DigitArq e o novo módulo de interoperabilidade OAI-PMH

Luís Miguel Ferros, Miguel Ferreira
KEEP SOLUTIONS, LDA
DI, Universidade do Minho
4710-057 Braga, Portugal
Tel: 253604461
E-mail: {lmferros, mferreira}@keep.pt

José Carlos Ramalho
Departamento de informática
Universidade do Minho
4710-057 Braga, Portugal
Tel: 253604461
E-mail: jcr@di.uminho.pt

PALAVRAS-CHAVE: Arquivos digitais, Arquivos definitivos, interoperabilidade, metainformação, OAI-PMH, Portal Português de Arquivos, DigitArq, *Data provider*, *Service Provider*, *harvester*.

RESUMO

Ao longo dos últimos anos, tem-se assistido à proliferação de entidades detentoras que disponibilizam através da Internet partes do seu acervo em formato digital (conteúdos de arquivo, registos bibliográficos, repositórios institucionais, objectos de ensino, etc.). O principal objectivo desta actividade é permitir aos seus utentes aceder ao catálogo do arquivo a partir de qualquer parte do mundo.

Apesar da elevada quantidade de informação arquivística e conteúdos que se pode encontrar na Internet, essa informação continua dispersa, debaixo de estruturas heterogéneas (cada entidade detentora possui a sua interface de pesquisa e a sua maneira de disponibilizar a informação) e, sobretudo, com níveis de qualidade muito variáveis.

No sentido de mitigar todos estes problemas surge a Rede Portuguesa de Arquivos, uma iniciativa da Direcção-Geral de Arquivos (DGARQ, 2009) que visa, entre outros objectivos, a criação de um portal de pesquisa que servirá de ponto de acesso privilegiado a toda a informação de arquivo produzida em território nacional. Trata-se do Portal Português de Arquivos.

A mais recente versão do DigitArq, versão 3.1.8, vem agora acompanhada de um módulo designado Módulo de interoperabilidade OAI-PMH fazendo com que este software seja o primeiro produto nacional de gestão de arquivos definitivos a ser compatível com o Portal Português de Arquivos, assim como com outros agregadores de carácter internacional.

Constituído por sete módulos funcionais, o DigitArq procura ir de encontro às necessidades globais de um profissional de arquivo, permitindo realizar tarefas diversas como a descrição arquivística, gestão de projectos de digitalização, publicação de catálogo na Web, navegação e pesquisa, gestão de utilizadores e gestão de produtividade. Este software pode ser descarregado gratuitamente a partir do sítio Web http://www.digitarq.pt.

O módulo de interoperabilidade OAI-PMH foi desenvolvido tendo como principal objectivo ser 100% compatível com as directrizes de adesão ao Portal Português de Arquivos tornando, assim, as instituições que o utilizam aptas do ponto de vista tecnológico para adesão à Rede Portuguesa de Arquivos (DGARQ, 2008a, 2008b).

Ao longo deste artigo irão ser abordados em detalhe os aspectos técnicos e políticos que estão subjacentes ao desenvolvimento do Módulo de interoperabilidade OAI-PMH, bem como a sua articulação com os agregadores, dos quais o Portal Português de Arquivos é um exemplo. Começaremos por dar uma visão alargada da arquitectura de funcionamento de um agregador OAI-PMH: como funciona, qual o seu papel face às entidades detentoras, que tecnologias é que o suportam, etc. Serão também apresentados os vários cenários de exploração.

No artigo serão ainda descritas detalhadamente as características do Módulo de interoperabilidade OAI-PMH, i.e. normas utilizadas, mapeamentos em vigor, os formatos de metainformação suportados (com ênfase no subconjunto do Dublin Core (DCMI, 2009c) usado pelo OAI-PMH) e os resultados obtidos após validação da sua interface OAI-PMH recorrendo à Ferramenta de verificação de conformidade disponibilizada pela Direcção-Geral de Arquivos.

Para concluir, serão apresentadas as vantagens da implementação deste módulo em aplicações que disponibilizam conteúdos na Internet focando a questão da interoperabilidade que se torna praticamente transparente para a entidade detentora.

