiro EADPART - O ficheiro EADPART referenciado no ficheiro METS tem que existir no SIP.

(f) Verificação da sintaxe do EADPART - A sintaxe é verificada automaticamente através de um parser gerado pelo XMLBeans a partir do *schema* de um ficheiro EADPART.

(g) Verificação da referência para o ficheiro PREMIS - O ficheiro PREMIS referenciado no ficheiro METS tem que existir no SIP.

(h) Verificação da sintaxe do PREMIS - A sintaxe é verificada automaticamente através de um parser gerado pelo XMLBeans a partir do *schema* de um ficheiro PREMIS.

(i) Verificação da existência de pelo menos uma representação - Um SIP contém um objecto digital que pode estar "materializado"em mais do que uma representação.

(j) Para cada representação referenciada no METS:

   i. Validação das referências para as secções do EADPART e do PREMIS - Cada representação descrita no ficheiro METS tem que fazer referência ao ficheiro EADPART e ao ficheiro PREMIS.

   ii. Validação da representação - Cada representação é verificada dependendo do tipo de representação. Os passos de verificação para os três tipos de representações são descritos nas secções 6.1.4, 6.1.5 e 6.1.6.

2. Ingestão - Depois de feitas todas as validações o SIP RODA é preparado e ingerido pelo repositório.

   (a) Criação do SIP `Diringest` - Os conteúdos do SIP são transformados em objectos Fedora; um *Objecto de Descrição*, um *Objecto de Preservação*, um ou mais *Objectos de Representação* e é criado um novo SIP `Diringest` que é ingerido pelo serviço `Diringest` do Fedora. Para mais informação sobre o serviço `Diringest` ver secção 6.1.1;

   (b) Registo do evento de ingestão - Depois de ingerido o SIP, é registado um evento de preservação correspondente ao processo de ingestão e adicionado ao *Objecto de preservação* ingerido.

   (c) Associação ao Fundo - O *Objecto de Descrição* criado é automaticamente associado a um fundo que por sua vez está associado ao produtor que ingeriu o SIP.

   (d) Normalização da representação - Caso a representação ingerida não se encontre no formato de preservação correspondente ao seu tipo é criada uma representação normalizada, derivada da representação original. Esta tarefa é feita pelo Gestor de Normalização (ver secção 6.1.7).

6.1.4 Validação de uma representação `digitalized_work`

De seguinda são descritos os passos de validação de representações do tipo `roda:r:digitalized_work` (como as do fundo AACC - ver anexo A).

1. Verificação da existência de 1 ficheiro METS - Uma representação do tipo `roda:r:digitalized_work` tem que conter um ficheiro METS que contém a metainformação estructural da representção.

2. Verificação da existência de pelo menos 1 ficheiro de imagem - Uma representação tem que conter no minimo uma imagem.

3. Verificação da sintaxe do METS - A sintaxe é verificada automaticamente através de um parser gerado pelo XMLBeans [Apache, 2007] a partir do schema de um ficheiro METS.

4. Verificação dos conteúdos da representação - Verifica-se se todos os ficheiros referenciados existem e se todos os ficheiros que existem estão referenciados.

5. Verificação das imagens - Todas as imagens são verificadas quanto à sua integridade e validade. Para a validação das imagens é usada a ferramenta JHOVE [JSTOR et al., 2003].

6.1.5 Validação de uma representação `structured_text`

De seguinda são descritos os passos de validação de representações do tipo `roda:r:structured_text`.

1. Verificação da existência de 1 ficheiro PDF - Uma representação do tipo `roda:r:structured_text` tem que conter um, e apenas um, ficheiro PDF.

2. Verificação do PDF - O ficheiro PDF é verificado quanto à sua integridade e validade. Para a validação é usada a ferramenta JHOVE.

6.1.6 Validação de uma representação `relational_database`

De seguinda são descritos os passos de validação de representações do tipo `roda:r:relational_database`.

1. Verificação da existência de 1 ficheiro DBML (XML) - Uma representação do tipo `roda:r:relational_database` tem que conter um, e apenas um, ficheiro DBML e eventualmente outros ficheiros que fazem parte da base de dados.

2. Verificação do DBML (XML) - O ficheiro DBML é verificado quanto à sua integridade e validade. Para a validação é usada a ferramenta JHOVE [JSTOR et al., 2003]. Actualmente ainda não são verificados outros ficheiros associados à base de dados.

A ingestão de bases de dados relacionais implica um passo anterior a todo o processo de ingestão que consiste em extrair a base de dados do ambiente de exploração (Access, SQL Server, etc) para o formato normalizado (DBML [Henriques et al., 2002]). Esta funcionalidade já foi desenvolvida numa ferramenta auxiliar que deverá no futuro ser integrada no criador de SIPs (figura 26).

6.1.7 Gestor de Normalização

O Gestor de Normalização é um serviço usado para criar representações normalizadas de representações já existentes. Por exemplo, o ingestor aceita representações originais de imagens em vários formatos de imagem (TIFF,

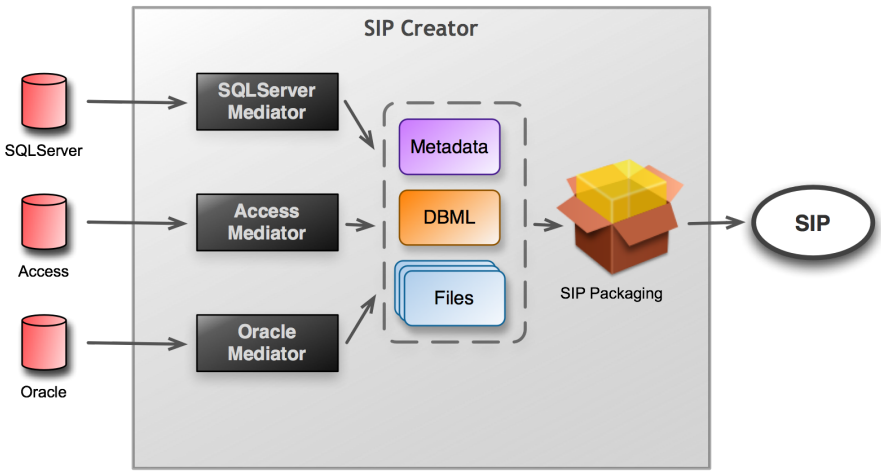


Figura 26: Esquema da criação do SIP de bases de dados

GIF, JPG, etc), no entanto, o formato normalizado para imagens é o TIFF [Adobe, 2002], portanto é necessário "pedir"ao Gestor de Normalização para criar uma versão normalizada da representação original que terá uma relação com o objecto de preservação já existente (criada pelo Gestor de Relações) e que será associada à representação original através de um evento PREMIS (criado pelo Gestor de Eventos).

6.2 Gestão

O componente de gestão é um dos componentes básicos do modelo OAIS. Este componente tem funcionalidades como a edição e gestão de metainformação descritiva, gestão de eventos de preservação, monitorização, gestão de utilizadores, etc. Neste momento encontram-se parcialmente implementadas duas funcionalidades básicas de gestão: a edição de metainformação descritiva e o evento de verificação de integridade.

6.2.1 Edição de Metainformação

O móldulo de edição de permite manipular a metainformação descritiva (ficheiros EADPART). Através do módulo de edição o arquivista pode criar ou alterar a estrutura de um fundo assim como criar ou alterar a descrição de qualquer componente da árvore de descrição. A edição da metainformação descritiva de um determinado componente de descrição é feita através da página apresentada na figura 27. Os campos são apresentados com a estrutura descrita na secção 5.4.1 e cada um dos campos pode ser editado bastando clicar no campo respectivo e alterar o texto que aparece na caixa de edição.

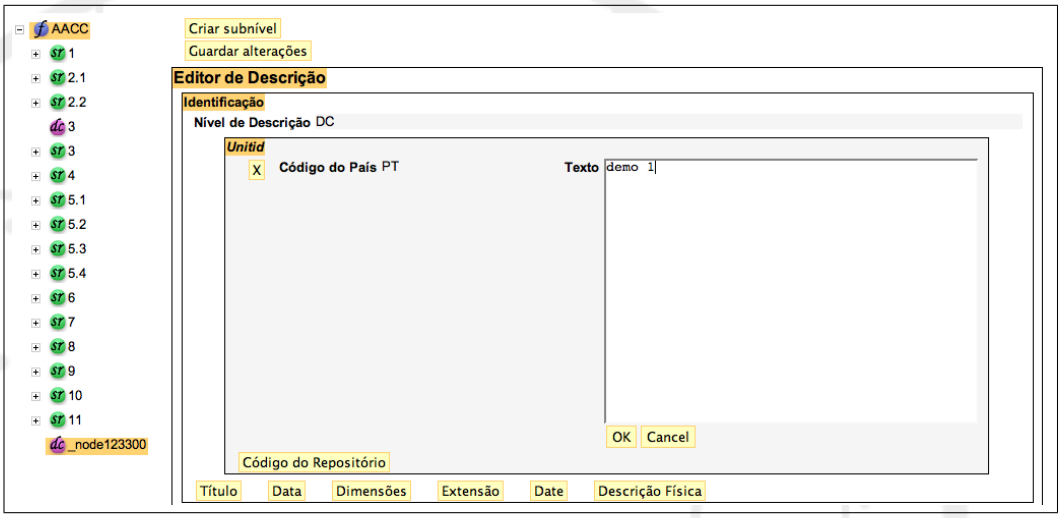


Figura 27: Editor EADPART

O botão `Guardar Alterações` guarda as alterações feitas alterando o *Objecto de Descrição* correspondente.

Novos níveis de descrição podem ser criados usando o botão `Criar subnível` e editando a decrição desse novo nível. Esta funcionalidade faz uso do Gestor de Relações para criar a relação entre o componente hierarquicamente superior e o novo componente (figura 27).

6.2.2 Eventos de preservação

O serviço de verificação de integridade é um serviço que verifica todos os ficheiros de todas as representações, comparando o resumo criptográfico do ficheiro em disco contra o armazenado no repositório.

Figura 28: PREMIS - Evento de verificação de integridade

Neste momento ainda não existe uma forma de espoletar este evento a partir do *Front Office*. Esta capacidade de lançar e monitorizar o estado dos eventos de preservação será adicionada ao *Front Office* no futuro.

6.3 Pesquisa

A forma mais directa de aceder à informação custodiada é recorrendo à funcionalidade de Pesquisa. A Pesquisa utiliza um serviço da *Fedora Service Framework*, o *Fedora Generic Search Service*. Este serviço permite as seguintes funcionalidades:

- Indexação dos objectos Fedora (registos FOXML), incluindo os conteúdos textuais das datastreams e os resultados de disseminadores.
- Procura no índice gerado.
- Plugin dos mecanismos de procura escolhidos, até agora o *Lucene* e o *Zebra*.

O mecanismo de procura escolhido nesta implementação foi o Lucene.

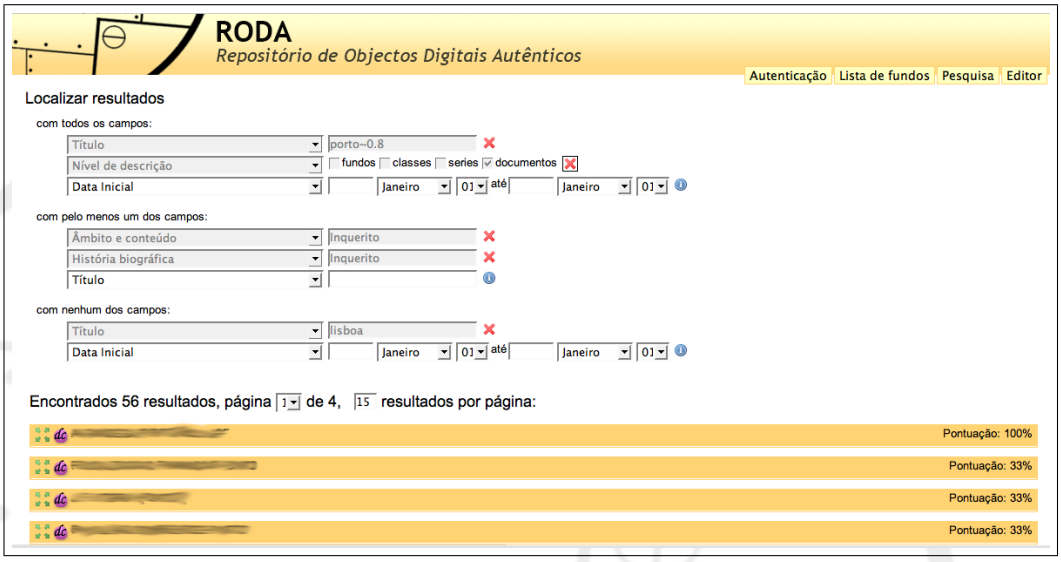


Figura 29: Pesquisa pelos vários campos

Um conjunto extenso de campos da metainformação descritiva estão disponíveis. A pesquisa é feita sempre mediante um campo. Os campos de pesquisa estão separados em três grupos (ver figura 29):

- **campos de procura exclusiva -** em que todos os campos têm de ser satisfeitos
- **campos de procura alternativa -** em que pelo menos um dos campos tem de ser satisfeito

- **campos de procura negativa -** em que nenhum dos campos pode ser satisfeito

Qualquer um destes grupos pode ter várias condições sobre qualquer um dos campos. A maneira de inserir a condição pode ser definida nos parâmetros e designa-se por *picker*. Existem já definidos *pickers* para campos de texto, níveis de descrição, datas e intervalo de datas. O picker para campos de texto, o mais genérico, permite todas as funcionalidades da pesquisa Lucene:

- **Procura Wildcard**
  Para utilizar a procura *wildcard* de uma única letra usa-se o símbolo '?'. Assim, com a procura `document?` é possível encontrar palavras como documento e documenta.
  Para utilizar a procura wildcard de múltiplas letras usa-se o símbolo '*'. Assim, com a procura `document*` é possível encontrar palavras como documento e documentos.

