XML e Preservação Digital

José Carlos Ramalho[1] and Miguel Ferreira[1] and Rui Castro[2] and Luis Faria[2] and Francisco Barbedo[2] and Luis Corujo[2]

[1] Dep. Informática, Universidade do Minho
[2] Instituto dos Arquivos Nacionais, Torre do Tombo

Resumo Actualmente estamos a substituir gradualmente os documentos físicos por documentos digitais. A constatação de que cada vez há mais informação apenas em suporte digital levanta uma série de preocupações relacionadas com a preservação de todo este manancial de informação. Não é de estranhar, portanto, que a Preservação Digital tenha emergido como uma área de investigação que tem adquirido cada vez mais importância. A principal preocupação de todos os que tentam contribuir para a área é como garantir uma maior longevidade ao material digital que é produzido diariamente. Por exemplo, material produzido está dependente de plataformas de hardware e software que normalmente se tornam obsoletos em 5 anos. O XML como formato neutro para a representação de informação surge naturalmente neste contexto. São já vários os dialectos XML produzidos e usados para e na Preservação Digital: PREMIS, METS, NISO MIX, etc. Neste artigo, caracterizámos o estado actual desta área, identificando várias normas e mostrando com um caso de estudo real (RODA: Repositório de Objectos Digitais Autênticos), como é que se podem usar e combinar aquelas normas numa solução de Preservação Digital.

1 Introdução

O Instituto de Arquivos Nacionais da Torre do Tombo (IAN/TT), tem na sua função de preservação histórica um grande desafio perante o crescimento da produção de documentos digitais pela função pública, devido à sua própria evolução no sentido do governo electrónico. 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.

Foi com esse objectivo que se iniciou o projecto RODA (Repositório de Objectos Digitais e Autênticos). Numa primeira instância o RODA visa a construção de um protótipo exemplificativo de uma solução de preservação digital. O protótipo pode, posteriormente, vir a ser desenvolvido na forma de produto na escala necessária para responder cabalmente às necessidades das organizações no que respeita à perservação de objectos digitais de conservação permanente.

Procura-se desta forma iniciar um processo que leve o IAN/TT a responder às solicitações governamentais e comunitárias no contexto do governo electrónico.

Neste artigo, apresentam-se de forma sintética os passos que conduziram à implementação do protótipo do RODA. Primeiro discute-se a informação que se vai guardar

e como é que esta irá ficar organizada. Depois apresenta-se a primeira abordagem à arquitectura do RODA. E, por fim, conclui-se indicando os passos e as escolhas que guiaram a sua implementação.

2 Arquitectura do Repositório

Qualquer repositório digital tem uma parte substancial da sua estrutura assente em normas de metainformação. Do repositório mais simples, como um repositório doméstico de música digital, até ao repositório governamental criado para preservar o conhecimento e o legado histórico de uma população, a metainformação é usada como suporte das funcionalidades mais básicas do repositório, e. g. facilitando o acesso aos objectos digitais armazenados.

Um repositório digital que tem como objectivo a preservação a longo prazo dos materiais custodiados tem que integrar diversos tipos de metainformação:

Descritiva – é essencial para a organização do repositório e para o acesso aos objectos digitais armazenados.

Preservação – é essencial para garantir a autenticidade dos objectos armazenados fornecendo evidências da sua proveniência e de todas as acções sobre eles realizadas.

Técnica – é essencial para garantir o bom estado dos objectos no repositório e para garantir o respectivo acesso continuado.

Estrutural – é essencial para organizar objectos digitais complexos (ex: um livro é composto por capítulos, estes por secções e estas por páginas).

Em termos funcionais, a grande maioria dos repositórios digitais segue o modelo de referência da OAIS ("*Open Archival Information System*") [TS04,fSDS02]. Este foi também o modelo escolhido para a implementação do RODA (fig.1).

