RODA: Repositório de Objectos Digitais Autênticos

Francisco Barbedo, Luís Corujo, Rui Castro, Luís Faria
Instituto dos Arquivos Nacionais
1649-010 Lisboa, Portugal
Tel: +351 217811500
E-mail: {frbarbedo, lcorujo, rcastro, lfaria}@iantt.pt

José Carlos Ramalho
Departamento de Informática da Universidade do Minho,
4710-057 Braga, Portugal
Tel: +351 253604479
E-mail: jcr@di.uminho.pt

Miguel Ferreira
Departamento de Sistemas de Informação da Universidade do Minho,
4800 Guimarães, Portugal
Tel: +351 253510261
E-mail: mferreira@dsi.uminho.pt

RESUMO

Um Arquivo Digital é uma estrutura que compreende tecnologia, recursos humanos e um conjunto de políticas para incorporar, gerir e acessibilizar, numa perspectiva continuada, objectos digitais de natureza arquivística. A informação de arquivo distingue-se de qualquer outra pelo facto de ser produzida com o propósito primário de constituir prova de uma actividade organizacional. Por esse facto a sua estabilidade e perenidade têm que ser asseguradas de forma a garantir as suas propriedades básicas ao longo do tempo: integridade, fiabilidade e autenticidade. Tudo isto é complicado pelo facto de o objecto digital ser extremamente volátil. Dependente de um sistema intermediário (software e hardware) integrado numa indústria altamente competitiva e evolutiva, observam-se prazos de retrocompatibilidade assegurados pelas empresas desenvolvedoras da ordem dos 5 anos. Isto significa que 5 anos é sensivelmente é o período de "auto-preservação" dos objectos digitais.

Neste contexto a prática de preservação digital deverá entrar nos planos de actividades e preocupações das instituições. O problema reside em como guardar de forma operacionalmente útil os objectos digitais que irão ser necessários às actividades da Organização durante períodos de tempo muito superiores ao prazo de "auto-preservação?

O RODA (Repositório de Objectos Digitais Autênticos) é um projecto lançado pelo Instituto dos Arquivos Nacionais/Torre do Tombo (Brevemente Direcção Geral de Arquivos – DGARQ) que conta com a colaboração da Universidade do Minho e que pretende abordar de forma sistemática estas questões no intuito de vir a colmatar um vazio actualmente existente relativamente à gestão continuada de objectos digitais.

PALAVRAS-CHAVE: Preservação digital, autenticidade, arquivo digital, objectos digitais, governo electrónico

INTRODUÇÃO

O Instituto de Arquivos Nacionais/Torre do Tombo (IAN/TT) assume na sua missão institucional a responsabilidade pela identificação e preservação de documentação de valor histórico como meio de garantir e fomentar a memória individual e colectiva nacional. Em paralelo, as iniciativas do Governo Electrónico determinam que a Administração Pública (AP) deverá, cada vez mais, basear a sua actividade em processos de negócio electrónicos com o intuito de agilizar e assegurar um serviço mais rápido, completo e transparente para o cidadão. Neste cenário torna-se evidente que assistiremos a um aumento da produção de informação digital, informação esta que, de acordo com a missão do IAN/TT, deverá ver assegurado o seu valor evidencial através da garantia da sua autenticidade.

Acontece, no entanto, que o IAN/TT ainda não dispõe de estruturas capazes de suportar a incorporação e gestão de informação de arquivo produzida em formatos electrónicos. Neste sentido, o IAN/TT deve empenhar-se de modo a desenvolver processos, ferramentas e recursos capazes de dar resposta às necessidades de preservação da informação digital produzida na Administração Pública cuja conservação continuada seja considerada como justificada.

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.

Neste artigo, apresentam-se e discutem-se as decisões que se tomaram para a implementação do repositório que irá suportar o RODA. Numa primeira secção enumeram-se os requisitos funcionais que serviram de ponto de partida a todo o trabalho realizado. A seguir e depois de discutir a arquitectura aplicacional apresentam-se as várias normas existentes para suportar a metainformação necessária à devida catalogação dos objectos digitais (descrição arquivística, descrição técnica, informação de preservação). Esta secção termina com a indicação dos elementos das várias normas que irão ser combinados no RODA para catalogar os objectos digitais. O passo seguinte consistiu na tomada de decisão sobre a estratégia a seguir para implementação do protótipo. O capítulo seguinte apresenta uma análise comparativa das duas

estruturas (consideradas as mais relevantes para este tipo de projecto), DSpace e Fedora, à luz dos requisitos funcionais discutidos no início. Por fim, no último capítulo apresentam-se as conclusões e as decisões tomadas para a implementação do RODA, que já está em curso. Neste capítulo englobam-se também as decisões tomadas relativamente ao tipo de objectos digitais que será possível armazenar no RODA. Nas conclusões e nas indicações do trabalho futuro indicam-se, em traços gerais, as estratégias que irão ser seguidas para suportar os vários tipos de objectos digitais.