- **Procura Fuzzy**
  A procura *fuzzy* é baseada no algoritmo da distância Levenshtein. Para a utilizar usa-se o símbolo '∼' no fim de uma palavra. Assim, com a procura `torto∼` é possível encontrar palavras como morto e tortos. É ainda possível adicionar uma parâmetro que especifica o grau de similaridade entre as palavras. Este valor está entre 0 e 1, com valores mais próximos de 1 só palavras com maior semelhança serão encontrados (exemplo: `torto∼0.8`). O parâmetro tem o valor de 0.5 por defeito.

- **Procura de Proximidade**
  É possível encontrar palavras que estão separadas por uma distância especifica. Para utilizar a procura de proximidade usa-se o símbolo '∼' no fim de uma frase. Por exemplo, para procurar por "objecto"e "digital"com uma separação máxima, entre os dois, de 10 palavras, dentro de um campo, usa-se a frase de procura: `"objecto digital"∼10`.

- **Procura num Intervalo**
  São procuradas todas as palavras que se encontrem entre duas palavras dadas, por ordem alfabética. Pode-se incluir ou excluir as palavras dadas da procura. Por exemplo, a frase de procura `[Alberto TO Vítor]` encontra todas as palavras que estão entre Alberto e Vítor, inclusivamente. A mesma procura, mas excluindo os limites, será:
  `{Alberto TO Vítor}`

- **Aumentar a relevância de uma palavra**
  Para aumentar a relevância de uma palavra na procura usa-se o símbolo '^' no fim da palavra, seguido do factor de relevância (um número). Quanto maior for este factor, mais relevante a palavra será para a procura. Exemplos:
  `objecto^4 digital`
  `"objecto digital"^4 "documento digital"`.

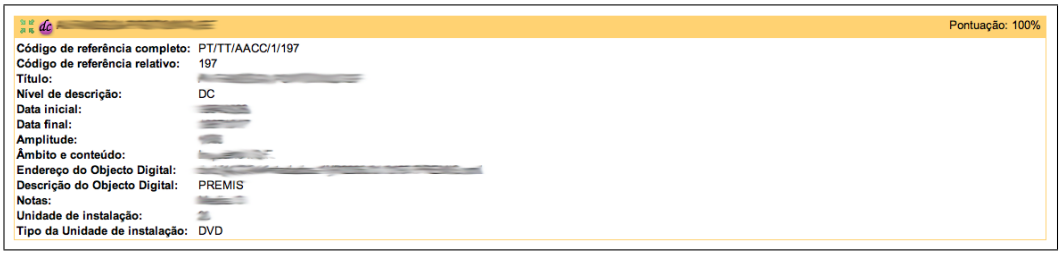


Figura 30: Expansão de um resultado da pesquisa

Qualquer resultado da procura pode ser expandido, de forma a visualizar imediatamente toda a informação indexada desse item (figura 30). Com um clique o navegador é chamado, expandindo o mesmo até que o objecto em causa esteja visível e a sua informação seja mostrada (figura 31).

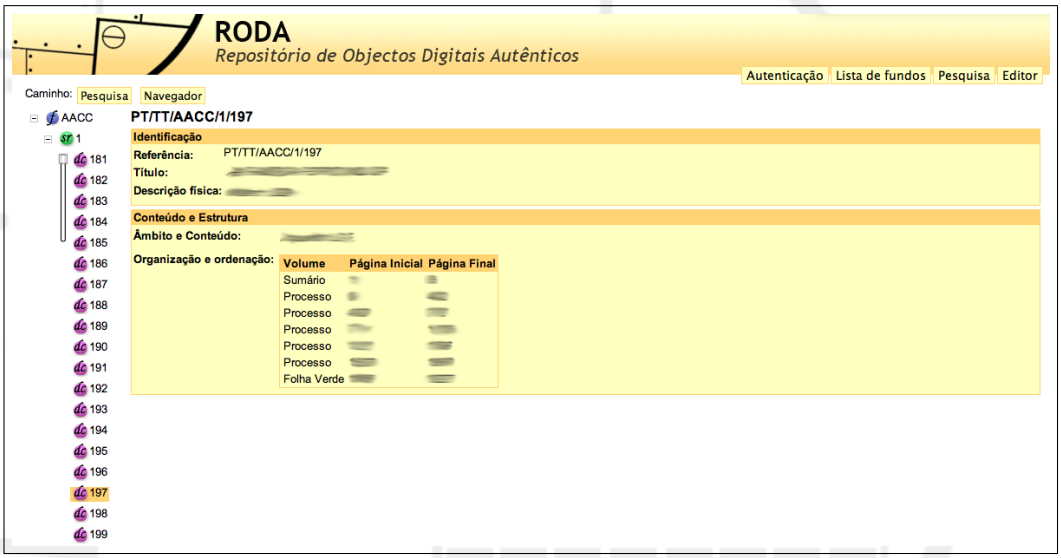


Figura 31: Resultado da pesquisa no navegador

6.4 Navegação

Outra forma de acesso à metainformação descritiva é pela Navegação. Aqui, a hierarquia da metainformação descritiva é exposta e o utilizador pode facilmente navegar por ela. Primeiro é dada uma listagem dos fundos disponíveis no repositório (figura 32). Esta lista pode ser facilmente diminuída usando o filtro disponível (no caso da figura, o filtro contém o valor de "alta autoridade"). O filtro remove da lista todos os fundos cuja a identificação ou o título não contenham todas as palavras do filtro. A pesquisa é feita de imediato, à medida que o valor do filtro é inserido. Deste modo é fácil encontrar o fundo desejado, mesmo que o repositório contenha um grande número de fundos.

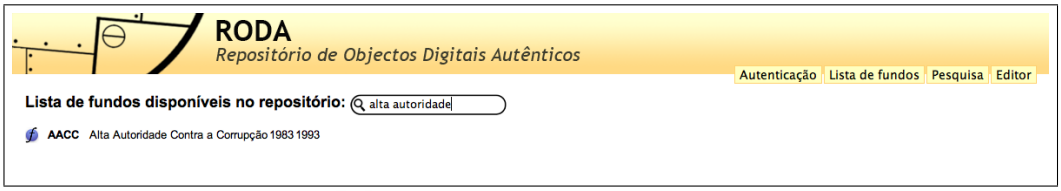


Figura 32: Listagem de todos os fundos disponíveis

Depois de escolher o fundo desejado na Lista de fundos, ele é mostrado no Navegador (figura 33). Do lado esquerdo do navegador é mostrada uma árvore com a hierarquia da metainformação descritiva. Como um fundo ou uma série pode ter milhares de filhos, é impraticável mostrar todos os subníveis de descrição de uma só vez. Por isso foi implementado um modo de restringir o número de filhos visíveis num momento por um valor parametrizável (por omissão 20). Para mover a janela de filhos visíveis é utilizado um elevador (ou *slider*), que se encontra por baixo do nível descritivo superior.

Do lado direito do Navegador é apresentada a metainformação do item escolhido do lado esquerdo. Por omissão a metainformação descritiva é apresentada. Um botão no topo direito pode ser usado para alternar entre a metainformação descritiva e a de preservação (figura 34).

No painel de metainformação de preservação, todos os objectos PREMIS do tipo Representação são mostrados. Estes listam todos os ficheiros que formam a representação, representações das quais derivam e eventos que actuaram sobre ele. Ao seleccionar o identificador de um ficheiro, representação, evento ou agente um painel com a metainformação de preservação do mesmo é apresentado (figuras 35 e 36).

Um último componente do Navegador é o painel de Disseminações, que se encontra por baixo da metainformação descritiva ou de preservação (ver figura 33 ou 34). Este componente será explicado na secção 6.5.

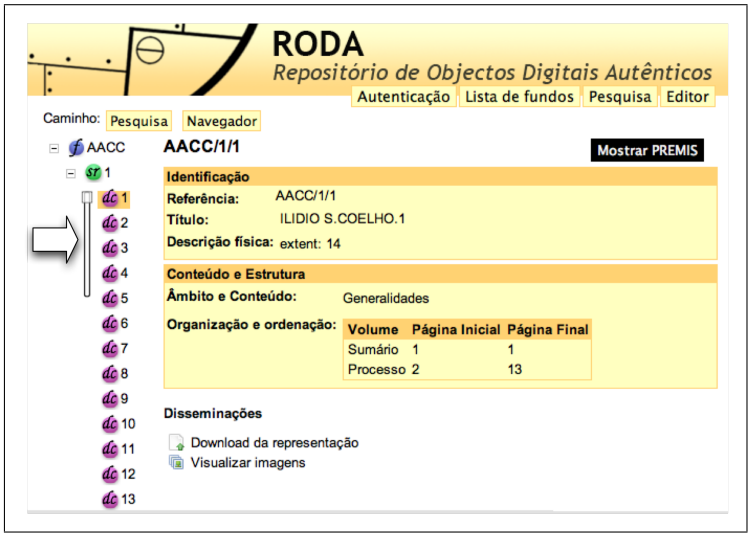


Figura 33: Navegação sobre um documento

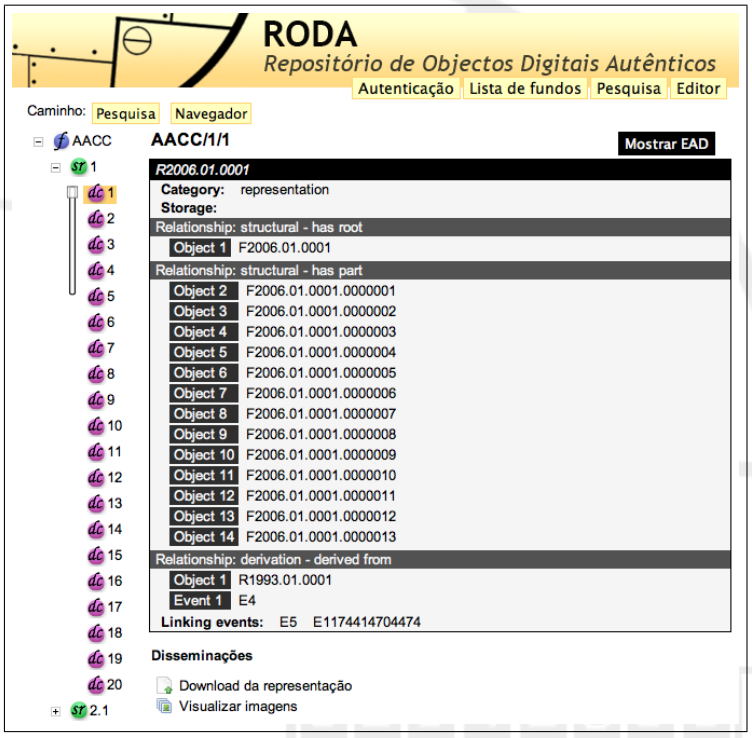


Figura 34: Metainformação de preservação no Navegador

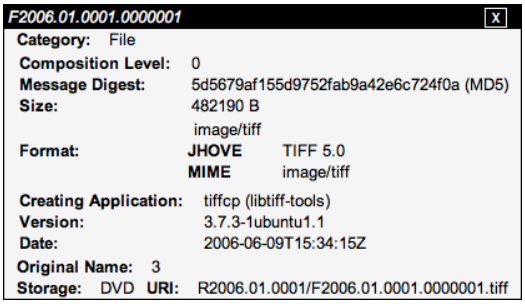


Figura 35: Painel com um objecto PREMIS

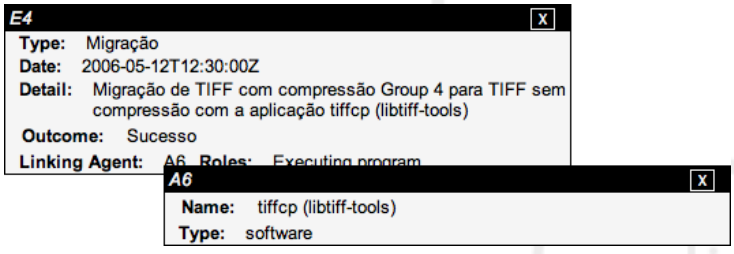


Figura 36: Painel com um evento e um agente PREMIS

6.5 Disseminação

O Fedora define um modelo de objecto digital genérico que pode ser usado para exprimir variados tipos de objectos, incluindo documentos, imagens, bases de dados, metainformação e muitas outras entidades. Para facilitar o acesso a estes objectos o Fedora oferece um mecanismo de associação de serviços a um objecto, para produzir conteúdo dinâmico ou computado a partir dos objectos digitais. O modelo é simples e flexível, de maneira a que variados tipos de objectos digitais possam ser criados e, no entanto, geridos e tratados de uma maneira consistente ([Fedora Project, 2005]).

Estes serviços associados aos objectos digitais são chamados de disseminadores. Internamente estes disseminadores apontam para um conjunto de serviços que são chamados pelo repositório para produzir "representações virtuais"do objecto. Uma "representação virtual"é um conteúdo que não está guardado explicitamente num objecto digital, mas que é produzido em tempo de execução ([Fedora Project, 2005]).