REDE PORTUGUESA DE ARQUIVOS

A Rede Portuguesa de Arquivos tem como objectivo promover a articulação e o inter-relacionamento entre entidades detentoras de arquivo. Esta rede corresponde a um conjunto de entidades detentoras que funcionam de modo integrado e articulado na prossecução de objectivos comuns. Tais objectivos passam pela

disponibilização, recolha e partilha de conteúdos de arquivo (DGARQ, 2008a). Realça-se aqui a importância da colaboração entre entidades aderentes, motivada por expectativas e interesses comuns.

A adesão à Rede Portuguesa de Arquivos pressupõe o cumprimento de um conjunto mínimo de requisitos:

a) *Requisitos administrativos*: as entidades aderentes devem dispor de autonomia administrativa. Caso tal não se verifique, a adesão poderá ser solicitada pela respectiva entidade de tutela, para si própria ou para uma ou várias das suas unidades orgânicas dependentes.

b) *Requisitos de acesso*: a posse de um sistema de arquivo é a condição de base que viabiliza a inclusão de uma entidade na rede. As entidades aderentes devem disponibilizar à Rede Portuguesa de Arquivos recursos de informação arquivística de acesso-livre.

c) *Requisitos técnicos*: a informação de arquivo disponibilizada à Rede Portuguesa de Arquivos deve: 1) representar convenientemente a complexidade e hierarquização da informação arquivística; 2) garantir a normalização estrutural básica da descrição da documentação de arquivo, independentemente da sua forma ou suporte; 3) permitir a interoperabilidade das descrições produzidas pelo conjunto das entidades aderentes à Rede; 4) a durabilidade dos dados contra a rápida obsolescência de software e de hardware; 5) a facilidade de armazenamento, processamento, transmissão e troca de dados arquivísticos; 6) a conversão de instrumentos de descrição não informatizados e a sua subsequente disponibilização em linha;

d) *Requisitos funcionais*: com o objectivo de assegurar o bom funcionamento do portal de pesquisa é obrigatório que as entidades aderentes cumpram os seguintes requisitos: 1) disponibilizem os seus registos de metainformação através de repositórios em acesso-livre; 2) disponham de uma ligação à internet que permita o acesso às descrições desses conteúdos; 3) implementem globalmente o protocolo OAI-PMH e os mapeamentos em vigor.

De acordo com os requisitos supracitados, as entidades detentoras de arquivo que desejarem participar na Rede Portuguesa de Arquivos e no Portal Português de Arquivos deverão cumprir os requisitos de carácter administrativo e dispor de software compatível com o protocolo OAI-PMH – *Open Archives Initiative – Protocol for Metadata Harvesting* (Open Archives Initiative, 2002). Adicionalmente, os metadados disponibilizados por estas entidades deverão estar de acordo com as directrizes definidas no âmbito do projecto sob pena de a entidade detentora ver a sua adesão à Rede rejeitada por falta de compatibilidade, e/ou qualidade e/ou completude dos seus metadados.

DIGITARQ – SOFTWARE DE GESTÃO DE ARQUIVO

O software DigitArq (Ferreira & Ramalho, 2004a, 2004b; Ramalho, Ferreira, Ferros, Lima, & Sousa, 2006), desenvolvido em conjunto pelo Arquivo Distrital do Porto, Direcção-Geral de Arquivos e Universidade do Minho, tem como objectivo a simplificação do trabalho num arquivo definitivo.

O desenvolvimento deste conjunto aplicacional permitiu disciplinar e, sobretudo, gerir o processo de produção de auxiliares de pesquisa no seio de um arquivo, bem como centralizar os resultados dessa actividade num único repositório de dados permitindo o acesso imediato e simultâneo a essa informação por parte de todos os utentes e funcionários da instituição.

A solução assenta em quatro normas internacionais:

1. ISAD(G) - International Standard Archival Description (International Council on Archives, 1999);
2. EAD - Encoded Archival Description (Library of Congress, 1998);
3. ISAAR - International Standard Archival Authorities Records (Corporate, Persons, Families) (International Council on Archives, 1995);
4. EAC - Encoded Archival Context (Encoded Archival Context Working Group, 2003)

As duas primeiras, harmonizadas desde a publicação da versão 2002 do EAD, destinam-se a suportar o processo de descrição arquivística. As restantes destinam-se a apoiar a produção de registos de autoridade.

Actualmente este software é constituído por 6 módulos funcionais que procuram ir de encontro às necessidades de um profissional de arquivo. Entre estas encontram-se a descrição arquivística, gestão de projectos de digitalização, publicação Web e a navegação e pesquisa.