OBJECTIVOS E REQUISITOS FUNCIONAIS (JCR)

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

- Requisitos funcionais para um arquivo digital, clientes e aplicações a integrar;
- Modelos conceptual, lógico e de dados de um arquivo digital;
- Estrutura de metainformação, de requisitos técnicos e organizacionais;
- Protótipo dum arquivo digital para preservar objectos digitais susceptíveis de conservação definitiva;
- 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.

Os produtos acima referidos serão acompanhados de relatórios de progresso e de um relatório final em que constem pormenorizadamente todos os processos e métodos de desenvolvimento utilizados na sua elaboração, sendo que dois destes já se encontram publicados no portal do projecto [1].

O protótipo de arquivo digital será planeado na perspectiva de obter um sistema capaz de assegurar todas as funcionalidades de um arquivo digital constantes da norma OAIS (Open Archival Information System) [2], nomeadamente, a integração (ingestão), a gestão e a disseminação de informação de arquivo. A limitação deste protótipo residirá na restrição de formatos a integrar.

Foram considerados nesta fase do projecto três classes de objectos digitais: documentos de texto (estruturado simples, estruturado com imagens e estruturado com tabelas), imagens bidimensionais e bases de dados relacionais.

O projecto contempla ainda alguns objectivos secundários, nomeadamente: a definição de uma política de arquivo para os objectos digitais produzidos pela Administração Pública nacional (avaliação e selecção); a definição de uma política de preservação para o arquivo digital; a criação ou identificação de modelos viáveis de financiamento para suportar o Arquivo Digital; a identificação e selecção dos esquemas de metainformação e a definição de uma taxionomia de propriedades significativas para cada uma das classes de objectos consideradas.

METAINFORMAÇÃO DE SUPORTE AO REPOSITÓRIO

Um repositório digital com as características do RODA tem necessariamente de ser suportado por um conjunto de esquemas de metainformação capazes de assegurar a realização de actividades elementares como a gestão de objectos digitais, facilitar a sua localização ou garantir a conservação do valor probatório.

Nesta secção é apresentada uma breve descrição dos diferentes esquemas de metainformação utilizados pelo RODA. Estes deverão ser compreendidos de acordo com a função desempenhada no interior do repositório (Figura 2).

A metainformação tem como objectivos: 1/ assegurar a autenticidade dos objectos digitais fixando e descrevendo

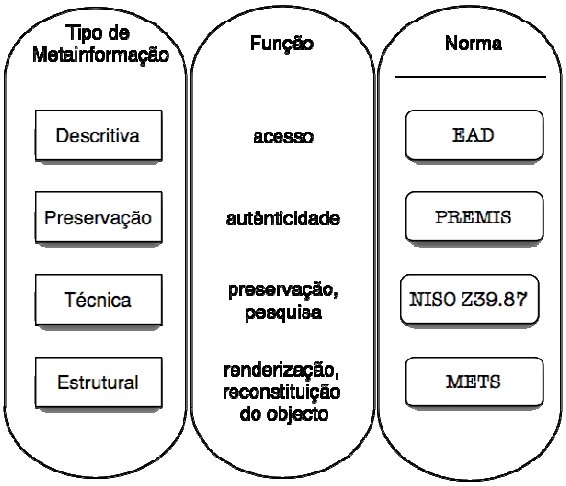

Figura 1 - Esquemas utilizados no esboço

propriedades essenciais destes, as quais não são imediatamente auto-explicativas. 2/ promover a eficácia na localização e recuperação dos objectos digitais e da informação neles contida, 3/ salvaguardar propriedades essenciais, intrínsecas e extrínsecas ao objecto digital, que permitam informar e assegurar o processo de preservar esses mesmos objectos ao longo do tempo

PREMIS Data Dictionary

Em 2003 a OCLC (*Online Computer Library Center*) e a RLG (*Research Libraries Group*) estabeleceram o grupo de trabalho designado *PREservation Metadata: Implementation Strategies* (PREMIS). Em Maio de 2005 o mesmo grupo apresentou o seu relatório final, o *Data Dictionary for Preservation Metadata* [3], que define o esquema de metainformação que passamos a apresentar.

O esquema está organizado segundo um modelo simples (**Figura**) identificando cinco tipos de entidades envolvidas nas actividades de preservação digital:

- *Object* - ou Objecto Digital, é a unidade discreta de informação no formato digital

- *Intellectual entity* – ou entidade intelectual, é um conjunto coerente de conteúdos, que pode ser razoavelmente descrito como uma unidade indissociável de informação (e.g. um livro, uma imagem, uma base de dados). Uma entidade intelectual pode conter outras entidades intelectuais no seu interior, por exemplo um livro pode conter uma imagem.