Pode-se descrever este modelo funcional da seguinte maneira: o produtor prepara a informação que quer preservar organizando-a num pacote especial (SIP - *"Submission Information Package"*); estes pacotes são enviados ao sistema que sabe analisá-los, verificá-los e retirar de lá a informação arquivando-a no repositório, transformando um SIP num AIP (*"Archival Information Package"*; o administrador pode realizar acções (verificação, correcção, preservação,...) sobre os AIPs; por sua vez, o consumidor pode realizar pesquisas sobre os AIPs e a dada altura pode solicitar um determinado objecto digital que lhe é oferecido na forma de um DIP (*"Dissemination Information Package"*), ou seja, o AIP solicitado é transformado num DIP que é fornecido ao consumidor.

Esta explicação resume o modelo OAIS e resume também o funcionamento pretendido no RODA.

Na implementação deste modelo começou-se por limitar o tipo de objectos digitais que se irão tratar (no contexto de um protótipo não seria viável contemplá-los todos). No protótipo inicial do RODA foi decidido que seriam contemplados os seguintes tipos de objectos digitais:

texto estruturado – que poderá conter tabelas e imagens fixas. O formato de preservação para texto estruturado será o PDF/A.
Este formato é capaz de preservar documentos de texto estruturado (com tabelas e imagens) mantendo o aspecto e paginação do documento original.

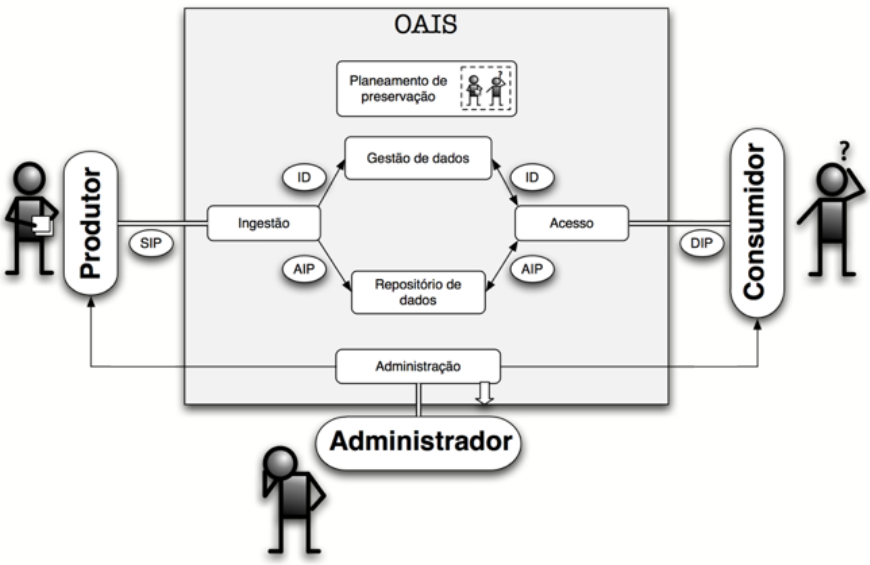

Figura 1. Modelo OAIS

as imagens bidimensionais fixas – fotografias, digitalizações, ... O formato de preservação para imagens fixas será o TIFF não comprimido. Este formato é aberto e muito bem suportado por inúmeras ferramentas de acesso livre e portanto é um bom formato de preservação.

bases de dados relacionais – para o protótipo prepararam-se interfaces com o MS Access, MS SQL Server e Oracle. O formato de preservação de bases de dados relacionais será o XML. Mais precisamente, as bases de dados serão migradas para DBML[JLRH02] ("*Data base Markup Language*"). Os alvos de preservação serão a estrutura da base de dados (tabelas e relações entre tabelas) e os dados. As funcionalidades associadas a *stored procedures* não serão alvo de preservação nesta fase. A eventual perda de informação que possa ocorrer ao não perservar outras funcionalidades (como os *stored procedures*) não é preocupante, porque a representação original de todos os objectos ingeridos será preservada e, portanto, será sempre possível no futuro derivar novas representações mais ricas que já incluam funcionalidades extra como as *stored procedures*.

Para mais detalhes sobre as resoluções relativas às taxionomias de objectos pode ser consultado [Bar06b]. As propriedades significativas a preservar sobre cada tipo de objecto foram alvo de alguma discussão e o resultado final é comparável ao estudo publicado por Ruusalepp[Ruu02] e é apresentado na secção seguinte.