PROTOCOLO OAI-PMH

O protocolo OAI-PMH foi desenvolvido pela *Open Archives Initiative*, uma entidade cujo principal objectivo é desenvolver e promover normas que facilitem a interoperabilidade e a livre circulação de metadados e conteúdos entre diferentes entidades e sistemas de informação.

O protocolo OAI-PMH define em detalhe como deve ser realizada a transferência de metadados entre duas entidades distintas: os *data providers* e os *service providers*. Os *data providers* participam na transferência fornecendo os seus metadados aos *service providers*. Estes, por sua vez, têm como principal objectivo fornecer serviços de valor acrescentado, tais como serviços de pesquisa e referência ou estatísticas tendo por base a metainformação recolhida dos *data providers*.

A interacção entre as duas entidades básicas do OAI-PMH pode ser vista na Figura 1. Pode-se observar que um *service provider* que deseja realizar uma recolha de metadados envia um pedido OAI-PMH (via HTTP) a um *data provider* que responderá com os metadados solicitados em formato XML (segundo o *schema* do OAI-PMH). Com base nos metadados recolhidos e armazenados centralmente, o *service provider* pode disponibilizar serviços de valor acrescentado como, por

exemplo, um sistema de pesquisa, estatísticas diversas, relatórios.

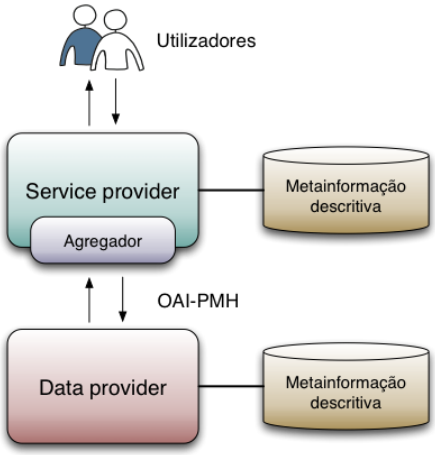

Figura 1 – Interacção entre as duas entidades do OAI-PMH

Definições e conceitos

Nesta secção são descritas as principais definições e conceitos do protocolo.

Agregador

Um agregador, ou *harvester*, é uma aplicação cliente que envia pedidos OAI-PMH a um *data provider* e que tem como objectivo recolher os seus metadados. Um *harvester* é geralmente parte integrante de um *service provider*.

Repositório

Um repositório, ou *repository*, é um servidor acessível através da Internet que pode processar os 6 pedidos OAI-PMH. Um repositório é gerido por um *data provider* e tem como missão expor metadados para serem recolhidos por *harvesters*. Podem ser distinguidas 3 entidades distintas relacionadas com os metadados acessíveis através de OAI-PMH:

Recurso – trata-se do objecto ou "coisa" à qual os metadados dizem respeito. A sua natureza, suporte físico ou forma estão fora do âmbito do OAI-PMH.

Item – Um item trata-se de um constituinte de um repositório ao qual é possível solicitar metadados. Um item descreve um recurso.

Registo – Um registo trata-se de um item materializado num formato de metadados específico compatível com o protocolo OAI-PMH. Um registo é retornado na resposta (codificada em XML) a um pedido de recolha de metadados para um item específico do repositório, item esse que descreve um determinado recurso.

Identificador único

Um identificador único identifica de forma unívoca um item no interior de um repositório. O identificador único é usado nos pedidos OAI-PMH para extrair metainformação de um item.

Set

Um *set* é um elemento opcional que permite agrupar itens do repositório e possibilitar a realização de recolhas selectivas de registos de metainformação. Os repositórios podem organizar os seus itens em diferentes *sets* (i.e. conjuntos) de acordo com os critérios que melhor entenderem, e.g. pode criar-se um *set* que contém os registos de uma determinada colecção, um *set* que identifica os registos em acesso-livre, ou até um *set* para distinguir os itens do repositório que possuem valor patrimonial daqueles que são puramente administrativos.

Pedidos e respostas do protocolo OAI-PMH

Os pedidos OAI-PMH são expressos sob a forma de pedidos HTTP, i.e. estes devem ser submetidos usando um dos métodos GET ou POST. O método POST tem a vantagem de não impor limitações ao comprimento dos argumentos, no entanto, os repositórios que implementem o protocolo OAI-PMH devem suportar ambos os métodos.