Para permitir isto, cada disseminador é associado a um objecto especial que contém uma descrição do serviço sob a forma de *Web Service Description Language* (WSDL). O repositório usa esta informação para fazer as chamadas aos serviços apropriados em tempo de execução e produzir as representações virtuais. Da perspectiva do cliente este sistema é transparente ([Fedora Project, 2005]).

Um disseminador é definido por dois objectos, uma **interface** e uma **implementação**.

Uma **interface** define um conjunto abstracto de métodos que um disseminador deve implementar. Um disseminador refere uma interface como forma de afirmar que este disseminador vai suportar estes métodos. Essencialmente, a interface define um "contracto de comportamento"que um ou mais objectos digitais podem subscrever ([Fedora Project, 2005]).

Uma **implementação** é a metainformação para o concreto mapeamento do serviço. Um disseminador refere uma implementação como forma de afirmar que este disseminador usa esta implementação concreta do serviço para executar os métodos definidos na interface. Uma implementação está relacionada com uma interface no sentido em que o primeiro define uma implementação dos métodos abstractos definidos no segundo ([Fedora Project, 2005]). Podem existir várias implementações para a mesma interface.

O *RODA Office* utiliza este mecanismo, criando *Visualizadores*[18] para cada definição de comportamento de um disseminador. Estes *Visualizadores*

[18]Um Visualizador é um componente que permite visualizar algo

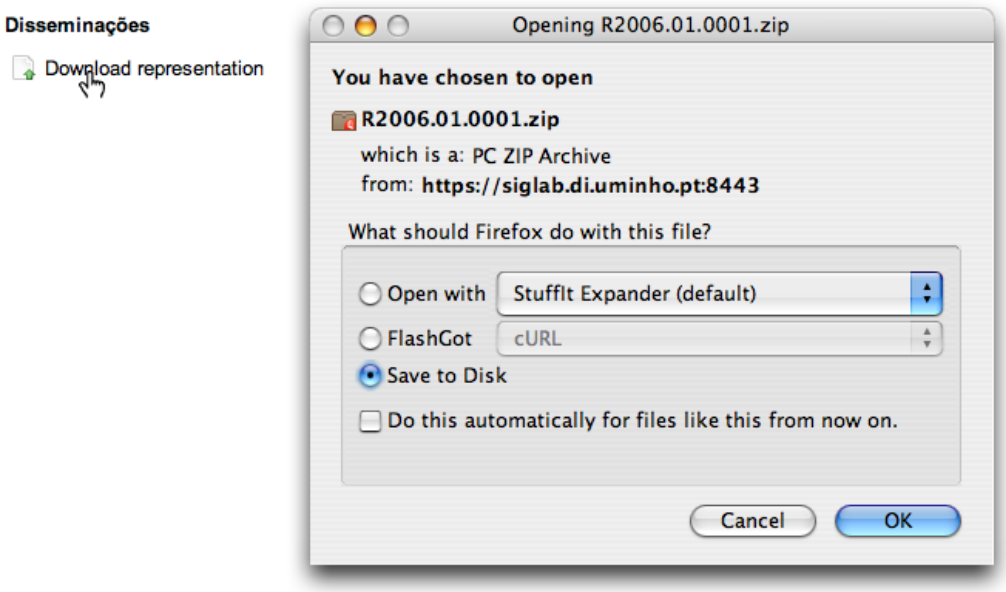


Figura 37: Disseminação na forma de um zip

utilizam os métodos de serviço definidos na definição de comportamento e apresentam o resultado ao utilizador.

Em forma de exemplo, foram criados no protótipo três disseminadores, úteis para as classes de objectos a preservar. Um deles, permite descarregar o AIP num arquivo comprimido, e pode ser associado a objectos de qualquer classe. Outro para visualizar imagens no Navegador em forma de *slideshow*. E ainda um disseminador de bases de dados relacionais que permite a navegação nos seus registos e relações.

O primeiro, permite descarregar a representação comprimida num ficheiro, caso as *datastreams* que constituem este disseminador forem múltiplas, ou a *datastream* em si, caso a representação for constituída por apenas uma *datastream*. O objecto de definição de comportamento deste disseminador define um único método, `oneOrZip()`, que devolve a *datastream* ou a representação comprimida. O objecto de mecanismo do comportamento define qual o serviço que deve ser chamado de modo a obter o resultado esperado. O *viewer* no *RODA Office* apenas apresenta um botão e uma legenda que possibilita descarregar o ficheiro devolvido (figura 37).

O segundo, o disseminador para imagens, deve ser associado a representações do tipo `digitalized-work` (imagens TIFF estruturadas por um METS) e permite a pré-visualização do conjunto de imagens em tamanho pequeno (figura 38) e o slideshow das mesmas imagens em tamanho grande (figura 39). A conversão das imagens do formato TIFF para o formato JPEG

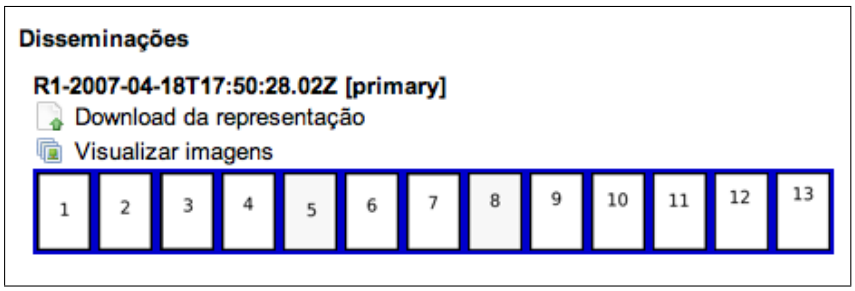


Figura 38: Representação apresentando a disseminação de imagens

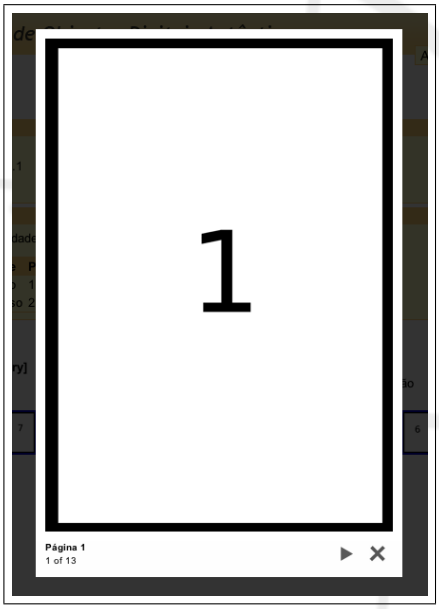


Figura 39: Visualizador de imagens

e a redução tamanho da imagem é transparente tanto para o utilizador como para o produtor pois o disseminador é automaticamente associado ao objecto digital no acto de ingestão.

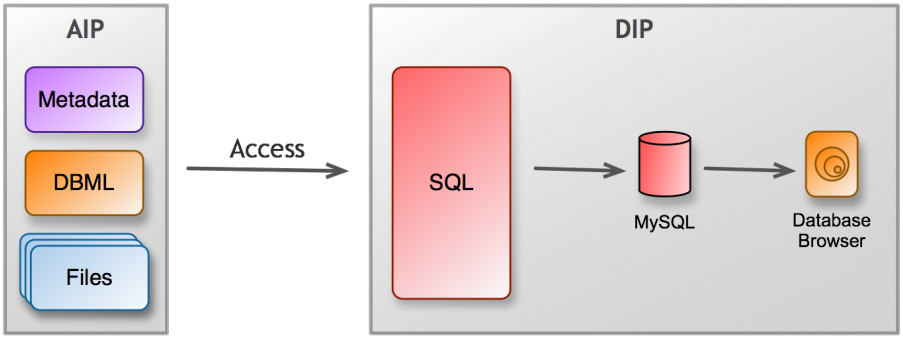


Figura 40: Esquema representativo da disseminação de uma base de dados

O último disseminador criado foi o de bases de dados. Este disseminador usa a representação da base de dados em DBML e os ficheiros por ele referidos para criar uma representação da base de dados em SQL (*Structured Query Language*). Posteriormente injecta este SQL num SGBD (Sistema de Gestão de Bases de Dados) de última geração. A navegação e procura na base de dados será facilitada por uma página web, representada na figura 40 por *Database Browser*.

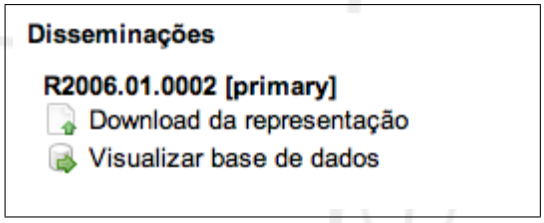


Figura 41: Representação apresentando a disseminação para bases de dados

Todo este processo é automático e transparente, tanto para o consumidor, como para o produtor e o administrador. Ao ingerir uma base de dados, o disseminador é criado automaticamente. Uma nova disseminação aparece disponível ao consumidor no navegador (figura 41). Ao clicar na disseminação o consumidor acede ao *Database Browser*.

O *Database Browser* apresenta inicialmente uma descrição global da base de dados (figura 42). A partir desta página inicial é possível aceder à lista de relações entre as tabelas ou ver a listagem e descrição de cada tabela (figura 43). A listagem completa dos registos de uma tabela está acessível a

partir desta última página. A listagem completa está segmentada em páginas, para que seja possível navegar mesmo que uma tabela contenha uma quantidade massiva de registos (figura 44). Caso o registo refira contenha uma chave estrangeira, ou seja, esteja relacionado com outro registo de outra tabela, é apresentado uma hiperligação para esse mesmo registo.

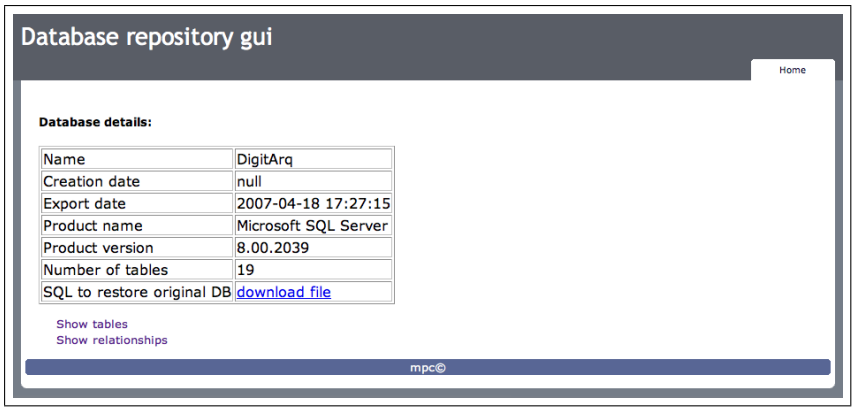


Figura 42: Visualizador de bases de dados - página inicial

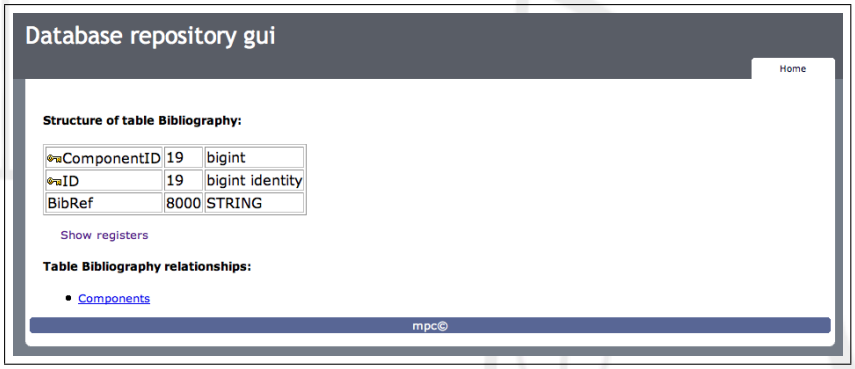


Figura 43: Visualizador de bases de dados - propriedades de uma tabela

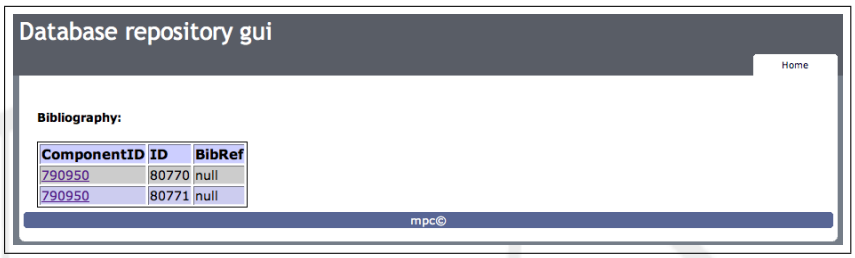


Figura 44: Visualizador de bases de dados - tabela com hiperligações para as chaves estrangeiras

7 Avaliação

Chegando ao término deste projecto, adequa-se agora uma avaliação do trabalho efectuado. Parte importante deste trabalho foi a investigação e escolha de normas e plataformas em que o repositório se iria basear.

As normas escolhidas são de grande importância pois a sua aceitação e uso generalizado garantem a sua persistência no tempo e assegura a interoperabilidade entre repositórios. As normas escolhidas têm de ser genéricas o suficiente para que se adeqúem aos objectivos do projecto, mas não demasiado genéricas pois a baixa entropia dificulta o seu manuseamento.

O PREMIS e o EAD adaptam-se perfeitamente aos objectivos deste projecto e demonstram-se fáceis de manusear. Tiveram, no entanto, que ser criados vocabulários controlados para permitir a especialização e validação destes esquemas (e.g. o atributo `@otherlevel` está restringido aos valores: `F`, `SF`, `C`, `SC`, `SR`, `SSR`, `DC`, `D`).