- *Agent* – ou agente, trata-se de uma pessoa, organização ou aplicação de software que participou num evento de preservação durante o tempo de vida de um Objecto.

- *Event* – ou evento, é uma acção que envolve pelo menos um objecto ou um agente conhecidos pelo repositório.

- *Rights* -, ou direitos, é um conjunto de um ou mais direitos ou permissões relativos a um Objecto e/ou Agente.

O RODA implementa todas as unidades semânticas apresentadas pelo PREMIS, i.e. *Objectos*, *Eventos*, *Agentes* e *Direitos,* excepto a *Entidade Intelectual.* A sua funcionalidade é assegurada pelo esquema de metainformação descritiva adoptado, i.e. o EAD.

No *PREMIS Data Dictionary* a entidade *Object* tem três subtipos: *file*, *bitstream* e *representation*. Um *ficheiro* (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 contíguo ou não contíguo de dados dentro de um *file*, que tem algumas propriedades comuns significativas para efeitos da preservação digital. Uma representação (*representation*) é um conjunto de ficheiros *(files)*, incluindo metainformação estrutural, necessários para a apresentação de uma Entidade Iintelectual (*Intellectual Entity*).

A entidade *Evento* (event) agrega metainformação sobre acções realizadas em torno dos objectos digitais custodiados. Um repositório com objectivos de preservação digital deverá registar *Eventos* por variadas razões. O registo deste tipo de informação permite assegurar a autenticidade dos objectos custodiados.

Todo o tipo de eventos poderão ser registados, sendo que as acções que provocam modificações nos objectos digitais assumem particular importância. Não obstante, actividades relacionadas com a criação de relações entre objectos, validações, análises de integridade, etc. poderão também ser registadas de modo a documentar as respectivas actividades de gestão.

EAD (Encoded Archival Description)

O EAD (*Encoded Archival Description*) [4] define metainformação descritiva e encontra-se na versão 2002. Esta metainformação permite descrever os objectos custodiados de forma contextualizada, ajudando os seus potenciais consumidores a categorizar e localizar a informação pretendida. Informação deste tipo é

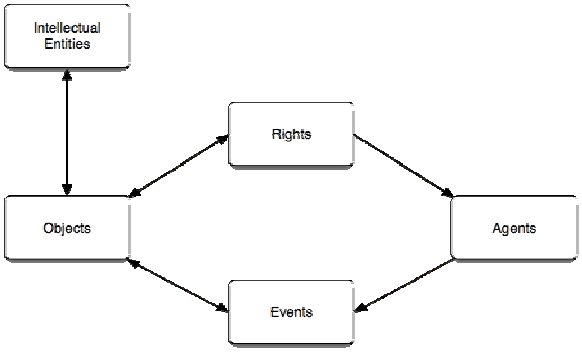

Figura 2 - PREMIS Data Model

vulgarmente utilizada por motores de busca para encontrar informação.

Uma instância EAD é constituída por três partes:

- eadheader - contém informação sobre a metainformação em si.

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

- archdesc - compreende informação sobre um fundo documental e sobre os respectivos materiais que o constituem.

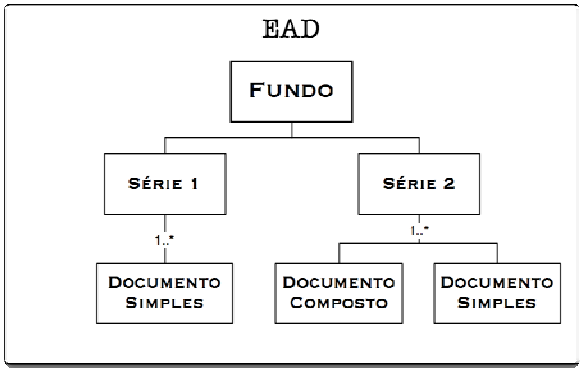

Figura 3 - Esquema de um EAD exemplo

Cada instância de um EAD contem um ou mais elementos XML do tipo <c> (i.e. *component*). Estes elementos podem ser aninhados de modo a criar uma estrutura hierárquica capaz de descrever um fundo documental na sua totalidade. Cada um destes elementos é caracterizado por um identificador único e um nível de descrição (atributo *level* do elemento <c>) que pode assumir um dos seguintes valores:

- Fundo - o conjunto de todos os documentos, independentemente da sua forma ou formato, organicamente criados e/ou acumulados por certa pessoa, família, ou instituição no decurso das suas actividades ou funções;

- Série - conjunto de documentos que dizem respeito a uma actividade funcional;

- Documento composto - uma unidade organizada de documentos agrupados para uso do criador ou no processo de organização arquivística por serem relativos a um mesmo assunto ou actividade;

- Documento simples - a unidade elementar de um arquivo, i.e. um documento intelectualmente indivisível.