3 Estrutura interna do repositório

Para definir a estrutura do repositório houve que decidir que informação se iria guardar e como é que esta seria estruturada. Além das representações físicas dos objectos é necessário guardar informação descritiva para facilitar o acesso e a procura, informação sobre as características originais do objectos (informação técnica) e informação de preservação sobre cada acção que é realizada sobre o objecto pois só assim se poderá garantir a autenticidade do mesmo e a sua possível reutilização como meio de prova ou outro. A estas há ainda a adicionar a informação estrutural que irá servir como agregador/organizador dos pacotes de informação que irão circular na arquitectura: SIP, AIP e DIP.

Ou seja, vamos ter uma solução que combina metainformação descritiva, técnica e de preservação. Dos objectivos do projecto cedo se concluiu que a metainformação descritiva e de preservação iriam representar um papel principal enquanto que a técnica teria um papel secundário. De realçar que todos os esquemas de metainformação adoptados e apresentados nas secções seguintes são dialectos XML.

3.1 Metainformação Descritiva

Há várias linguagens de anotação XML para especificar metainformação descritiva: Dublin Core[Wei97], MARC[mar06], Encoded Archival Description (EAD)[ead98].

Como o RODA será um arquivo digital interessava adoptar um esquema de metainformação descritiva multinível que permitisse descrever de forma estruturada as entidades intervenientes e as suas relações hierárquicas, pelo que a escolha recaiu sobre o EAD.

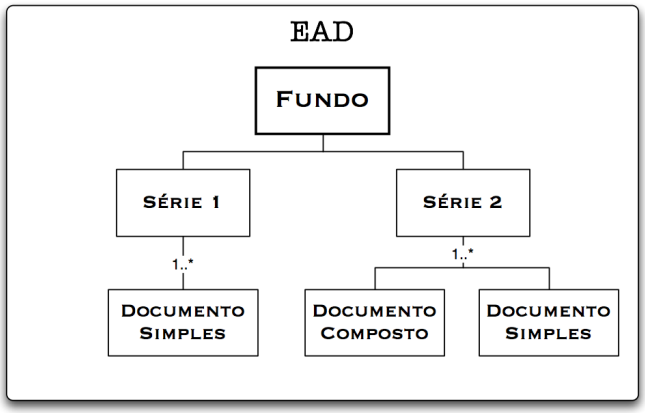

Figura 2. Esquema de um EAD exemplo

O EAD (*Encoded Archival Description*) permite definir metainformação descritiva. Esta metainformação descreve a informação de forma contextual, ajudando a categorizar e a procurar a informação. A informação deste tipo é utilizada por motores de busca para procurar informação.

Uma instância EAD contém três partes:

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

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

<archdesc> - contém a descrição de um fundo documental e a informação contextual e administrativa associada.

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 (*level*, fig. 2), que pode ter o valor[3]:

fundo - o conjunto de todos os arquivos, 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 reunidos porque fazem parte da mesma acumulação, processo de inserção ou actividade,ou partilham uma forma particular, ou outra qualquer relação relativa a sua criação, ingestão ou uso.

documento composto - uma unidade organizada de documentos agrupados para uso do criador ou no processo de organização arquivística, porque são relativos ao mesmo assunto ou actividade. É normalmente a unidade básica de uma série.

documento simples - A mais pequena e intelectualmente indivisível unidade de arquivo.

Cada nível de descrição contém informação descritiva adequada, seguindo o modelo da ISAD(G)[oA99]. Como exemplos de campos existem: título, datas importantes, historia biográfica, história custodial, âmbito e conteúdo, existência e localização dos originais e cópias, etc.

3.2 Metainformação de Preservação

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*[Gro05], que define o esquema que se resolveu adoptar no RODA.

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

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

[3] estes valores são os considerados pela nossa implementação, usando o atributo *otherlevel* do EAD, na definição oficial do EAD estes valores diferem um pouco

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 associado com 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 existente (EAD, MARC, MODS, Dublin Core, etc.) e porque é demasiado específica do domínio em consideração.