Outra decisão importante foi a escolha da plataforma de desenvolvimento. O Fedora correspondeu às espectativas de flexibilidade. Revelou, no entanto, alguma imaturidade na sua implementação, pois foram encontrados variados *bugs* em algumas funcionalidades essenciais. Isto obrigou a alguns trabalho a circundar os bugs ou mesmo a corrigi-los. Ainda não temos conclusões acerca da sua capacidade de ligar com um grande número de utilizadores e objectos digitais.

O modelo de dados do RODA no Fedora mostrou-se rápido e eficiente, devido à separação da entidade intelectual, da metainformação de preservação e da representação em diferentes objectos fedora, ligados pela sua relação em RDF. No entanto, para maximizar o aproveitamento destas ligações vai ser necessária uma pequena alteração no modelo de dados, dividindo a metainformação de preservação de diferentes representações da mesma entidade intelectual em vários objectos fedora. A mudança no modelo de dados tem algumas implicações nas funcionalidades já implementadas, razão porque se deixará a alteração no modelo de dados para trabalho futuro.

Por falta de tempo, alguns dos objectivos propostos para o prótotipo não foram atingidos:

- Suporte para a negociação na Ingestão
- A elaboração de uma ferramenta, enquanto módulo da anterior, capaz de se "acoplar"com sistemas de gestão documental existentes na AP

e assegurar funções de preservação digital numa perspectiva de gestão administrativa.

- Modelo(s) de financiamento que poderia(m) suportar o Arquivo Digital;

8 Trabalho Futuro

O projecto RODA será seguido pelo projecto RODA 2. O objectivo deste projecto é colocar em exploração uma solução de preservação digital, para os tipos de documentos já estudados no projecto RODA; texto estruturado, imagens bidimencionais e bases de dados relacionais. Para que esta solução seja uma realidade, do ponto de vista tecnológico, é necessário desenvolver muito trabalho no sentido de tornar o protótipo num produto estável e fácil de usar por todos os intervenientes no processo de preservação; o produtor, o administrador e o consumidor.

As tarefas mais relevantes que precisam de ser levadas a cabo no decorrer do projecto RODA 2 são as seguintes:

Efectuar testes de performance com o Fedora - o repositório irá lidar com grandes quantidades de informação, tanto em número de documentos como em volume de dados, e é preciso ter a certeza que a plataforma escolhida (o Fedora) é capaz de lidar com essas quantidades de informação e que consegue funcionar em condições aceitáveis perante tais requisitos.

Data Centre - a volume da informação é também uma preocupação no que respeita ao suporte para guardar e manter os materiais custodiados. E necessário investigar e implementar solução flexivel e escalável que possa acomodar de forma segura toda a informação à medida que vá sendo inserida no repositório.

Workflow de ingestão - é necessário implementar uma funcionalidade para auxiliar o processo de ingestão e toda a negociação anterior à ingestão propriamente dita dos documentos no repositório.

Criador SIPs - é necessário criar uma ferramenta para ajudar os produtores a criar SIPs. Esta ferramenta terá uma versão para a *web* e uma versão *standalone*.

Módulos de Gestão - é necessário desenvolver vários módulos de gestão que são essenciais para o funcionamento do repositório, como a gestão de utilizadores e politicas de permissões, a gestão de eventos de preservação, integração das funcionalidades do CRiB[19] [Ferreira et al., 2006].

Indentificadores persistentes - implementar um sistema de identificadores presistentes para os documentos.

[19]CRiB - Conversion and Recommendation of Digital Object Formats

Alteração do modelo de dados (PREMIS) - É necessário alterar a estrutura de dados interna no que respeita à metainformação de preservação por motivos de facilidade e eficiência na gestão deste tipo de metainformação. No modelo actual a metainformação de preservação respeitante a uma Entidade Intelectual, independentemente do número de representações do mesma, é guardada apenas num objecto. Esta abordagem torna dificil a gestão da metainformação de preservação por parte dos componentes de software. É preciso estudar um novo modelo para guardar a metainformação de preservação que torne mais simples a gestão da mesma.

Melhoramentos e correção de erros nos módulos já existentes - Procura, Edição de metainformação, Visualizador de imagens, Visualizador de bases de dados, Navegador.

Melhoramento do aspecto a interface gráfica - para tal iremos contar com um designer.

Referências

[pdf, 2007] (2007). The pdf/a competence center webpage. http://www.pdfa.org.

[Adobe, 2002] Adobe (2002). Tiff developer information site. http://partners.adobe.com/public/developer/tiff.

[Apache, 2007] Apache (2007). Página oficial do xmlbeans. http://xmlbeans.apache.org.

[Barbedo, 2006a] Barbedo, F. (2006a). Especificação de requisitos. Technical Report 41012-005, IAN/TT.

[Barbedo, 2006b] Barbedo, F. (2006b). Taxionomias de objectos digitais a integrar no roda. Technical Report 41012-006, IAN/TT.

[CCSDS, 2002] CCSDS (2002). Reference model for an open archival information system (oais) - blue book. *Washington: National Aeronautics and Space Administration*.

[Comissão ao Conselho et al., 2003] Comissão ao Conselho, Parlamento Europeu, Comité Económico e Social Europeu, and Comité das Regiões (2003). Papel da administração em linha (egoverno) no futuro da europa.

[Fedora Project, 2005] Fedora Project (2005). The fedora digital object model. http://www.fedora.info/download/2.0/userdocs/digitalobjects/objectModel.html.

[Ferreira, 2006] Ferreira, M. (2006). *Introdução à Preservação Digital: Conceitos, estratégias e actuais consensos*. Escola de Engenharia da Universidade do Minho. ISBN 972-8692-30-7, 978-972-8692-30-8.

[Ferreira et al., 2006] Ferreira, M., Baptista, A. A., and Ramalho, J. C. (2006). Crib : a service oriented architecture for digital preservation outsourcing. *Paper presented at the XATA - XML: Aplicações e Tecnologias Associadas*.

[Grisoft, 2007] Grisoft (2007). Página oficial do avg, versão grátis. http://free.grisoft.com.

[Henriques et al., 2002] Henriques, M., Libreotto, G., Ramalho, J., and Henriques, P. (2002). Bidirectional conversion between xml documents and relational data bases. *International conference on CSCW in design*.

[International Council on Archives, 1999] International Council on Archives (1999). *ISAD(G): General International Standard Archival Description*, 2nd edition.

[JSTOR et al., 2003] JSTOR et al. (2003). Página oficial jstor/harvard object validation environment. http://hul.harvard.edu/jhove.

[Lavoie, 2004] Lavoie, B. F. (2004). The open archival information system reference model: Introductory guide. *Digital Preservation Coalition*.

[NISO, 2006] NISO (2006). Niso z39.87-200x development page. http://www.niso.org/standards/standard_detail.cfm?std_id=731.

[OCLC, 1995] OCLC (1995). Dublin core metadata initiative. http://dublincore.org.

[OCLC and RLG, 2005] OCLC and RLG (2005). *PREMIS: Data Dictionary for Preservation Metadata*.

[RLG, 2002] RLG (2002). Rlg ead report card. http://www.rlg.org/ead-report-card.

[RLG EAD Advisory Group, 2002] RLG EAD Advisory Group (2002). *RLG Best Practive Guidelines for Encoded Archival Description*.

[Saramago, 2004] Saramago, M. (2004). Metadados para a preservação digital e aplicação do modelo oais. In *VIII Congresso da BAD*.

[Society of American Archivists, 2003] Society of American Archivists (2003). Society of american archivists home page. http://www.archivists.org.

[Society of American Archivists, 2006] Society of American Archivists (2006). Ead tools survey. http://www.archivists.org/saagroups/ead.

[The Fedora Project, 2007] The Fedora Project (2007). Fedora directory ingest service.

[The InterPARES Project, 2007] The InterPARES Project (2007). Interpares project. http://www.interpares.org.

[The Library of Congress, 2002a] The Library of Congress (2002a). Página oficial do ead versão de 2002. http://www.loc.gov/ead/.

[The Library of Congress, 2002b] The Library of Congress (2002b). Página oficial do ead versão de 2002. http://www.loc.gov/ead/.

[The Library of Congress, 2004] The Library of Congress (2004). Niso metadata for images in xml schema official web site. http://www.loc.gov/standards/mix.

[The Library of Congress, 2005] The Library of Congress (2005). Marc standards. http://www.loc.gov/marc.

[The Library of Congress, 2006a] The Library of Congress (2006a). Página oficial do metadata object description schema (mods). http://www.loc.gov/standards/mods.

[The Library of Congress, 2006b] The Library of Congress (2006b). Página oficial do mets. http://www.loc.gov/standards/mets.

[The Library of Congress, 2006c] The Library of Congress (2006c). Página oficial do mets. http://www.loc.gov/standards/mets.

[The Library of Congress, 2006d] The Library of Congress (2006d). Página oficial do premis. http://www.loc.gov/standards/premis.

[Wikipedia, 2007a] Wikipedia (2007a). Google web toolkit. http://en.wikipedia.org/wiki/Google_Code#Google_Web_Toolkit.

[Wikipedia, 2007b] Wikipedia (2007b). Informação sobre j2ee. http://pt.wikipedia.org/wiki/J2EE.

9 Glossário

Administrador
Termo genérico usado para designar os diferentes actores que intervêm na gestão e controlo das operações e eventos que ocorrem diariamente no repositório.

Consumidor
Actor (pessoa ou sistema) que interage com o repositório digital com o intuito de pesquisar e aceder a informação nele preservada.

Entidade Intelectual
Conjunto coerente de conteúdos que pode ser descrito como uma unidade ao nível de documento simples ou composto (ex. uma imagem, um processo, uma base de dados). Uma entidade intelectual pode conter outras entidades intelectuais, da mesma forma que os documentos compostos contêm documentos simples.

Ficheiro
Sequência de *bytes* com ordem e nome, reconhecida por um sistema operativo. Um ficheiro tem propriedades como permissões, tamanho e data da última modificação.

Metainformação Descritiva
Informação que identifica e descreve as propriedades intelectuais de um recursos ou conjunto de recursos, tendo em vista auxiliar a sua recuperação e facilitar a sua intelegibilidade.

Metainformação Estrutural
Informação de como uma Entidade Intelectual é construída ou organizada. Descreve, basicamente, a estrutura de uma Entidade Intelectual, como um livro, de maneira a que todas as páginas desse item possam ser apresentadas na ordem correcta. Metainformação Estrutural pode também incluír informação que suporte a navegação entre componentes de um objecto complexo. Exemplos incluem percorrer as páginas de um livro, saltar para um capítulo ou página particular ou alternar entre imagens e o texto correspondente.

Metainformação de Preservação
Informação necessária para suportar o processo de preservação digital. Esta informação inclui a documentação de todos processos que ocorrem dentro de um repositório digital com a finalidade de preservar uma

representação, e.g. ingestão, migrações, auditorias, refrescamentos, etc. Esta informação assegura a autenticidade da informação custodiada.

Metainformação Técnica
Informação que descreve os atributos ou propriedades físicas dos Ficheiros. Algumas propriedades na Metainformação Técnica são especificas do formato do Ficheiro (e.g. palete de cores associada com uma imagem TIFF) enquanto outras são independentes do formato (ou seja, são atributos de todos os Ficheiros independentemente do seu formato, e.g. o tamanho em bytes). A Metainformação Técnica permite assegurar a qualidade de uma migração pois permite comparar as propriedades que são comuns entre o formato alvo e o de origem, e.g. a resolução de uma imagem ao ser convertida de TIFF para JPEG2000 numa migração.

Migração
Estratégia de preservação que consiste num conjunto de tarefas definidas com o intuito de transferir periodicamente os materiais digitais de uma configuração *hardware/software* em perigo de obsolescência para uma outra mais actual.

Contrariamente a outras estratégias de preservação, a Migração não tem como objectivo manter a Representação no seu estado original. Em vez disso, converte a Representação de um formato em perigo de obsolescência para um outro que os computadores actuais podem interpretar.

Objecto de Descrição
Objecto Fedora que guarda metainformação descritiva de um Objecto de Representação ou de um conjunto de Objectos de Representação. Relaciona-se com um ou vários Objectos de Preservação. Tem uma relação hierárquica com outros Objectos de Descrição.

Objecto de Preservação
Objecto Fedora que contêm metainformação de preservação acerca das Representações contidas nos Objectos de Representação relacionados com este. Um Objecto de Preservação está sempre relacionado com um Objecto de Descrição.

Objecto de Representação
Objecto Fedora que contêm uma Representação. A metainformação de preservação desta Representação está contida no Objecto de Preservação relacionado com este. A metainformação descritiva desta Re-

presentação está contida no Objecto de Descrição relacionado com o dito Objecto de Preservação.

Objecto Digital

Unidade discreta de informação em formato digital. Se bem que a definição de objecto digital varia nos contextos associados ao RODA, por exemplo no PREMIS um objecto digital é uma Representação, Ficheiro, Bitstream ou Filestream e no Fedora um Objecto Digital é um conjunto de datastreams e metainformação encapsuladas num objecto Fedora, no contexto do RODA a definição de Objecto Digital adquire um caractér mais prático, estando associado a uma Representação no seu total, juntando assim as definições das duas partes.

Produtor

Actor (pessoa ou sistema) que interage com o repositório digital fornecendo-lhe informação a preservar.