Cada nível de descrição contém informação descritiva adequada, seguindo o modelo da ISAD(G) [5]. Como exemplos deste tipo de informação podemos realçar o título, datas extremas, historia biográfica, história custodial, âmbito e conteúdo, existência e localização dos originais e cópias, etc.

Para mais informações sobre o esquema EAD, é possível consultar as seguintes fontes de informação:

- Official EAD Version 2002 Web Site[1]
- Society of American Archivists[2]
- RLG Best Practices Guidelines for Encoded Archival Description[3]
- EAD Tools Survey[4]
- RLG EAD Report Card[5]

NISO Z39.87

Este esquema de metainformação define um conjunto de elementos capazes de caracterizar imagens digitais. O esquema utilizado data de 2002, no entanto, encontra-se neste momento em período de avaliação pública a versão 2005 que define uma reorganização dos seus elementos de modo a se tornar mais compatível com o esquema PREMIS.

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

Basic Image Parameters

Esta secção agrupa elementos fundamentais para a reconstrução e apresentação do objecto digital em interfaces gráficas. Eis alguns dos seus subelementos:

- MIMEType - o formato da imagem;
- ByteOrder - a ordem pela qual a informação binária se encontra representada;
- Compression - a tecnologia de compressão e o nível de compressão utilizados;
- ColorSpace - a palete de cores utilizada;
- DisplayOrientation - a orientação em que a imagem deve ser apresentada num monitor convencional;

Image Creation

Esta secção regista informação sobre aspectos logísticos e condições administrativas relativas à captura da imagem digital. Alguns dos seus subelementos são apresentados de seguida:

- SourceType - 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;
- DigitalCameraCapture - propriedades relevantes da câmara digital usada na captura da imagem.

Imaging performance assessment

O objectivo desta secção é descrever os atributos da imagem relacionados com a sua qualidade. Estes elementos servem como métricas para assegurar a fidelidade da imagem após a aplicação de técnicas de preservação, especialmente aquelas baseadas em migração. Exemplos de subelementos:

- XSamplingFrequency e YSamplingFrequency - a resolução da imagem nos dois eixos;
- ImageWidth e ImageLength - o tamanho da imagem nos dois eixos.

Change history

Esta secção tem como função documentar os processos de intervenção sobre a imagem durante todo o ciclo de vida da mesma.

- Image Processing - 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, consultar as seguintes fontes de informação:

- NISO Metadata for Images in XML Schema Official Web Site[6]
- NISO Z39.87 -200x Development page[7]

METS (Metadata Encoding & Transmition Standard)

O METS (*Metadata Encoding & Transmition Standard*) [6] trata-se de uma norma que permite associar

[1] http://www.loc.gov/ead
[2] http://www.archivists.org
[3] http://www.rlg.org/en/pdfs/bpg.pdf
[4] http://www.archivists.org/saagroups/ead
[5] http://www.rlg.org/ead-report-card

[6] http://www.loc.gov/standards/mix
[7] http://www.niso.org/standards/standard_detail.cfm?std_id=731

metainformação descritiva, administrativa e estrutural sobre objectos digitais.

Um documento METS é constituído por sete secções principais:

1. Cabeçalho METS - O cabeçalho METS contém metadados que descrevem o documento METS em si, incluindo informação sobre o criador, editor, etc.

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), conter metadados descritivos embebidos, ou ambos. Múltiplas instâncias de metadados descritivos, tanto internas como externas, podem coexistir num mesmo documento METS.

2. Metadados administrativos - A secção de metadados administrativos regista informação sobre como os ficheiros que constituem um objecto digital foram criados e armazenados, direitos de propriedade intelectual, metadados sobre o objecto original a partir do qual derivações poderão ter sido produzidas e informação sobre a proveniência dos mesmos (i.e., relações entre ficheiros originais e suas derivadas e informação sobre possíveis transformações aplicadas aos mesmos). Tal como os metadados descritivos, os metadados administrativos podem ser tanto externos ao documento METS como descritos internamente.

3. 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>, de modo a permitir a subdivisão de ficheiros por versão do objecto.

4. Mapa estrutural - O Mapa Estrutural é o núcleo do documento METS. Este esboça uma estrutura hierárquica para o objecto digital e liga os elementos dessa estrutura aos respectivos ficheiros, bem como aos metadados referentes a cada elemento.

5. Ligações estruturais - Esta secção permite aos criadores de METS registar a existência de hiperligações entre nós na hierarquia esboçada no Mapa estrutural. Esta secção tem particular interesse para quem pretende arquivar sítios *Web*.