No *PREMIS Data Dictionary* a entidade *Object* tem três subtipos: *representation*, *file* e *bitstream*. Um *file* é uma sequência de *bytes* com ordem e nome, reconhecida por um sistema operativo. Um *file* tem propriedades como permissões, tamanho e data da última modificação. Um *bitstream* é um conjunto de dados dentro de um *file*, que tem algumas propriedades comuns significativas para efeitos da preservação digital. Uma *representation* é um conjunto de *files*, incluindo metainformação estrutural, necessários para renderização 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 (ou seja, criam uma nova versão) de um objecto digital é fundamental para manter a proveniência digital, um elemento chave para a autenticidade. Acções que criam relações ou que modificam relações existentes são importantes para explicar as mesmas relações. Até acções que não alteram nada, como validações e análises à integridade nos objectos, podem ser importantes registar para efeitos de gestão.

3.3 Metainformação Técnica

É com esta metainformação que se descrevem as características técnicas dos ficheiros e dos seus formatos. De momento, no RODA, utiliza-se NIZO MIX[nis06] para imagens e documentos de texto e DBML para as bases de dados relacionais.

O NIZO MIX 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 comentário a versão de 2005, que vem trazer uma nova organização ainda 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.
2. **Image Creation** - algo como metainformação técnica descritiva, dá informação sobre aspectos logísticos e condições administrativas relativas à captura da imagem digital.

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

O DBML[JLRH02] permite descrever uma base de dados no seu todo: informação sobre o ambiente de funcionamento, estrutura da base de dados e a informação presente na base de dados no momento da captura.

Os objectos que irão circular nos fluxos do repositório têm uma estrutura abstracta ilustrada na figura 3.

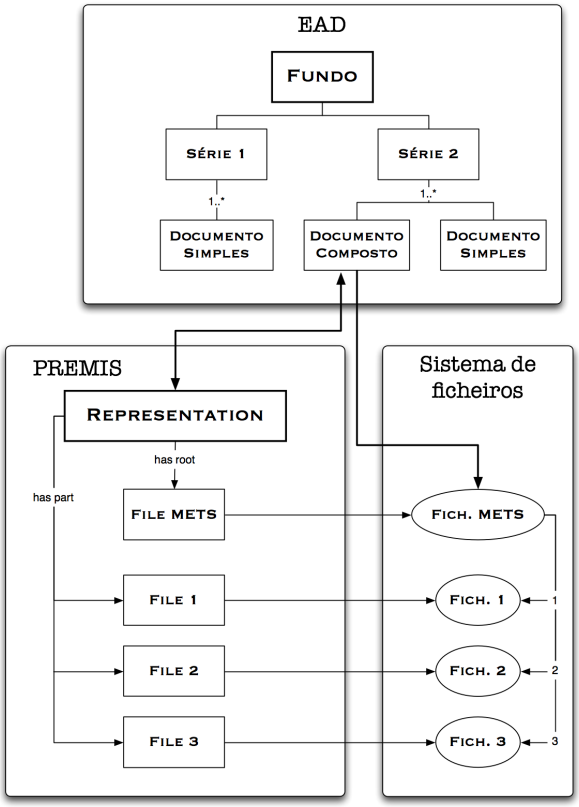

Figura 3. Esquema completo (excepto m.i. técnica)

Esta estrutura é depois implementada com algumas variações materializando os objectos que se designaram por SIP, AIP e DIP, que se definem nas secções seguintes.

3.4 SIP, AIP e DIP

Tendo decidido os esquemas de metainformação a usar no RODA passa a ser possível definir os dados que vão circular nos fluxos do protótipo.

O SIP ("Submission Information Package") é o formato usado para ingerir conteúdos no repositório digital. É composto pelo objecto digital (a sua representação física), por um conjunto de metainformação (descritiva em EAD, técnica e de preservação (em PREMIS – todo o registo de eventos previamente aplicados àquele objecto) e por uma descrição estrutural (uma espécie de lista descritiva do que vai dentro do pacote). Esta descrição estrutural é feita usando outra linguagem XML, o METS[met06,McD06] (Metadata Encoding & Transmition Standard) é uma norma para codificar metainformação descritiva, administrativa e estrutural sobre objectos guardados num repositório digital).