Representação

Objecto Digital que instancia ou corporaliza uma Entidade Intelectual. Uma representação é o conjunto de ficheiros e metainformação estrutural necessária para uma apresentação completa e razoável da Entidade Intelectual.

Uma representação é uma "materialização"de uma Entidade Intelectual. Dentro do repositório e do SIP RODA cada tipo de representação tem um identificador, para que o repositório e o ingestor possam saber de que tipo é a representação para "chamar"os procedimentos adequados para lidar com uma determinada representação.

A Caso de estudo: AACC

No inicio do projecto RODA foi efectuado um caso de estudo com o restauro do fundo da *Alta Autoridade Contra a Corrupção* (AACC), guardado pela Torre do Tombo.

A documentação da AACC em formato digital sobre a qual se debruçou esta tarefa foi incorporada na Torre do Tombo (TT) em 1993 em 81 discos ópticos de dupla face. Em 1999 a empresa MEGADOC realizou um refrescamento da informação para CDs. Passaram a estar na TT os 81 discos ópticos originais e 161 CDs que são cópias exactas das 161 faces gravadas dos discos ópticos originais. Estes CDs contém imagens resultantes da digitalização de todas as páginas de documentação da AACC.

A ingestão (ou integração) do fundo foi feita manualmente, recorrendo a vários programas, desenvolvidos pelos signatários, que analisaram, validaram e transformaram os dados e metadados. O processo de migração decorreu de uma forma sequencial descrita em seguida, por secções.

A.1 Análise básica do fundo

O primeiro passo deste processo foi fazer uma cópia local (para os computadores pessoais) de uma amostra de 10 CDs. A cópia foi efectuada por um pequeno script que para além de copiar os conteúdos integrais dos CDs, guardava o resultado da cópia em ficheiros de texto para que houvesse um registo dos erros deste processo.

Seguidamente, foi feita uma análise da árvore dos sistemas de ficheiros dessa mesma amostra em conjunto com a documentação para separar dados de metadados e ainda tirar alguma informação da própria árvore. Depois desta análise verificou-se que o CD 159 continha digitalizações de desenhos e esquemas, o CD 160 continha os sistemas operativos das máquinas usadas para a digitalização e o CD 161 continha ficheiros variados que foram adicionados no fim do processo de digitalização. Nenhum destes CDs continha metainformação de algum tipo. Nos restantes 158 CDs foi encontrado um conjunto de directórios mais ou menos constante (em média 60) cujo nome é um número que segue sequencialmente pelas pastas. As pastas têm um máximo de 256 ficheiros por pasta. Cada ficheiro corresponde a uma imagem e o nome do ficheiro é um número que itera sequencialmente pelo CD e transversalmente ao directório a que pertence. Na raiz da árvore do sistema de ficheiros de cada CD existe um ficheiro de texto, com o nome 'sado' e a extensão relativa ao número do CD (na sequência dos 158 CDs), que contêm, em formato CSV ('comma separated values'), os metadados, que analisaremos mais tarde e ao qual referir-me-ei como *ficheiro sado*.

Tendo conhecimento que os CDs resultaram de um refrescamento (cópia directa) em 1999, a partir de um conjunto de discos ópticos gravados em 1993, resultantes de um processo de digitalização auxiliados por um sistema de indexação de imagens baseado em **LIS5000**, concluiu-se que os directórios criados deveram-se a uma limitação do sistema de ficheiros utilizado na altura, ou de um processo automático do sistema de indexação e que não têm qualquer significado. Pelo que cada CD pode ser visto como um conjunto plano de ficheiros, descritos pelo *ficheiro sado* respectivo.

A.2 Análise dos ficheiros de metadados

Os *ficheiros sado* encontrados na raiz de cada CD contêm toda a metainformação relativa à documentação guardada nos CDs da AACC.

Cada *ficheiro sado* é um ficheiro de texto em formato CSV, ou seja, valores separados por vírgulas. Depois de uma análise detalhada, por parte de arquivistas e da equipa de desenvolvimento, da metainformação em formato digital e em papel e de algumas imagens dos CDs, concluiu-se que o conjunto total do fundo está dividido em várias séries (Processos, Ofícios, Documentação Interna, etc.), cada uma delas com diferente número de valores por conjunto e diferentes significados para cada. No entanto, três valores dentro do conjunto tinham um significado comum em todas as séries, aparecendo sempre na parte final do conjunto, pela mesma ordem, estes são o número do disco, o número da primeira imagem (*imagem*) e o número de páginas (*paginas*). O número do disco corresponde ao número do CD (e consequentemente ao número da face do disco óptico original). O número da primeira imagem refere o nome do ficheiro, dentro do disco referido, que faz a primeira página do conjunto de ficheiros a que esta linha se refere. O número de páginas refere-se ao número de ficheiros cujo nome se sequencia numericamente ao da primeira página e que completam os ficheiros referenciados por esta linha. Assim cada linha aglomera um conjunto de ficheiros, relacionando os metadados descritos pelos outros valores do conjunto com os dados que descrevem, além de completar com metainformação estrutural (uma sequência implícita, limitada pelo intervalo $[imagem, imagem + paginas]$) o conjunto de ficheiros antes disperso.

Os arquivistas identificaram as várias entidades intelectuais e a sua relação com os metadados e consequentemente com dados. Na maior parte das séries identificadas, os *ficheiros sado* correspondentes tem um valor no conjunto (normalmente o primeiro), que tem o significado de código de referência que *'identifica uniquivocamente a unidade de descrição e providencia uma relação com a descrição que a representa'* [International Council on Archives, 1999]. O grupo de linhas (conjuntos de valores), dos *ficheiros sado* da mesma série,

que têm o mesmo código de referência, são referentes à mesma entidade intelectual. As séries que não têm um código de referência no conjunto de valores, são interpretadas como uma entidade intelectual por conjunto, segundo os arquivistas.

A.3 Validação

A.3.1 Validação dos metadados

Depois de ter a correcta interpretação da estrutura de metadados presente no fundo a ingerir (Secção A.2), procedemos à validação desta estrutura. Esta validação, que é impraticável manualmente, foi conseguida através de um script em Perl (escolhido pelas suas capacidades de manipulação de texto) que faz a interpretação (*'parsing'*) do *ficheiro sado* e o avalia segundo as seguintes regras:

- Todas as linhas tem o formato CSV, ou seja, têm valores separados por vírgulas e o número de valores é igual ao esperado, segundo a interpretação dos arquivistas
- Os dois últimos campos (*imagem* e *paginas*) são do tipo numérico e formam um intervalo $[imagem, imagem + paginas]$
- Os intervalos a que cada linha se refere não se sobrepõem
- Não existem ficheiros que não pertencem a nenhum intervalo

Com este script conseguimos detectar um grande número de erros nos *ficheiros sado*, que tivemos de corrigir, observando e interpretando as imagens da AACC e onde perdemos a maior parte do tempo gasto neste restauro. Numa ingestão oficial pelo repositório, estes erros não poderiam ser admitidos e o fundo teria de ser devolvido à procedência.

A.3.2 Validação dos ficheiros

Os ficheiros guardados necessitam de uma validação para garantir que não foram corrompidos e que continuam com a mesma informação que era suposto conterem. Já que não temos acesso a um *Message Digest* efectuado anteriormente, teremos de encontrar outras técnicas que avaliem a qualidade do ficheiro. Sabendo que o tipo esperado dos ficheiros é imagem, mais especificamente TIFF, tentamos então através de programas de inferência do tipo de ficheiro, validar o tipo de ficheiro inferido contra o esperado.

Tentamos primeiro o programa file (UNIX) que através de heurísticas tenta inferir o tipo de ficheiro. No entanto, sendo a validação por este método

insuficiente para os nossos propósitos, pois não garantia que o TIFF e os metadados embebidos estavam bem formados, estudamos um projecto de análise de ficheiros apropriado para os nossas intenções: o JHOVE [JSTOR et al., 2003].

O JHOVE, JSTOR/Harvard Object Validation Environment, é um projecto de uma colaboração entre a JSTOR (*Journal Storage, the scholarly journal archive*) e a Biblioteca Universitária de Harvard para desenvolver uma *framework* extensível para validações de formatos. JHOVE contém funcionalidades para identificar, validar e caracterizar objectos digitais. Nesta secção interessa-nos as funções de identificação e validação, mas mais tarde (Secção A.4.1) utilizaremos a função de caracterização.

Utilizando o JHOVE analisamos a cópia do fundo no disco rígido e descobrimos que 70 ficheiros estavam corrompidos (Anexo A.6). Para análise mais profunda deste facto, analisamos as cópias dos mesmos ficheiros nos CDs e até nos discos ópticos e descobrimos que estes ficheiros foram corrompidos antes mesmo de serem copiados para CDs (foram criados já com erro ou degradaram-se entre 1993 e 1999, embora esta última hipótese seja altamente improvável).

Quanto ao resto dos ficheiros, apenas encontramos um pequeno pormenor, que embora o JHOVE apontasse como erro, não foi considerado representativo para justificar uma mudança no original. Um dos critérios de validação que o JHOVE utiliza para analisar um ficheiro TIFF é se a *tag DateTime* tem o formato *"YYYY:MM:DD HH:MM:SS"*. No entanto as imagens apresentam a *tag DateTime* no formato *"YYYY:MM:DD***T***HH:MM:SS"*.

A.4 Migração

Após a validação completa entramos num processo de migração dos dados e metadados, para que estes sigam o esboço do repositório explicado na Secção 5.1.

A.4.1 Transformação e criação de metadados

Após a correcção e validação dos *ficheiros sado* obtivemos um conjunto de metainformação descritiva e estrutural juntos num formato CSV. Segundo o esboço do repositório, a metainformação descritiva pertence ao esquema EAD, enquanto que a estrutural pertence ao esquema METS e a parte do PREMIS (pois este tem que agregar o conjunto de ficheiros sobre uma representação). Após uma interpretação (*parsing*) para uma estrutura intermédia, em que os dados dos *ficheiros sado* foram agrupados segundo as considerações sobre entidades intelectuais supracitadas na Secção A.2, foram implementa-

dos vários programas que a partir desta estrutura intermédia criam os ficheiros EAD, PREMIS e METS, com as relações explicadas na Secção 5.1.

No entanto existe informação que é necessária no PREMIS e na sua extensão com NISO Z39.87 que não se encontram directamente disponíveis. Para este efeito foi utilizada a função de caracterização do JHOVE, que estende a metainformação inferida com o esquema NISO Z39.87 e providência todo o resto da informação necessária ao PREMIS. Com o resultado do JHOVE e utilizando transformações XSL, foi facilmente construído o resto do PREMIS necessário.

A.4.2 Esquema de identificadores

Por obrigações dos esquemas usados, foi necessário criar novos identificadores para os objectos digitais guardados. Se bem que houve muita discussão sobre identificadores persistentes, tomou-se a decisão que não se implementariam devido às infra-estruturas necessárias. Escolhemos, assim, identificadores criados para este caso específico apenas com o objectivo de servir de identificador único e não de localizador automático.

Figura 45: Estrutura de um identificador

O identificador criado é constituído por campos separados por pontos. Existe uma distinção do nível do tipo do objecto digital que o identificador se refere, R para *Representation*, F para *File*. Segue-se logo após uma distinção da versão do ficheiro ou representação, diferenciando os objectos originais (R1993) e os objectos derivados (pela migração da Secção A.4.3, no ano corrente, R2006). De seguida descreve-se na árvore da hierarquia EAD, distinguindo a série (Processos - 1, Despachos - 2.1, Ofícios - 2.2, etc.) e a entidade intelectual a que a respectiva representação se refere (exemplo para a representação relativa à entidade intelectual '2' da série Processos, versão 2006: R2006.01.0002). Para identificar os ficheiros constitutivos de uma representação, baseamos-nos no identificador da representação respectiva, alterando o identificador de R para F e adiciona-se uma distinção relativa à sua ordem na representação (exemplo para a terceira página: F2006.01.0002.0000003 (como esquematizado na figura 45). Para os ficheiros

METS que guardam a metainformação estrutural de uma representação não se usa o número de página (neste caso, o METS: F2006.01.0002).

A.4.3 Migração dos ficheiros

Todos os ficheiros guardados no fundo AACC são do tipo TIFF com compressão *Group 4*.

Como a compressão é um inibidor à renderização do objecto, tendo em conta que o algoritmo de compressão poderá tornar-se obsoleto mais rapidamente o próprio formato impedindo ou tornando mais difícil o acesso à representação, foi tomada a decisão de nunca preservar ficheiros comprimidos, migrando estes para o seu formato sem compressão.

Esta política adoptada pelo repositório obrigou os ficheiros do fundo AACC a serem migrados para o mesmo formato, mas sem compressão. Este evento fez com que o fundo ocupasse muito mais espaço que anteriormente, passando de 31.5 GB para perto de 800 GB. Foi utilizado o *tiffcp* da *libtiff* para executar esta descompressão.

A.5 Refrescamento

O último passo do processo foi o refrescamento de todo o fundo AACC para DVDs. Foi utilizado o esquema de metadados do repositório, que fica guardado em dois DVDs de metadados (AACC-Metadados-1 e AACC-Metadados-2). No DVD 1 existe um ficheiro para o EAD e um ficheiro PREMIS para cada representação de 2006. O segundo DVD de metadados contém os ficheiros PREMIS para as representações de 1993 das quais derivam as de 2006. Em cada DVD com dados, existe um directório por representação, em que o nome do directório corresponde ao identificador da própria representação (ver figura 46). Cada directório contém os ficheiros da representação com nomes iguais aos seus identificadores mais a extensão correspondente ao tipo do ficheiro (.tiff para as imagens e .METS.xml para o ficheiro de meta informação estrutural). Na raiz do DVD existe uma duplicação dos objectos PREMIS que existem no DVD de metadados, cujas representações a que se referem se encontram no DVD em questão.