6. Comportamento - Esta secção pode ser usada para associar comportamentos executáveis ao conteúdo no objecto METS. Cada *comportamento* é representado na interface por um elemento que define o subconjunto de comportamentos associado a esse comportamento particular. Cada comportamento também tem um mecanismo que identifica um módulo de código executável capaz de implementar e executar aquele subconjunto de comportamentos

Combinando diferentes esquemas

O RODA irá usar dois esquemas de metainformação primários, o EAD, para guardar a metainformação descritiva, e o PREMIS para guardar metainformação de preservação. Para além destes irão ser usados vários esquemas secundários que servirão para guardar metainformação técnica que não existe no PREMIS, ou que existe mas de forma insuficientemente detalhada como, por exemplo, o NISO Z39.87 para imagens fixas digitais.

Para cada tipo de documentos que o repositório irá armazenar poderá haver, caso seja necessário, um esquema de metainformação técnica para refinar o PREMIS. Irá ser usado ainda outro tipo de metainformação, trata-se de um esquema de metainformação estrutural que permite organizar objectos digitais constituídos por vários ficheiros. Um dos esquemas utilizados para este fim é o METS.

Na **Figura** podemos ver o *PREMIS Data Model* que faz referência a 5 entidades: Objectos, Eventos, Agentes, Direitos e Entidades Intelectuais, no entanto esta última não é descrita pelo PREMIS, sendo apenas referenciada porque está fora do domínio de aplicação deste esquema que se prende apenas com questões de preservação. Para descrever as entidades intelectuais é utilizado o EAD.

O EAD descreve a Entidade Intelectual a vários níveis, sendo o mais baixo aquele em que se efectiva a ligação ao PREMIS. Esta ligação é assegurada pela propriedade *linking Intellectual Entity Identifier* associada a cada representação descrita ao nível do PREMIS. Considera-se a representação uma manifestação "física" de uma entidade intelectual, sendo que uma dada entidade intelectual poderá assumir várias representações (e.g. um livro poderá ser representado através de um conjunto de imagens TIF, em PDF, em HTML, etc.).

O PREMIS guarda informação necessária à preservação de objectos digitais ao longo do tempo. Este esquema descreve representações, agregados de ficheiros necessários para apresentar uma entidade intelectual e descreve também estes mesmos ficheiros (i.e., o

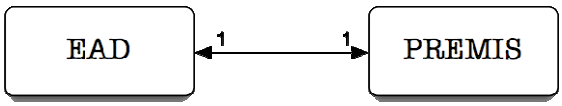

Figura 2 - Relação do EAD com o PREMIS

verdadeiro alvo dos processos de preservação). No entanto o PREMIS é um esquema generalista e não guarda metainformação técnica específica de um tipo de ficheiros, logo, para completar a metainformação técnica dos vários tipos de ficheiros irá ser usado o esquema de metainformação técnica adequado a cada tipo específico de ficheiros a ser preservado.

No caso de imagens digitais foi escolhido o esquema NISO Z39.87 (versão de 2002). Esta informação articula-se com o PREMIS sendo embebida no interior do campo *ObjectCharacteristics* do PREMIS.

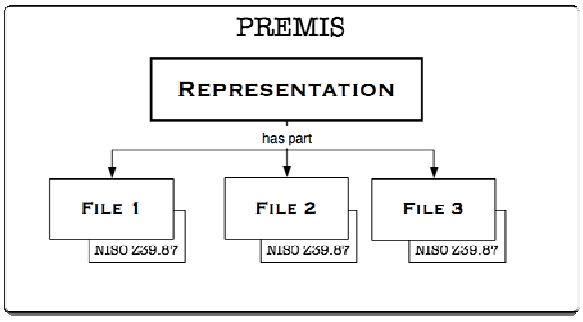

Figura 3 - Esquema exemplo PREMIS com extensão NISO Z39.87

Uma representação é descrita no esquema PREMIS por um elemento do tipo representação. Este elemento poderá conter elementos do tipo *Ficheiro* ou do tipo representação. Cada elemento *Ficheiro* descreve um ficheiro especificando propriedades como o seu tamanho, formato, data de criação, localização no sistema de ficheiros, etc. (Figura 4).