Na ingestão do SIP pelo repositório, a metainformação é dividida nos seus componentes funcionais que recebem um tratamento diferenciado. Neste processo, um SIP é transformado num AIP ("Archival Information Package"). Um AIP representa a forma de como um objecto digital é armazenado no repositório. O AIP tem a ele associadas a metainformação técnica e de preservação que são essenciais para a execução dos eventos de preservação. A metainformação descritiva é armazenada separadamente numa base de dados pois tem outros objectivos como facilitar o acesso e a procura de objectos digitais no repositório.

O DIP ("Dissemination Information Package") ao contrário dos outros dois tem um formato mais livre. O DIP é o formato de saída do repositório e podemos ter vários. Numa fase inicial serão suportados três DIP: um que corresponde a uma cópia da representação original, outro que corresponde a uma vista Web sobre um determinado objecto e outro que corresponde a um empacotamento num ficheiro comprimido em formato ZIP do objecto digital.

4 Implementação

Implementar um repositório de raiz é um trabalho bastante extenso e fora dos objectivos do RODA. Existem várias iniciativas *open-source* nos quais um repositório deste tipo se pode basear, mas há dois candidatos que se destacam no panorama actual: DSpace e Fedora.

O DSpace[4] 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 para instituições de investigação o poderem utilizar na sua forma original, ou modificar e estender conforme as necessidades. Muitas instituições de investigação por todo mundo utilizam o DSpace como solução para os mais variados tipos de arquivos digitais.

O Fedora é uma plataforma tecnológica *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

[4] `http://dspace.org`

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.

4.1 DSpace ou Fedora?

O passo seguinte foi a selecção da plataforma tecnológica sobre a qual seria desenvolvido o protótipo. Com esse objectivo elaborou-se um caderno de encargos bastante detalhado sobre os requisitos funcionais do projecto[Bar06a]. À luz destes requisitos fez-se um estudo comparativo das duas plataformas cujos resultados se apresentam de forma resumida no gráfico da figura 4.

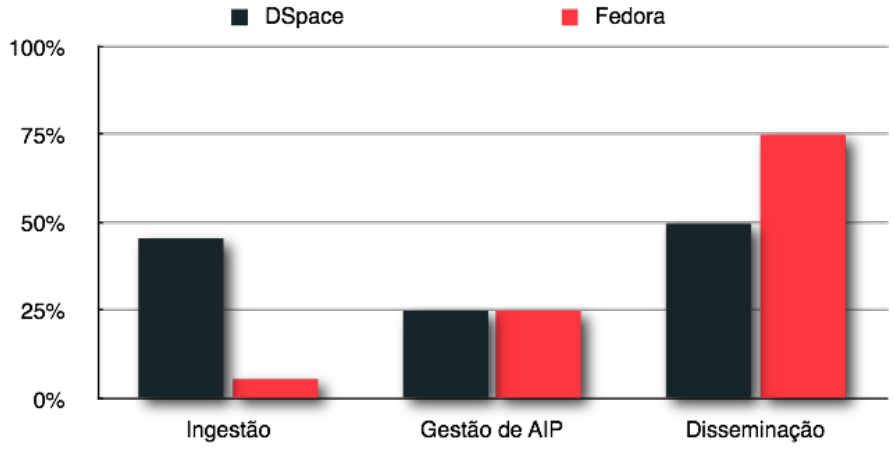

Figura 4. Comparação de requisitos entre as 2 plataformas

Como se pode ver nesta mostra de resultados, os requisitos foram agrupados nas três componentes funcionais principais do modelo OAIS: Ingestão, Gestão de AIP e Disseminação.

Observando a figura 4 seria de prever que o DSpace seria a plataforma seleccionada uma vez que ganha no cumprimento do componente principal que é a ingestão e está a par na manutenção do repositório.