As interfaces OAI-PMH são acessíveis através de um endereço base sobre o qual todos os pedidos são efectuados. Esse endereço possui a forma: http://host:port/path

Os pedidos OAI-PMH são definidos através de parâmetros passados ao endereço base sob a forma de "parâmetro=valor". Os parâmetros podem aparecer por qualquer ordem e múltiplos parâmetros são separados pelo carácter "&". Cada pedido OAI-PMH deve ter pelo menos um par "parâmetro=valor" que especifica o pedido OAI-PMH enviado pelo agregador: verb=**valor**, onde **valor** identifica o pedido OAI-PMH. Os valores possíveis são: Identify, ListMetadataFormats, ListSets, GetRecord, ListRecords e ListIdentifiers.

As respostas aos pedidos são formatadas como respostas HTTP, com os campos de cabeçalho HTTP apropriados.

Descrevem-se de seguida os pedidos, ou *verbs*, definidos no âmbito do protocolo OAI-PMH.

Identify

Este *verb* é utilizado para recuperar informação administrativa e técnica sobre um repositório, e.g. nome, identificador, e-mail do administrador, informações sobre a propriedade intelectual dos dados contidos no repositório, etc.

ListMetadataFormats (identifier)

Este pedido é utilizado para obter os formatos de metainformação suportados pelo repositório. É obrigatória a implementação de, pelo menos, o formato de metainformação Dublin Core.

ListSets

Utilizado para recuperar a lista de *sets* suportados pelo repositório.

GetRecord (identifier, metadataPrefix)

Este *verb* é utilizado para recuperar um registo específico de metainformação existente no repositório. Os argumentos devem especificar o identificador do item e o formato da metainformação que se pretende obter.

ListRecords (metadataPrefix, from, until, set)

Este *verb* é utilizado para recolher registos de um repositório.

ListIdentifiers (metadataPrefix, from, until, set)

Trata-se de uma forma abreviada de ListRecords e recupera apenas cabeçalhos dos registos de metainformação.

Recolhas selectivas

A recolha selectiva permite aos agregadores limitar a recolha de registos a apenas subconjuntos de toda a metainformação disponível no repositório. O OAI-PMH suporta recolhas selectivas através de dois tipos de critérios que podem ser combinados num mesmo pedido: *datestamps* e *sets*.

Datestamps - os agregadores podem usar *datestamps* para recolher apenas os registos que foram criados, eliminados ou alterados dentro do intervalo de datas específico. Os *datestamps* podem ser incluídos como valores dos argumentos opcionais, *from* e *until*, nos pedidos *ListRecords* e *ListIdentifiers*.

from – registos com data maior ou igual que

until – registos com data menor ou igual que

O valor do argumento *from* deve ser menor ou igual que o valor do argumento *until*, caso contrário o repositório deve devolver um erro do tipo *badArgument*.

Sets - os agregadores podem especificar um *set* como critério para recolha selectiva. Para especificar um *set* basta definir o valor do parâmetro *setSpec* nos pedidos *ListRecords* e *ListIdentifiers*.

NORMA EAD (*Encoded Archival Description*)

O EAD é uma norma que permite codificar auxiliares de pesquisa em XML (Library of Congress, 1998). Esta norma de metainformação permite descrever os objectos custodiados num arquivo de forma hierárquica e contextualizada, ajudando os seus potenciais consumidores a categorizar e localizar a informação pretendida.

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.

Cada instância de um EAD contém 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 com respectivos subníveis: fundo, secção, série, unidade de instalação, documento composto, documento simples[1].

Cada nível de descrição contém informação descritiva adequada à unidade orgânica ou documental à qual dizem respeito (ICA, 2008). Como exemplos deste tipo de informação podemos realçar o título, datas extremas, história biográfica, história custodial, âmbito e conteúdo, existência e localização dos originais e cópias, etc.