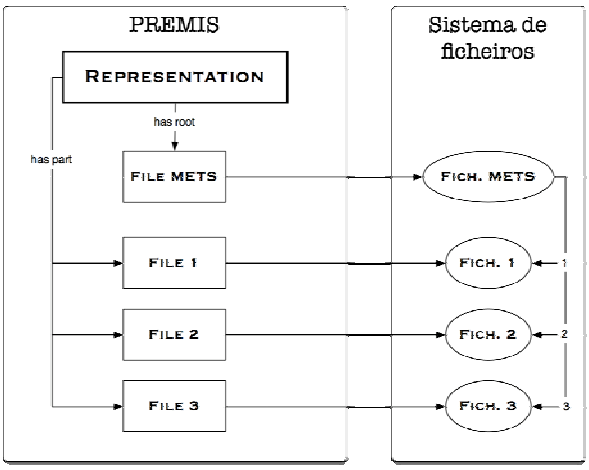

Figura 4 - Relação entre o PREMIS e o sistema de ficheiros

Embora todos os ficheiros de uma representação estejam referenciados no elemento correspondente, falta ainda metainformação estrutural que permita um único ponto de acesso a esta representação e a navegação, de forma ordenada, pelos vários ficheiros da representação. Apesar do PREMIS oferecer algumas formas de descrever a estrutura de uma representação, esta não seria suficiente para todos os casos possíveis. Foi, portanto, decidido utilizar outro esquema de metainformação e guardar no PREMIS uma referência para o mesmo através da propriedade *hasRoot*. No nosso caso de estudo utilizamos o METS como esquema de metainformação estrutural para imagens digitais. Outras classes de objectos, e.g. bases de dados relacionais, irão necessitar de esquemas mais adequados. A selecção desses esquemas ainda não fora realizada até data de escrita deste artigo.

Em suma, o PREMIS compreende metainformação técnica e de preservação associada a uma representação. Cada representação possui uma ligação a um componente EAD que contêm a metainformação descritiva. O EAD, por sua vez, aponta também para a representação PREMIS e para o ficheiro que serve de ponto de entrada na representação. Esta articulação resulta numa estrutura flexível e segura. Qualquer que seja a perspectiva sobre a representação (descritiva ou de preservação) é sempre possível alcançar e reunir toda a metainformação bem como os ficheiros que constituem a representação (**Figura 5**).

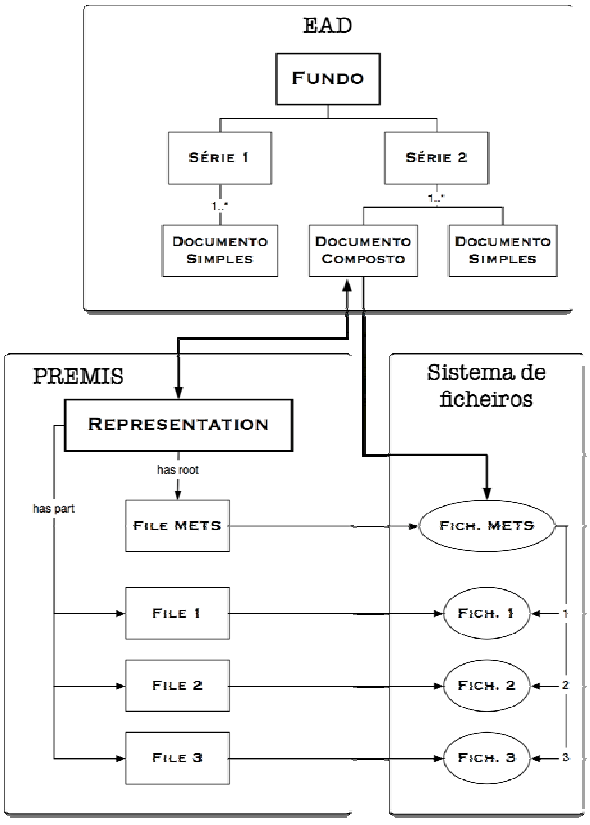

Figura 5 - Esquema completo (excepto NISO Z39.87)

DSPACE VS FEDORA

Implementar um repositório a partir do nada é um trabalho oneroso e lento. Dado o limite temporal inerente a este projecto, decidiu-se procurar junto dos diversos projectos *open-source* existentes um repositório que pudesse servir de plataforma de desenvolvimento para o RODA. No panorama do software livre há, efectivamente, dois candidatos que se destacam: DSpace [7] e Fedora [8].

Uma comparação destes dois sistemas foi realizada tendo por base nos requisitos funcionais do RODA. Estes requisitos estão divididos em três tipos distintos e encontram-se descritos nas tabelas 1, 2 e 3.