No entanto, depois de se comparar a estrutura abstracta pretendida para o RODA (figura 3) com as estruturas que as duas plataformas implementam verificou-se que a estrutura do DSpace seria completamente desadequada para o fim pretendido.

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).

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. Esta decisão coincide com a conclusão de um estudo comparativo de repositórios baseados em software livre levado a cabo no âmbito do projecto "Open Access Repositories in New Zeland"[iNZ06].

4.2 Fedora: implementação

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:

PID - identificador único persistente.

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

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.

4.3 Tipos de Objectos

O RODA irá distinguir 3 tipos de objectos: *Objectos de Descrição (OD)*, *Objectos de Representação (OR)* e *Objectos de Preservação (OP)*. Os *Objectos de Descrição* guardarão informação descritiva, os *Objectos de Representação* conterão uma representação de um documento descrito num *Objecto de Descrição* e os *Objectos de Preservação* serão usados para guardar metainformação de preservação (PREMIS).

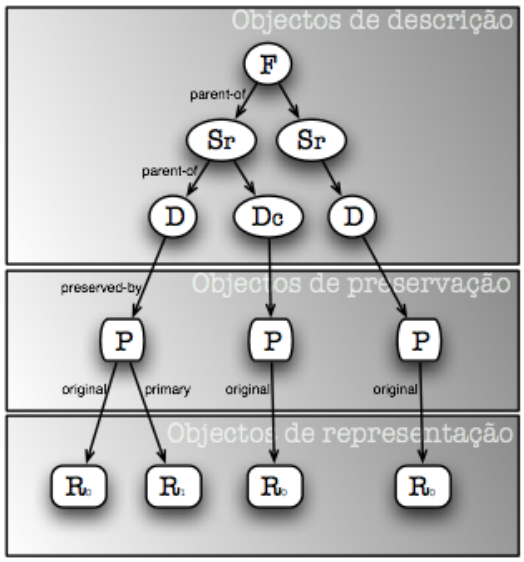

Figura 5. Objectos do RODA divididos por camadas

Todos os objectos relativos a um determinado fundo presente no repositório estarão relacionados através do mecanismo de relações do Fedora (usando RDF[5] – Resource Description Framework) formando uma árvore de descrição arquivística. Na raiz da árvore estará um *OD* que descreve o fundo, conectados a este estarão as descrições dos sub-fundos ou séries e estes por sua vez estarão conectados às unidades de descrição que forem necessárias para descrever adequadamente o fundo. Os *OP* estarão ligados às folhas dos *OD* e guardarão toda a metainformação de preservação relativa às representações da entidade intelectual descrita no *OD*. Por sua vez, os *OR* estarão associados aos *OP* e conterão as representações da entidade intelectual.

4.4 Conteúdos dos Objectos

Para cada tipo de objecto do repositório, *OD*, *OR* e *OP*, é necessário definir o seu conteúdo.

Os OD *são constituídos por:*

RELS-EXT - documento XML/RDF com informação sobre as relações entre o objecto e outros objectos do repositório.

DC - ficheiro com metainformação Dublin Core em formato XML. Este ficheiro está presente em todos os objectos Fedora por omissão. No RODA apenas terá o titulo (title) e o identificador (identifier) que representam o mínimo exigido pelo Fedora.

[5] `http://www.w3.org/RDF`

EAD - ficheiro com metainformação descritiva EAD em formato XML. Este ficheiro XML não será um EAD, mas um EADPART. Este formato será basicamente um elemento `<c>` do EAD mas chamado `<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 (para efeitos de optimização houve que optar por uma granularidade mais fina).

Por sua vez, os OP *são constituídos por:*

RELS-EXT - (ver descrição acima).
DC - (ver descrição acima).
PREMIS+ - ficheiro(s) com metainformação de preservação PREMIS.

Por fim, os OR *têm a seguinte estrutura :*

RELS-EXT - (ver descrição acima).
DC - (ver descrição acima).
METS - ficheiro com metainformação estrutural caso a representação seja composta por mais do que um ficheiro.
FICHEIRO+ - ficheiro(s) que compõem a representação.