No entanto existem representações que ocupam mais do que o espaço disponível num DVD (*Single-Layer*, 4483 MB de espaço). Para resolver este problema, cada representação foi dividida pelos DVDs que necessitava, duplicando de novo os objectos PREMIS (os objectos representação são duplicados, mas os objectos ficheiros só aparecem caso eles existam no DVD em questão). O ficheiro METS é duplicado em todos os DVDs que contenham a representação porque este referência todos os outros ficheiros da mesma.

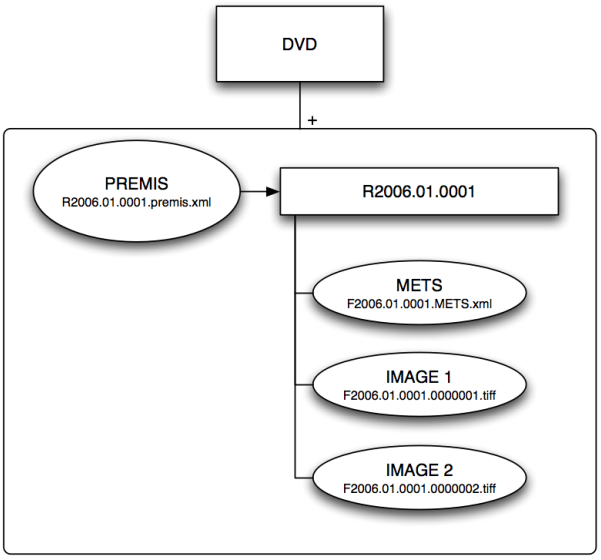


Figura 46: Esquema da estrutura do sistema de ficheiros dos DVDs de dados

O processo de refrescamento para DVDs foi colapsado com a migração dos ficheiros e com a criação de metainformação porque a situação ideal, em que cada um dos passos seria feito de uma vez só para a totalidade do fundo, ocuparia mais espaço (800Gb) que o que havia disponível nas máquinas de trabalho (cerca de 100Gb). Os processos de migração dos ficheiros, criação da metainformação e criação das imagens dos DVDs foram feitas para cada DVD individualmente e sempre que o espaço nos computadores de trabalho era esgotado os DVDs produzidos eram gravados e as imagens produzidas apagadas para dar lugar às seguintes.

Uma das dificuldades sentidas neste processo foi estimar o tamanho exacto dos ficheiros descomprimidos. As imagens do fundo eram suposto terem todas as mesmas dimensões (1640x239 pixeis) e, portanto, a estimativa para o tamanho das imagens descomprimidas era sempre igual para todas, mas na realidade havia algumas imagens que eram maiores e esse facto fez com que as estimativas do tamanho das imagens depois de descomprimidas fosse desajustado à realidade e a que os DVDs produzidos inicialmente fossem maiores que o permitido. Outro método de estimação teve que ser implementado e desta vez baseado na análise das dimensões de cada uma das imagens.

O resultado do refrescamento foram 196 DVDs *Single Layer* de dados (i.e. as imagens) e 2 DVDs *Dual Layer* com os metadados. Os 196 DVDs de dados estão nomeados como AACC-[número do DVD] (e.g. AACC-001 para o DVD 1) e os DVDs de metadados estão nomeados AACC-Metadados-1 e

AACC-Metadados-2, respectivamente para o DVD 1 e 2 de metadados.

A.6 Ficheiros Corrompidos

A tabela 3 identifica os ficheiros corrompidos detectados na Secção A.3.2.

Tabela 3: Ficheiros Corrompidos

CD	Ficheiro	Directório
30	4560	17
	5494	21
	7573	29
31	5020	19
	5214	20
32	4554	17
34	7839	30
50	5808	22
	5810	22
	5812	22
	5816	22
	5821	22
	5825	22
	5829	22
	5831	22
	5832	22
	5835	22
	5836	22
	5839	22
	5843	23
	5845	23
	5846	23
	5849	23
	5850	23
	7398	29
	7414	29
	7416	29
	7420	29
	7422	29
	7424	29
	7426	29
	7428	29
	7432	29
	7456	29
	7476	29

CD	Ficheiro	Directório
50	7492	29
	7500	29
	7502	29
	7504	29
	7508	29
	7510	29
	7512	29
	7514	29
	7516	29
	7518	29
	7520	29
	7522	29
	7524	29
	7526	29
52	4512	17
92	12113	47
95	5802	22
109	2940	11
	2942	11
119	65	0
	92	0
121	8865	34
133	8808	34
137	12173	47
141	10061	39
159	6	0
	288	1
	293	1
	337	1
	391	1
	442	1
	546	2
	1036	5
	1040	5
	1057	5

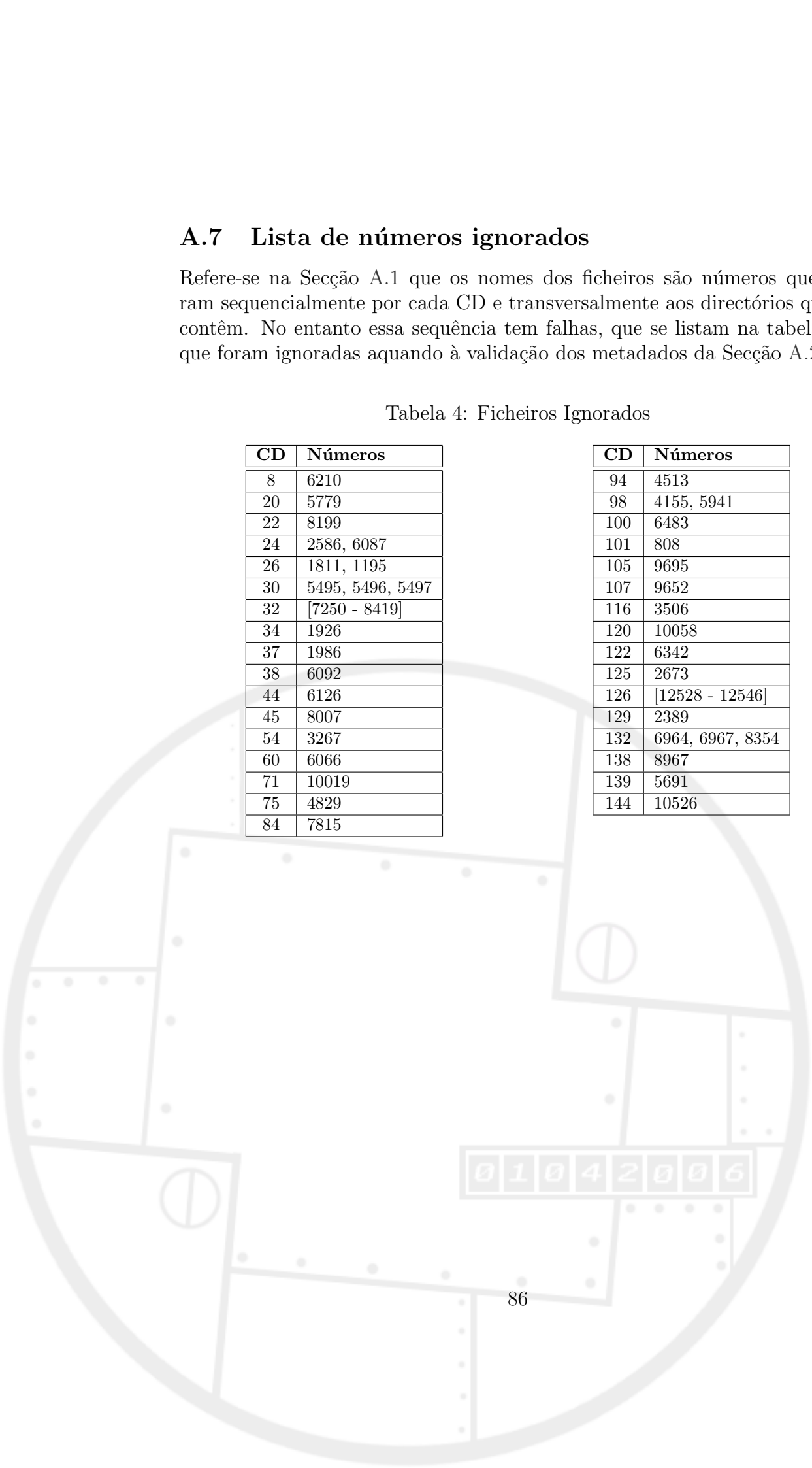
A.7 Lista de números ignorados

Refere-se na Secção A.1 que os nomes dos ficheiros são números que iteram sequencialmente por cada CD e transversalmente aos directórios que os contêm. No entanto essa sequência tem falhas, que se listam na tabela 4 e que foram ignoradas aquando à validação dos metadados da Secção A.2.

Tabela 4: Ficheiros Ignorados

CD	Números
8	6210
20	5779
22	8199
24	2586, 6087
26	1811, 1195
30	5495, 5496, 5497
32	[7250 - 8419]
34	1926
37	1986
38	6092
44	6126
45	8007
54	3267
60	6066
71	10019
75	4829
84	7815

CD	Números
94	4513
98	4155, 5941
100	6483
101	808
105	9695
107	9652
116	3506
120	10058
122	6342
125	2673
126	[12528 - 12546]
129	2389
132	6964, 6967, 8354
138	8967
139	5691
144	10526



B Análise de Requisitos

Os requisitos podem ser divididos os três meta-processos definidos no OAIS: Ingestão, Gestão e Disseminação. Nos três sub-capitulos seguintes são apresentadas as tabelas 5, 6 e 7 que listam os requisitos identificados para cada um dos meta-processos respectivos. Para cada requisito é identificado o componente ou funcionalidade que seria necessário para o satisfazer e determinado se o DSpace ou Fedora satisfaz (✔) ou não (✖) o mesmo.

B.1 Processo de Ingestão

Nº	Descrição	Componente	D	F
1.1	O RODA tem que desenvolver uma interface intuitiva que suporte as transacções previstas durante o processo de ingestão.	Interface Gráfica para processo de ingestão.	✓	✖
1.2	O RODA tem de ter a capacidade de registar informação administrativa sobre o cliente.	Registo de Produtores	✓	✖
1.3	O RODA deve ter a capacidade tecnológica de produzir SIP e de fornecer ferramentas para o cliente produzir SIP de acordo com procedimentos normalizados do Repositório Digital.	Ferramenta de auxílio à produção de SIPs	✖	✖
1.4	O RODA deve produzir documentos notificativos e disponibilizar essa informação através da interface, do resultado da avaliação preliminar (sub-processo 1.1).	Feedback sobre a viabilidade do processo de ingestão	✓	✖
1.5	O RODA deve ter a capacidade de integrar apenas parte dos SIP propostos para ingestão, devendo identificar a coerência tecnológica e intelectual desses SIP.	O Repositório deve permitir que parte de um conjunto de SIPs possa não ser incorporado (rejeitado) mas que essa informação seja guardada.	✓	✓
1.6	O RODA deve dispor de espaços físicos alocados na plataforma tecnológica para colocar SIP em fase de pré-integração. Estes espaços devem prever a colocação de SIP provenientes de diferentes clientes possibilitando a realização simultânea de vários processos de ingestão.	Entenda-se "espaços físicos"por "espaços lógicos". Um local de quarentena para que possam ser validados antes de concluir a ingestão.	✓	✖
1.7	O RODA deve ter a capacidade de analisar os objectos digitais dentro de diversos parâmetros criando uma ou mais rotinas de análise aplicáveis a qualquer SIP (por ex. a despistagem de vírus).	Validação dos SIPs.	✖	✖

Nº	Descrição	Componente	D	F
1.8	O RODA deve assegurar a capacidade de evitar a eliminação involuntária dos ficheiros admitidos à pré-integração.	Undelete, Trash.	✖	✖
1.9	Deve ser realizado sobre estes ficheiros um processo de autenticação sumário para garantir o controlo de eventuais corrupções involuntárias.	Validação dos SIPs / Criação de checksums.	✔	✖
1.10	O RODA deve ter a capacidade de emitir notificações para o cliente de forma automática, descrevendo os erros detectados.	Interface de notificação de erros no processo de ingestão.	✔	✖
1.11	O RODA tem de assegurar a conversão dos SIP de acordo com a política e estratégia de preservação digital preconizada.	Normalização dos conteúdos dos SIPs / Conversão para formatos normalizados.	✖	✖
1.12	O RODA tem de assegurar a atribuição controlada de MI aos SIP e AIP resultantes.	Gestão controlada de metainformação.	✔	✖
1.13	O RODA tem de ter a capacidade de produzir e manter identificadores únicos persistentes baseados em especificações internacionais.	PIDs standard.	✔[20]	✖
1.14	O RODA tem de ter documentação (MI) sobre os conversores/transformadores utilizados no processo de criação de AIP.	Metainformação sobre Agentes (Software).	✖	✖
1.15	O RODA deve produzir um relatório com dados sobre o subprocesso de integração a ser apresentado ao Cliente e à Administração.	Produção de relatórios sobre o processo de ingestão.	✖	✖

Tabela 5: Requisitos do processo de Ingestão

[20] DSpace usa o Handle System

Nº	Descrição	Componente	D	F
2.1	O RODA deve possuir uma interface que permita a interacção entre o gestor do RODA e os AIP utilizando todas as ferramentas necessárias para a realização das operações de gestão previstas.	Ferramenta de Administração.	✔	✔[21]
2.2	O RODA deve atribuir a cada EVENTO um fluxo de acções pré-determinadas e desenvolvidas de acordo com um fluxo sequencial e/ou concorrente que serão herdadas pelas diversas instâncias desse evento.	Definição de *workflows* para tarefas.	✖	✔[22]
2.3	O RODA deve ter a capacidade de gerar de forma automática MI para cada evento despoletado.	Gerar eventos PREMIS (event).	✖[23]	✖[24]
2.4	Deve ser assegurada a capacidade de atribuir MI descritiva a nível agregado (classes de AIP).	Fazer descrição de conjuntos de Representações (AIPs).	✔	✖
2.5	O RODA deverá desenvolver eventos de tipo rotina de auditoria (verificação e prevenção) para a gestão da infraestrutura tecnológica de suporte.	Monitorização do Sistema (Prevenção de falhas) e dos AIP.	✖[25]	✖
2.6	O RODA deve emitir automaticamente alertas (avisos) relativos a um conjunto de eventos que devem ser despoletados. Por exemplo: data de actualização de AIP (migração); data de refrescamento prevista, etc.	Notificações periódicas.	✖	✖

[21] Fedora tem uma ferramenta de administração mas não é muito amigável

[22] Fedora pode criar *behaviours* e *mechanisms* para a definição de tarefas. *Workflows* só podem ser implementados de maneira implícita dentro dos mesmos *behaviours* e *mechanisms*.