Tabela 1 - Processo 1: Ingestão

Requisito funcional; *componente*
O RODA tem que desenvolver uma interface intuitiva que suporte as transacções previstas durante o processo de ingestão; *Interface Gráfica para processo de Ingestão*
o RODA tem de ter a capacidade de registar informação administrativa sobre o cliente; *Registo de Produtores*
os documentos de orientação e regulamentação do RODA deverão estar disponíveis publicamente para consulta em linha; *página de ajuda do repositório ou site de apoio à incorporação*
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 auxiliar para produzir SIPs*
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*
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*
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*
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*
O RODA deve assegurar a capacidade de evitar a eliminação involuntário dos ficheiros admitidos à pré-integração; *Undelete, Trash*
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*
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*
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*
O RODA tem de assegurar a atribuição controlada de MI aos SIP e AIP resultantes; *Gestão controlada de metainformação*
O RODA tem de ter a capacidade de produzir e manter identificadores únicos persistentes baseados em especificações internacionais; *PIDs standard*
O RODA tem de ter documentação (MI) sobre os conversores/transformadores utilizados no processo de criação de AIP; *Metainformação sobre Agentes (Software)*
O RODA tem de assegurar o armazenamento dos SIP e AIP de acordo com a estratégia de armazenamento definida para cada tipologia (taxionomia) de SIP admitida; *Documento do Francisco "Taxionomias de Objectos Digitais a integrar no RODA"*
O RODA deve produzir um relatório com dados sobre o sub-processo de integração a ser apresentado ao Cliente e à Administração; *Produção de relatórios sobre o processo de ingestão*

Tabela 2 - Processo 2: Gestão

Requisito funcional; *componente*
O RODA deve possuir uma interface que permita a interacção entre o gestor do RODA e os AIP utilizando todos as ferramentas necessárias para a realização das operações de gestão previstas; *Ferramenta de Administração*
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*
O RODA deve ter a capacidade de gerar de forma automática MI para cada evento desencadeado; *Gerar eventos PREMIS (<event>)*
Deve ser assegurada a capacidade de atribuir MI descritiva a nível agregado (classes de AIP); *Fazer descrição de conjuntos de Representações (AIPs)*
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*
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*
O RODA deve assegurar que na sequência de um evento será adicionada MI aos objectos (AIP; Infra-estrutura tec.) alvo desse evento; *Actualização da MI PREMIS relativa a eventos*
Após um evento de actualização (migração) os AIP resultantes deverão ser confirmados quanto à

sua integridade e inteligibilidade sendo para isso sujeitos a processo de validação; *Efectuar uma validação no final de qualquer acção de altere os AIPs (Workflow supracitado)*

Tabela 3 - Processo 3: Disseminação

Requisito funcional; *componente*
O RODA tem de manter uma interface amigável para gerir o processo de disseminação e interagir com o cliente; *Interface Gráfica para Disseminação*
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*
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*
A certificação do DIP tem de respeitar os métodos legalmente reconhecidos em Portugal (Assinatura digital); *Assinaturas digitais nos DIPs*

Como podemos observar o DSpace implementa mais requisitos do que o Fedora, o que nos leva a pensar que será o mais apropriado. No entanto é necessário estudar a adequação da arquitectura de cada repositório para incluir os esquemas de metainformação escolhidos (PREMIS e EAD).

Conclusão

Essencialmente, o facto do DSpace apresentar uma estrutura demasiado específica dificulta a possibilidade de o adaptar às necessidades do nosso arquivo digital. Nesta medida, a contrapartida existente entre as funcionalidades oferecidas pelo DSpace e a consequente rigidez da sua arquitectura contra a flexibilidade mas ausência de ferramentas oferecidas pelo Fedora, levaram a equipa de projecto a optar pela segunda opção.

TAXONOMIAS DOS OBJECTOS DIGITAIS

As classes de objectos digitais a integrar nesta primeira fase do RODA serão: documentos de texto (estruturado), imagem fixa (i.e. *bitmap*) e as bases de dados relacionais, tendo sido considerado para cada uma delas requisitos em termos de preservação das suas propriedades que garantissem a autenticidade e fiabilidade, assim como o seu valor evidencial.

Documentos de texto

O texto poderá ser estruturado, i.e. conter elementos embebidos como tabelas e imagens fixas.

Considerou-se essencial a preservação do conteúdo e da estrutura, tanto ao nível da estruturação como da formatação do texto. Foram também considerados relevantes os atributos de formatação, como negritos ou sublinhados visto comportarem significado semântico que não deve ser ignorado.

Atendendo a esta determinação optou-se pela utilização de PDF (Portable Document Format) como formato de preservação, formato este que será migrado para PDF/A (Portable Document Format/Archival) assim que surjam no mercado ferramentas que o permitam.

O PDF/A é um formato produzido pela Adobe, sendo um software proprietário com código publicamente acessível, sendo uma norma da ISO - ISO 19005-1:2005 que, muito embora não seja considerada como uma norma totalmente adequada a preservação digital a sua utilização para preservação digital, é considerada como aceitável por entidades internacionalmente conotadas com esta área (e.g. NARA). Este formato é especificamente destinado a texto e imagem embora apenas alguns formatos de imagens sejam suportados. Essa limitação não interfere com os objectivos do RODA.

Imagem digital

Do ponto de vista de taxionomia não foi necessária qualquer categorização visto que a imagem é um objecto digital simples e consistente e sem grandes variações estruturais.