4.5 Relações entre Objectos

As relações entre os vários objectos estarão descritas nos ficheiros RELS-EXT de cada um dos objectos que possui ligações a outros. Este(s) ficheiro(s) está(ão) no formato XML/RDF e descreve(m) 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 também de 3 tipos e serão as seguintes:

parent-of - relação entre dois *OD*.
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 (*SIP*), a representação primária (*AIP*) e eventuais representações alternativas (*DIP*).

5 Arquitectura funcional do protótipo

O Fedora oferece uma interface ao utilizador muito pobre, pouco amigável e insuficiente para os requisitos do repositório. Além disso existe um conjunto de requisitos que não estão satisfeitos pelo Fedora, como por exemplo: ingestão com *workflow*, tarefas de gestão dos AIPs, sistemas de procura e navegação da informação baseados nos novos esquemas de metainformação, etc. O Fedora foi desenvolvido como uma arquitectura de serviços e todas as funcionalidades associadas são disponibilizadas como *web services*. Esta arquitectura permite que o repositório cresça em funcionalidades sem que

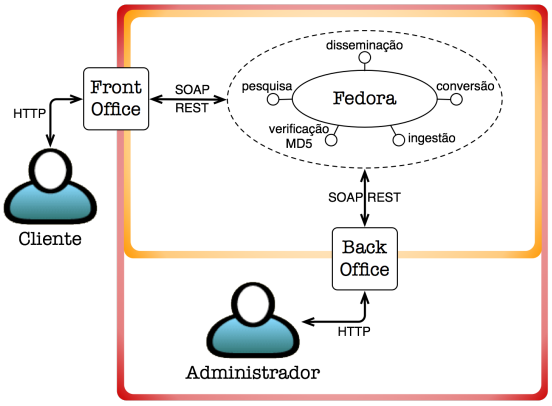

Figura 6. Arquitectura funcional do protótipo: *Front* e *Back Office*

o seu núcleo seja alterado e torna cada novo serviço independente da tecnologia usada noutros serviços já existentes não prejudicando no entanto a interoperabilidade entre os mesmos. Todas as componentes necessárias ao cumprimento dos requisitos serão desenvolvidas através de *web services* do Fedora e uma interface gráfica será disponibilizada para estes serviços na forma de um *Front Office*, para o acesso por parte dos clientes, e de um *Back Office*, para administração.

Como se pode observar na figura 6 foram identificados dois actores principais: o Cliente e o Administrador, que usam respectivamente um *Front Office* e um *Back Office*. Note-se que ambos são uma porta para uma zona mais reservada e que o Administrador já se encontra numa dessas zonas. Estas zonas de segurança são delimitadas por *firewalls*, que impedem o contacto directo com o Fedora e asseguram que a comunicação entre o Administrador e o *Back Office* seja realizada numa zona mais segura e controlada, diminuindo a possibilidade de ataques à integridade e confidencialidade do arquivo.

O Fedora possibilita que novos disseminadores de conteúdo sejam criados em tempo de execução e consequentemente novas formas de acesso aos mesmos conteúdos. Como o *Front Office* revela estes conteúdos por meio de disseminadores ao cliente, este terá de ter mecanismos para aceder a estes disseminadores e um meio de adicionar funcionalidade em tempo de execução para tratar estes disseminadores como necessário.

Para exemplificar este caso imaginemos que foi criado um novo disseminador para uma representação de um livro que oferece um método para obter uma imagem derivada de uma página do livro (cujo número é argumento do método). É claro que deve ser criada uma interface gráfica que usa este método para dar acesso ao conteúdo do livro ao cliente e que esta interface gráfica deve ser adicionada ao *Front Office* sem que haja necessidade de este ser recompilado.

Este é um caso clássico da *pattern Component Configurator* [SSRB00] que possibilita uma aplicação adicionar e remover os seus componentes em tempo de execução sem ter que modificar, recompilar ou estáticamente adicionar os componentes à aplicação.

Por falta de bases de desenvolvimento adequadas aos requisitos do projecto o *Front Office* e *Back Office* estão a ser desenvolvidos de raiz.