[23] DSpace tem um protótipo de sistema de eventos sobre o qual pode assentar este componente

[24] Em desenvolvimento - existe uma especificação para um sistema de eventos baseado em eventos PREMIS para o Fedora

[25] Só tem validação de sumário MD5

Nº	Descrição	Componente	D	F
2.7	O RODA deve assegurar que na sequência de um evento será adicionada MI aos objectos (AIP; Infra-estrutura técnica) alvo desse evento.	Actualização da MI PREMIS relativa a eventos.	✖[26]	✖[27]
2.8	Após um evento de actualização (migração) os AIP resultantes deverão ser verificados quanto à sua integridade e inteligibilidade sendo para isso sujeitos a processo de validação.	Efectuar uma validação no final de qualquer acção que altere os AIPs (*Workflow* supracitado).	✖	✖

Tabela 6: Requisitos do processo de Gestão de AIP

[26]Mas faz parte do sistema de eventos em desenvolvimento

[27]Mas tem implementação prevista

Nº	Descrição	Componente	D	F
3.1	O RODA tem de manter uma interface amigável para gerir o processo de disseminação e interactuar com o cliente.	Interface Gráfica para Disseminação.	✔[28]	✔[29]
3.2	O RODA tem de ter a capacidade de assegurar que são recuperadas apenas as componentes de um AIP e respectiva MI necessárias para satisfazer o pedido do utilizador, sem acrescentar ou diminuir informação.	Permitir transformações de formatos e MI entre os AIP e os DIP.	✖	✔
3.3	O RODA tem que permitir que a informação disponibilizada ao utilizador seja essencialmente descritiva e que o ponto de referência para recuperar o AIP e produzir o DIP seja o identificador. Para o gestor a informação obtida tem de permitir identificar, localizar e recuperar todas as componentes que eventualmente constituam o AIP.	O utilizador(consumidor) pesquisa apenas no EAD e usa o identificador único para recuperar o objecto.	✔	✔
3.4	A certificação do DIP tem de respeitar os métodos legalmente reconhecidos em Portugal (Assinatura digital).	Assinaturas digitais nos DIPs.	✖	✖

Tabela 7: Requisitos do processo de Disseminação

[28]Tem procura e navegador mas não tem disseminadores

[29]Mas não é amigável

C METS de um SIP **RODA**

```
<?xml version="1.0" encoding="UTF-8"?>
<mets xmlns="http://www.loc.gov/METS/"
      xmlns:xlink="http://www.w3.org/1999/xlink"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.loc.gov/METS/
                          http://www.loc.gov/standards/mets/mets.xsd"
      PROFILE="RODA_SIP" OBJID="PT/TT/AACC/1/1">

    <metsHdr CREATEDATE="2006-11-17T19:00:00">
        <agent ROLE="CREATOR">
            <name>Rui Castro</name>
        </agent>
    </metsHdr>

    <dmdSec ID="PT-TT-AACC-1-1.EAD" ADMID="R2006.01.0001.premis.AMDSEC">
        <mdRef LOCTYPE="URL" MDTYPE="OTHER" xlink:href="PT-TT-AACC-1-1.ead"
               LABEL="roda:d:dc" />
    </dmdSec>

    <amdSec ID="R2006.01.0001.premis.AMDSEC">
        <techMD ID="PT-TT-AACC-1-1.PREMIS">
            <mdRef LOCTYPE="URL" MDTYPE="PREMIS"
                xlink:href="R2006.01.0001.premis" />
        </techMD>
    </amdSec>

    <fileSec>

        <fileGrp>
            <file ID="F2006.01.0001" MIMETYPE="text/xml" CHECKSUMTYPE="MD5"
              CHECKSUM="d20627bc137dbcb616af020ec7d45ded">
                <FLocat LOCTYPE="URL"
                    xlink:href="R2006.01.0001/F2006.01.0001.METS.xml" />
            </file>
            <file ID="F2006.01.0001.0000001" MIMETYPE="image/tiff"
            CHECKSUMTYPE="MD5"
            CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                 xlink:href="R2006.01.0001/F2006.01.0001.0000001.tiff" />
            </file>
            <file ID="F2006.01.0001.0000002" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                    xlink:href="R2006.01.0001/F2006.01.0001.0000002.tiff" />
            </file>
            <file ID="F2006.01.0001.0000003" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                    xlink:href="R2006.01.0001/F2006.01.0001.0000003.tiff" />
            </file>
            <file ID="F2006.01.0001.0000004" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                    xlink:href="R2006.01.0001/F2006.01.0001.0000004.tiff" />
            </file>
            <file ID="F2006.01.0001.0000005" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                    xlink:href="R2006.01.0001/F2006.01.0001.0000005.tiff" />
            </file>
            <file ID="F2006.01.0001.0000006" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                    xlink:href="R2006.01.0001/F2006.01.0001.0000006.tiff" />
            </file>
            <file ID="F2006.01.0001.0000007" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                    xlink:href="R2006.01.0001/F2006.01.0001.0000007.tiff" />
            </file>
            <file ID="F2006.01.0001.0000008" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
```

```
                xlink:href="R2006.01.0001/F2006.01.0001.0000008.tiff"/>
            </file>
            <file ID="F2006.01.0001.0000009" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                xlink:href="R2006.01.0001/F2006.01.0001.0000009.tiff"/>
            </file>
            <file ID="F2006.01.0001.0000010" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                xlink:href="R2006.01.0001/F2006.01.0001.0000010.tiff"/>
            </file>
            <file ID="F2006.01.0001.0000011" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                xlink:href="R2006.01.0001/F2006.01.0001.0000011.tiff"/>
            </file>
            <file ID="F2006.01.0001.0000012" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                xlink:href="R2006.01.0001/F2006.01.0001.0000012.tiff"/>
            </file>
            <file ID="F2006.01.0001.0000013" MIMETYPE="image/tiff"
              CHECKSUMTYPE="MD5"
              CHECKSUM="5d76e70c855f3d6be4e845ab0f408f0c">
                <FLocat LOCTYPE="URL"
                xlink:href="R2006.01.0001/F2006.01.0001.0000013.tiff"/>
            </file>
        </fileGrp>
    </fileSec>

    <structMap>
        <div ID="R2006.01.0001" DMDID="PT-TT-AACC-1-1.EAD"
          ADMID="R2006.01.0001.premis.AMDSEC" TYPE="digitalized_work">
            <fptr FILEID="F2006.01.0001"/>
            <fptr FILEID="F2006.01.0001.0000001"/>
            <fptr FILEID="F2006.01.0001.0000002"/>
            <fptr FILEID="F2006.01.0001.0000003"/>
            <fptr FILEID="F2006.01.0001.0000004"/>
            <fptr FILEID="F2006.01.0001.0000005"/>
            <fptr FILEID="F2006.01.0001.0000006"/>
            <fptr FILEID="F2006.01.0001.0000007"/>
            <fptr FILEID="F2006.01.0001.0000008"/>
            <fptr FILEID="F2006.01.0001.0000009"/>
            <fptr FILEID="F2006.01.0001.0000010"/>
            <fptr FILEID="F2006.01.0001.0000011"/>
            <fptr FILEID="F2006.01.0001.0000012"/>
            <fptr FILEID="F2006.01.0001.0000013"/>
        </div>
    </structMap>

</mets>
```

D METS de um SIP Diringest

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<METS:mets xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
           xmlns:METS="http://www.loc.gov/METS/"
           xmlns:xlink="http://www.w3.org/1999/xlink"
           xmlns:premis="http://www.loc.gov/standards/premis/v1"
           xmlns:eadpart="http://roda.iantt.pt/eadpart/BetaSchema20070112">
  <METS:fileSec>
    <METS:fileGrp>
      <METS:file MIMETYPE="text/xml" ID="d1e14">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://PT-TT-AACC-1-1.ead"/>
      </METS:file>
      <METS:file ID="R2006.01.0001.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://R2006.01.0001.PREMIS"/>
      </METS:file>
      <METS:file ID="E4.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL" xlink:href="file://E4.PREMIS"/>
      </METS:file>
      <METS:file ID="A6.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL" xlink:href="file://A6.PREMIS"/>
      </METS:file>
      <METS:file ID="E5.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL" xlink:href="file://E5.PREMIS"/>
      </METS:file>
      <METS:file ID="A7.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL" xlink:href="file://A7.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000001.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000001.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000002.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000002.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000003.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000003.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000004.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000004.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000005.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000005.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000006.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000006.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000007.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000007.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000008.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000008.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000009.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000009.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000010.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000010.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000011.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000011.PREMIS"/>
      </METS:file>
      <METS:file ID="F2006.01.0001.0000012.PREMIS" MIMETYPE="text/xml">
        <METS:FLocat LOCTYPE="URL"
       xlink:href="file://F2006.01.0001.0000012.PREMIS"/>
      </METS:file>
```

```
        <METS:file ID="F2006.01.0001.0000013.PREMIS" MIMETYPE="text/xml">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://F2006.01.0001.0000013.PREMIS"/>
        </METS:file>
        <METS:file ID="F2006.01.0001" MIMETYPE="text/xml">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.METS.xml"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000001" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000001.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000002" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000002.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000003" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000003.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000004" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000004.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000005" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000005.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000006" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000006.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000007" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000007.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000008" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000008.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000009" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000009.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000010" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000010.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000011" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000011.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000012" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000012.tiff"/>
        </METS:file>
        <METS:file ID="F2006.01.0001.0000013" MIMETYPE="image/tiff">
            <METS:FLocat LOCTYPE="URL"
        xlink:href="file://R2006.01.0001/F2006.01.0001.0000013.tiff"/>
        </METS:file>
    </METS:fileGrp>
</METS:fileSec>
<METS:structMap>
    <METS:div LABEL="1" TYPE="roda:d:dc">
        <METS:div LABEL="Encoded Archival Description Part"
      TYPE="roda:f:EAD">
            <METS:fptr FILEID="d1e14"/>
        </METS:div>
        <METS:div TYPE="roda:p" LABEL="PT-TT-AACC-1-1.PREMIS">
            <METS:div LABEL="Object R2006.01.0001" TYPE="roda:f:PREMIS">
                <METS:fptr FILEID="R2006.01.0001.PREMIS"/>
            </METS:div>
            <METS:div LABEL="Event E4" TYPE="roda:f:PREMIS">
                <METS:fptr FILEID="E4.PREMIS"/>
            </METS:div>
            <METS:div LABEL="Agent A6" TYPE="roda:f:PREMIS">
                <METS:fptr FILEID="A6.PREMIS"/>
            </METS:div>
            <METS:div LABEL="Event E5" TYPE="roda:f:PREMIS">
                <METS:fptr FILEID="E5.PREMIS"/>
            </METS:div>
            <METS:div LABEL="Agent A7" TYPE="roda:f:PREMIS">
                <METS:fptr FILEID="A7.PREMIS"/>
```

```
</METS:div>
<METS:div LABEL="Object F2006.01.0001" TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000001"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000001.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000002"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000002.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000003"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000003.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000004"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000004.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000005"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000005.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000006"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000006.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000007"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000007.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000008"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000008.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000009"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000009.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000010"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000010.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000011"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000011.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000012"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000012.PREMIS"/>
</METS:div>
<METS:div LABEL="Object F2006.01.0001.0000013"
TYPE="roda:f:PREMIS">
    <METS:fptr FILEID="F2006.01.0001.0000013.PREMIS"/>
</METS:div>
<METS:div LABEL="R2006.01.0001" TYPE="roda:r:digitalized_work">
    <METS:div TYPE="roda:f" LABEL="F2006.01.0001">
        <METS:fptr FILEID="F2006.01.0001"/>
    </METS:div>
    <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000001">
        <METS:fptr FILEID="F2006.01.0001.0000001"/>
    </METS:div>
    <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000002">
        <METS:fptr FILEID="F2006.01.0001.0000002"/>
    </METS:div>
    <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000003">
        <METS:fptr FILEID="F2006.01.0001.0000003"/>
    </METS:div>
    <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000004">
        <METS:fptr FILEID="F2006.01.0001.0000004"/>
    </METS:div>
    <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000005">
        <METS:fptr FILEID="F2006.01.0001.0000005"/>
    </METS:div>
    <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000006">
        <METS:fptr FILEID="F2006.01.0001.0000006"/>
    </METS:div>
    <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000007">
        <METS:fptr FILEID="F2006.01.0001.0000007"/>
    </METS:div>
    <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000008">
        <METS:fptr FILEID="F2006.01.0001.0000008"/>
```