Considerou-se, no entanto, que se deve assegurar a preservação do cabeçalho, componente estrutural comum a todas as imagens, como parte integrante do ficheiro de imagem.

Devido ao facto dos cabeçalhos de alguns formatos não serem totalmente conciliáveis com o formato de preservação escolhido, TIFF sem compressão, seja por conter poucos elementos informativos e/ou por serem diferentes dos elementos previstos no formato de preservação, decidiu-se que se deve recolher a metainformação existente no cabeçalho e integrá-la dentro do esquema PREMIS e só depois migrar o formato para TIF mantendo no novo cabeçalho toda a metainformação que for possível transpor, sem preocupação de assegurar que a sua totalidade seja incluída no novo formato. A perda de alguma metainformação do cabeçalho da imagem é aceitável desde que seja conservada na metainformação de preservação externa.

Bases de dados relacionais

A primeira ponto a considerar no contexto da integração de bases de dados relacionais foi, se seria possível separar os dados das funcionalidades oferecidas pelo motor da base de dados, ou seja, diferenciar e avaliar separadamente a parte estática (tabelas, dados, relações) da parte dinâmica (interacções com os dados, procedimentos, interrogações, etc.) Havia funcionalidades que este último assegura que poderiam à partida ser

consideradas como elementos importantes do ponto de vista da autenticidade e fiabilidade da informação. No entanto, atendendo à multiplicidade de soluções seguidas pelos diversos fabricantes de software, dependentes quase exclusivamente de soluções de programação fechadas, afastou-se a possibilidade técnica de preservar este subsistema.

A questão da metainformação de acessibilidade ser normalmente guardada como dados numa tabela incluída na base de dados foi um argumento válido para diminuir a relevância do SGBD. O principal problema é que a sua não preservação limita consideravelmente as opções de manipulação posteriormente oferecidas ao utilizador.

Considerou-se, no entanto, ser impossível neste momento, devido à complexidade técnica inerente, preservar a componente dinâmica do SGBD, conclusão que está de acordo com documentos e soluções internacionais produzidas no domínio de preservação de bases de dados. Nestas circunstâncias consideram-se como objectos de preservação a estrutura da base de dados (tabelas, relações entre tabelas) e os dados propriamente ditos.

Definiu-se ainda como formato de preservação o XML de acordo com esquemas a desenvolver.

CONCLUSÃO E TRABALHO FUTURO

O projecto actual é necessariamente incompleto pois que se trata de um protótipo. Há no entanto questões que se prendem vir a ser analisadas na segunda fase do projecto (RODA 2) para o qual se prevê o início em Maio de 2007.[8]
Estes problemas são de natureza organizacional e técnica.
Abordando-se a questão do modelo financeiro mais adequado à sustentação de um Repositório digital,; o aspecto organizacional onde se incluem questões sobre o modelo de gestão mais apropriado para uma estrutura deste tipo e qual o forma de distribuição territorial de um arquivo digital com estas características.
Pretende-se ainda alargar o conjunto de tipologias de objectos digitais a integrar, de acordo com as tendências identificadas na administração pública.
A infra-estrutura física cujas bases serão elencadas e definidas ainda nesta fase do projecto, deverá ser adquirida e implementada de acordo como o respectivo plano de desenvolvimento.
O objectivo é montar uma estrutura definida de acordo com modelo OAIS, capaz de dar resposta efectiva às actuais carências na administração pública relativamente a custódia e preservação de informação de arquivo electrónica.

[8] Encontram-se já assegurados os meios indispensáveis para assegurar a continuidade do RODA durante o ano de 2007

AGRADECIMENTOS

O desenvolvimento deste projecto foi financiado pelo POAP (Programa Operacional da Administração Pública)

REFERÊNCIAS

[1] National Archives (Instituto dos Arquivos Nacionais/Torre do Tombo) and University of Minho, "RODA (Repositório de Objectos Digitais Autênticos) Web site," vol. 2006, 2006.

[2] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS) - Blue Book*. Washington: National Aeronautics and Space Administration, 2002.

[3] PREMIS Working Group, "Data dictionary for preservation metadata: final report of the PREMIS Working Group," OCLC Online Computer Library Center & Research Libraries Group, Dublin, Ohio, USA, Final report 2005.

[4] Library of Congress, "EAD - Encoded Archival Description," vol. 2004: Library of Congress.

[5] International Council on Archives, "ISAD(G): General International Standard Archival Description, Second edition," International Council on Archives 0-9696035-6-8, 1999.

[6] Library of Congress, "METS - Metadata Encoding & Transmission Standard."

[7] Hewlett-Packard Company and MIT Libraries, "DSpace Web site," vol. 2005.

[8] University of Virginia and Cornell University, "Fedora Web site," vol. 2005.