6 Conclusão

Os objectivos do RODA eram a construção de um protótipo. Para a construção desse protótipo especificaram-se requisitos funcionais e estruturais. À luz destes requisitos foi possível desenvolver estudos que levaram à especificação do modelo de dados e a uma série de decisões de implementação. Muitas destas decisões foram posteriormente validadas quer por experimentação da equipe do projecto quer por comparação com os resultados de equipes internacionais a trabalhar em projectos semelhantes.

Neste momento, a preservação digital é uma área que está a suscitar interesse na comunidade internacional. Naquilo que tenta fazer, o RODA é um projecto pioneiro: trabalha com normas que ainda estão a ser especificadas, algumas incompletas e tenta resolver tecnologicamente problemas que ainda não foram resolvidos na plataforma seleccionada (Fedora).

Algumas das ideias aqui discutidas foram já validadas com casos de estudo reais o que tem permitido à equipe avançar com mais segurança.

Podemos resumir o estado actual do RODA da seguinte forma: a sua especificação está feita, a solução tecnológica está planeada, o repositório central já está em exploração, alguns serviços (disseminadores) já estão operacionais, o *front-office* e o *back-office* finais estão em desenvolvimento, o ingestor de bases de dados está concluído (falta integrá-lo) e foi desenvolvido o primeiro protótipo de um construtor de SIP.

O que falta fazer pode resumir-se a: melhorar interfaces, criar serviços de administração e, mais importante que tudo o resto, criar as estruturas físicas e humanas para a implementação final do RODA (equipamento, recursos humanos e um plano para funcionamento em autonomia).

Uma coisa é certa, o RODA continuará a "rolar"por mais uns tempos...

Referências

[Bar06a] Francisco Barbedo. Especificação de requisitos. Technical Report 41012-005, IAN/TT, 2006.

[Bar06b] Francisco Barbedo. Taxionomias de objectos digitais a integrar no roda. Technical Report 41012-006, IAN/TT, 2006.

[ead98] Ead - encoded archival description. http://www.loc.gov/ead/, 1998.

[fSDS02] Consultative Committee for Space Data Systems. Washington: National Aeronautics and Space Administration, 2002.

[Gro05] PREMIS Working Group. Data dictionary for preservation metadata: final report of the premis working group. Technical Report Final report, OCLC Online Computer Library Center & Research Libraries Group, 2005.

[iNZ06] Project: Open Acess Repositories in New Zeland. Technical evaluation of selected open source repository solutions. Technical Report Version 1.3, Tertiary Education Commission of New Zeland, 2006.

[JLRH02] M. H. Jacinto, G. R. Librelotto, J. C. Ramalho, and P. R. Henriques. Bidirectional conversion between xml documents and relational data bases. In *7th International Conference on CSCW in Design*, Rio de Janeiro - Brasil, 2002.

[mar06] Marcxml. http://www.loc.gov/standards/marcxml, 2006.

[McD06] J. P. McDonough. Mets: standardized encoding for digital library objects. *International Journal on Digital Libraries*, 6(2), 2006.

[met06] Mets: An overview & tutorial. http://www.loc.gov/standards/mets/METSOverview.v2.html, 2006.

[nis06] Niso metadata for images in xml. http://www.loc.gov/standards/mix/, 2006.

[oA99] International Council on Archives. volume 0-9696035-6-8. International Council on Archives, 1999.

[Ruu02] Raivo Ruusalepp. Ahds preservation metadata framework. Technical report, Estonian Business Archives, Ltd, 2002.

[SSRB00] Douglas Schmidt, Michael Stal, Hans Rohnert, and Frank Buschmann. *Pattern-Oriented Software Architecture*, volume 2. John Wiley and Sons, 2000.

[TS04] K. P. Thomaz and A. J. Soares. A preservação digital e o modelo de referência open archival information system (oais). *DataGramaZero - Revista de Ciência da Informação*, 5, 2004.

[Wei97] S. Weibel. The dublin core: a simple content designation model for electronic resources. *Bulletin of the American Society for Information Science*, 24(1):9–11, 1997.