```
                </METS:div>
                <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000009">
                    <METS:fptr FILEID="F2006.01.0001.0000009"/>
                </METS:div>
                <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000010">
                    <METS:fptr FILEID="F2006.01.0001.0000010"/>
                </METS:div>
                <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000011">
                    <METS:fptr FILEID="F2006.01.0001.0000011"/>
                </METS:div>
                <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000012">
                    <METS:fptr FILEID="F2006.01.0001.0000012"/>
                </METS:div>
                <METS:div TYPE="roda:f" LABEL="F2006.01.0001.0000013">
                    <METS:fptr FILEID="F2006.01.0001.0000013"/>
                </METS:div>
            </METS:div>
        </METS:div>
    </METS:div>
  </METS:structMap>
</METS:mets>
```

E METS estrutural de uma representação

```
<?xml version="1.0" encoding="UTF-8"?>
<mets xmlns="http://www.loc.gov/METS/"
   xmlns:xlink="http://www.w3.org/1999/xlink"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.loc.gov/METS/
                       http://www.loc.gov/standards/mets/mets.xsd"
   OBJID="F2006.01.0002" LABEL="Processo XXX" TYPE="Processo de Averiguação">
   <fileSec>
     <fileGrp>
       <file ID="F2006.01.0002.0000001" MIMETYPE="image/tiff">
         <FLocat LOCTYPE="URL" xlink:href="F2006.01.0002.0000001.tiff"/>
       </file>
       <file ID="F2006.01.0002.0000002" MIMETYPE="image/tiff">
         <FLocat LOCTYPE="URL" xlink:href="F2006.01.0002.0000002.tiff"/>
       </file>
       <file ID="F2006.01.0002.0000003" MIMETYPE="image/tiff">
         <FLocat LOCTYPE="URL" xlink:href="F2006.01.0002.0000003.tiff"/>
       </file>
       <file ID="F2006.01.0002.0000004" MIMETYPE="image/tiff">
         <FLocat LOCTYPE="URL" xlink:href="F2006.01.0002.0000004.tiff"/>
       </file>
       <file ID="F2006.01.0002.0000005" MIMETYPE="image/tiff">
         <FLocat LOCTYPE="URL" xlink:href="F2006.01.0002.0000005.tiff"/>
       </file>
       <file ID="F2006.01.0002.0000006" MIMETYPE="image/tiff">
         <FLocat LOCTYPE="URL" xlink:href="F2006.01.0002.0000006.tiff"/>
       </file>
     </fileGrp>
   </fileSec>
   <structMap>
     <div>
       <div ORDER="1" LABEL="Sumário">
         <fptr>
           <seq>
             <area FILEID="F2006.01.0002.0000001"/>
           </seq>
         </fptr>
       </div>
       <div ORDER="2" LABEL="Processo">
         <fptr>
           <seq>
             <area FILEID="F2006.01.0002.0000002"/>
             <area FILEID="F2006.01.0002.0000003"/>
             <area FILEID="F2006.01.0002.0000004"/>
             <area FILEID="F2006.01.0002.0000005"/>
             <area FILEID="F2006.01.0002.0000006"/>
           </seq>
         </fptr>
       </div>
     </div>
   </structMap>
</mets>
```

F Exemplo de um ficheiro EAD

```
<?xml version="1.0" encoding="UTF-8"?>
<ead xmlns="urn:isbn:1-931666-22-9"
     xmlns:xlink="http://www.w3.org/1999/xlink"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="urn:isbn:1-931666-22-9 http://www.loc.gov/ead/ead.xsd
     http://www.w3.org/1999/xlink http://www.loc.gov/ead/xlink.xsd">
  <eadheader>
    <eadid/>
    <filedesc>
      <titlestmt>
        <titleproper/>
      </titlestmt>
    </filedesc>
  </eadheader>
  <archdesc level="otherlevel" otherlevel="F">
    <did>
      <abstract/>
      <unitid countrycode="PT"
        repositorycode="PT-adporto">ALL/CMTROFA2</unitid>
      <physdesc>
        <extent unit="livro">0</extent>
        <extent unit="capilha">0</extent>
        <extent unit="capa">0</extent>
        <extent unit="pasta">0</extent>
        <extent unit="macete">0</extent>
        <extent unit="maco">0</extent>
        <extent unit="ml">0</extent>
        <extent unit="rolo">0</extent>
        <extent unit="outro">0</extent>
        <extent unit="pagina">0</extent>
        <extent unit="folha">0</extent>
      </physdesc>
      <langmaterial>Português</langmaterial>
      <repository>Arquivo Distrital do Porto</repository>
      <unittitle>Câmara Municipal de Trofa 2</unittitle>
    </did>
    <processinfo>
      <p>
        <date normal="2004-06-14"/>
        <name>admin</name>
      </p>
    </processinfo>
    <dsc>
      <c level="otherlevel" otherlevel="SC">
        <did>
          <unitid countrycode="PT"
          repositorycode="PT-adporto">p00001</unitid>
          <physdesc>
            <extent unit="livro">0</extent>
            <extent unit="capilha">0</extent>
            <extent unit="capa">0</extent>
            <extent unit="pasta">0</extent>
            <extent unit="macete">0</extent>
            <extent unit="maco">0</extent>
            <extent unit="ml">0</extent>
            <extent unit="rolo">0</extent>
            <extent unit="outro">0</extent>
            <extent unit="pagina">0</extent>
            <extent unit="folha">0</extent>
          </physdesc>
          <langmaterial>Português</langmaterial>
          <repository>Arquivo Distrital do Porto</repository>
          <unittitle type="original">Sem título</unittitle>
        </did>
        <processinfo>
          <p>
            <date normal="2004-06-14"/>
            <name>admin</name>
          </p>
        </processinfo>
      </c>
    </dsc>
  </archdesc>
</ead>
```

G Exemplo de um ficheiro PREMIS contendo metainformação técnica NISO Z39.87

```
<?xml version="1.0" encoding="UTF-8"?>
<premis xmlns="http://www.loc.gov/standards/premis/v1"
   xmlns:mix="http://www.loc.gov/mix/"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.loc.gov/standards/premis/v1
         http://www.loc.gov/standards/premis/v1/PREMIS-v1-1.xsd
         http://www.loc.gov/mix/ http://www.loc.gov/mix/mix.xsd">
  <object>
    <objectIdentifier>
      <objectIdentifierType>Custom</objectIdentifierType>
      <objectIdentifierValue>R2006.01.0001</objectIdentifierValue>
    </objectIdentifier>
    <preservationLevel>full</preservationLevel>
    <objectCategory>representation</objectCategory>
    <relationship>
      <relationshipType>structural</relationshipType>
      <relationshipSubType>has root</relationshipSubType>
      <relatedObjectIdentification>
        <relatedObjectIdentifierType>Custom</relatedObjectIdentifierType>
        <relatedObjectIdentifierValue>F2006.01.0001</relatedObjectIdentifierValue>
        <relatedObjectSequence>0</relatedObjectSequence>
      </relatedObjectIdentification>
    </relationship>
    <relationship>
      <relationshipType>structural</relationshipType>
      <relationshipSubType>has part</relationshipSubType>
      <relatedObjectIdentification>
       <relatedObjectIdentifierType>Custom</relatedObjectIdentifierType>
       <relatedObjectIdentifierValue>F2006.01.0001.0000001</relatedObjectIdentifierValue>
       <relatedObjectSequence>1</relatedObjectSequence>
      </relatedObjectIdentification>
    </relationship>
    <linkingIntellectualEntityIdentifier>
      <linkingIntellectualEntityIdentifierType>Custom
</linkingIntellectualEntityIdentifierType>
      <linkingIntellectualEntityIdentifierValue>1
</linkingIntellectualEntityIdentifierValue>
    </linkingIntellectualEntityIdentifier>
  </object>
  <object>
    <objectIdentifier>
      <objectIdentifierType>Custom</objectIdentifierType>
      <objectIdentifierValue>F2006.01.0001</objectIdentifierValue>
    </objectIdentifier>
    <preservationLevel>full</preservationLevel>
    <objectCategory>File</objectCategory>
    <objectCharacteristics>
      <compositionLevel>1</compositionLevel>
      <size>3156</size>
      <format>
        <formatDesignation>
          <formatName>image/xml</formatName>
        </formatDesignation>
        <formatRegistry>
          <formatRegistryName>JHOVE</formatRegistryName>
          <formatRegistryKey>XML 1.0</formatRegistryKey>
        </formatRegistry>
        <formatRegistry>
          <formatRegistryName>MIME</formatRegistryName>
          <formatRegistryKey>text/xml</formatRegistryKey>
        </formatRegistry>
      </format>
    </objectCharacteristics>
    <creatingApplication>
      <creatingApplicationName>MakeDVDImage.pl</creatingApplicationName>
      <dateCreatedByApplication>2006-06-09T15:34:18Z</dateCreatedByApplication>
    </creatingApplication>
    <storage>
      <contentLocation>
        <contentLocationType>URI</contentLocationType>
        <contentLocationValue>R2006.01.0001/F2006.01.0001.mets</contentLocationValue>
      </contentLocation>
      <storageMedium>DVD</storageMedium>
    </storage>
  </object>

  <object>
```

```
    <objectIdentifier>
      <objectIdentifierType>Custom</objectIdentifierType>
      <objectIdentifierValue>F2006.01.0001.0000001</objectIdentifierValue>
    </objectIdentifier>
    <preservationLevel>full</preservationLevel>
    <objectCategory>File</objectCategory>
    <objectCharacteristics>
      <compositionLevel>0</compositionLevel>
      <fixity>
        <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
        <messageDigest>5d5679af155d9752fab9a42e6c724f0a</messageDigest>
      </fixity>
      <size>482190</size>
      <format>
        <formatDesignation>
          <formatName>image/tiff</formatName>
        </formatDesignation>
        <formatRegistry>
          <formatRegistryName>JHOVE</formatRegistryName>
          <formatRegistryKey>TIFF 5.0</formatRegistryKey>
        </formatRegistry>
        <formatRegistry>
          <formatRegistryName>MIME</formatRegistryName>
          <formatRegistryKey>image/tiff</formatRegistryKey>
        </formatRegistry>
      </format>
      <significantProperties>
        <mix:mix>
          <mix:BasicImageParameters>
            <mix:Format>
              <mix:MIMEType>image/tiff</mix:MIMEType>
              <mix:ByteOrder>little-endian</mix:ByteOrder>
              <mix:Compression>
                <mix:CompressionScheme>1</mix:CompressionScheme>
              </mix:Compression>
              <mix:PhotometricInterpretation>
                <mix:ColorSpace>0</mix:ColorSpace>
              </mix:PhotometricInterpretation>
              <mix:Segments>
                <mix:StripOffsets>8</mix:StripOffsets>
                <mix:RowsPerStrip>2339</mix:RowsPerStrip>
                <mix:StripByteCounts>481834</mix:StripByteCounts>
              </mix:Segments>
              <mix:PlanarConfiguration>1</mix:PlanarConfiguration>
            </mix:Format>
            <mix:File>
              <mix:Orientation>1</mix:Orientation>
            </mix:File>
          </mix:BasicImageParameters>
          <mix:ImageCreation>
          <mix:ScanningSystemCapture>
            <mix:ScanningSystemSoftware>
              <mix:ScanningSoftware>Kofax Image Processing Platform v1.10.018;
KIPP TIFF Storage Filter v1.10.018</mix:ScanningSoftware>
            </mix:ScanningSystemSoftware>
          </mix:ScanningSystemCapture>
            <mix:DateTimeCreated>1992-05-09T15:10:57</mix:DateTimeCreated>
          </mix:ImageCreation>
          <mix:ImagingPerformanceAssessment>
            <mix:SpatialMetrics>
              <mix:SamplingFrequencyUnit>2</mix:SamplingFrequencyUnit>
              <mix:XSamplingFrequency>200</mix:XSamplingFrequency>
              <mix:YSamplingFrequency>200</mix:YSamplingFrequency>
              <mix:ImageWidth>1648</mix:ImageWidth>
              <mix:ImageLength>2339</mix:ImageLength>
            </mix:SpatialMetrics>
            <mix:Energetics>
              <mix:BitsPerSample>1</mix:BitsPerSample>
              <mix:SamplesPerPixel>1</mix:SamplesPerPixel>
            </mix:Energetics>
          </mix:ImagingPerformanceAssessment>
        </mix:mix>
      </significantProperties>
    </objectCharacteristics>
    <creatingApplication>
      <creatingApplicationName>tiffcp (libtiff-tools)</creatingApplicationName>
      <creatingApplicationVersion>3.7.3-1ubuntu1.1</creatingApplicationVersion>
      <dateCreatedByApplication>2006-06-09T15:34:15Z</dateCreatedByApplication>
    </creatingApplication>
    <originalName>3</originalName>
    <storage>
      <contentLocation>
        <contentLocationType>URI</contentLocationType>
```

```
        <contentLocationValue>R2006.01.0001/F2006.01.0001.0000001.tiff
</contentLocationValue>
      </contentLocation>
      <storageMedium>DVD</storageMedium>
    </storage>
  </object>

</premis>
```