A norma EAD é flexível permitindo várias opções e soluções alternativas relativamente aos seus múltiplos elementos. Para mais informações sobre o esquema EAD, é possível consultar as seguintes fontes de informação:

- Official EAD Version 2002 Web Site (Congress, 2009)
- Society of American Archivists (SAA, 2009b)
- RLG Best Practices Guidelines for Encoded Archival Description (RLG, 2002)
- EAD Tools Survey (SAA, 2009a)

FORMATO DC (Dublin Core)

O Dublin Core (DCMI, 2009c) é um formato de metadados que visa descrever recursos electrónicos, sejam eles textos, vídeos, imagens, sons, bases de dados ou websites. As características mais relevantes deste formato são a sua simplicidade, interoperabilidade semântica, a sua modularidade/extensibilidade e o consenso internacional que se gerou em torno deste formato no que toca à sua capacidade de garantir interoperabilidade entre sistemas de metainformação.

Elemento	Descrição
Title	Nome pelo qual o recurso é conhecido.
Creator	Entidade (individuo ou instituição) responsável pela criação do recurso.
Subject	Tópicos sobre o conteúdo do recurso.
Description	Descrição do conteúdo do recurso.
Publisher	Entidade responsável por tornar o recurso acessível.
Contributor	Entidade responsável por qualquer contribuição para o conteúdo do recurso.
Date	Data associada a um evento do ciclo de vida do recurso.
Type	Natureza do conteúdo do recurso.
Format	Manifestação física ou digital do recurso. Pode incluir a identificação das aplicações ou equipamento necessários à utilização do recurso bem como as dimensões (tamanho e duração) do mesmo.
Identifier	Código de referência do recurso. Geralmente baseado num sistema de identificação formal, como URL, DOI ou ISBN.
Source	Referência para o recurso de onde o recurso descrito possa ter derivado.
Language	Idioma do conteúdo intelectual do recurso.
Relation	Referência a um recurso relacionado.
Coverage	Extensão ou alcance do recurso.
Rights	Informação sobre os direitos associados ao recurso descrito (direitos de autor, de propriedade intelectual ou outros).

Tabela 1 – Elementos do Dublin Core simples

A OCLC - *Online Computer Library Center* (OCLC, 2009), em 1995, liderou a primeira de várias reuniões com a Biblioteca do Congresso, Universidades e Organizações Não Governamentais, que deram origem a este formato. Esta reunião realizada em Dublin, Ohio,

[1] Os níveis de descrição supracitados são prescritos pela Direcção-Geral de Arquivos. Poderão existir outros níveis de descrição, porém estes são aqueles que serão suportados pelo Portal Português de Arquivos.

gerou a designação Dublin Core. Como resultado dos trabalhos levados a cabo, foram definidos um conjunto mínimo de elementos para a identificação e descrição de recursos[2].

Actualmente, o formato é mantido pela DCMI – *Dublin Core Metadata Initiative* (DCMI, 2009a), que visa desenvolver normas que fomentem a interoperabilidade entre diversos sistemas, facilitando a recuperação, partilha e gestão da informação (DCMI, 2009a).

Actualmente o Dublin Core possui dois níveis de especificação: o simples e o qualificado. O simples é constituído por um conjunto de 15 elementos (ver Tabela 1) conforme o DCMES – *Dublin Core Metadata Element Set* (DCMI, 2009b) e o qualificado inclui alguns elementos adicionais e um conjunto de "qualificadores" que permitem refinar a semântica dos elementos anteriormente existentes. Todos os elementos são opcionais e repetíveis.

EAD VS DUBLIN CORE NO CONTEXTO DO OAI-PMH

O formato EAD é o formato por eleição para representar a natureza complexa e hierárquica da informação de arquivo. Os registos descritivos encontram-se organizados segundo uma estrutura hierárquica onde é efectuada uma descrição do mais geral (i.e. fundo ou entidade produtora de informação) para o mais particular (documento ou item de informação indivisível). Um determinado registo de descrição só tem sentido quando enquadrado no respectivo contexto hierárquico. A sua descrição é complementada por todos os registos que o ascendem na hierarquia, i.e. até ao nível de descrição fundo.

Uma vez que um registo do EAD só está totalmente descrito quando se considera toda a estrutura orgânica ao qual este pertence, surge de imediato uma questão - como efectuar a recolha de registos EAD através do protocolo OAI-PMH?

Este problema pode ser resolvido de 3 formas distintas (Figura 2): (1) recolher o fundo completo ao qual pertence esse registo, (2) recolher o ramo da árvore desde o fundo até ao registo alterado, ou (3) recolher apenas o registo alterado (Ferros, Ramalho, & Ferreira, 2008).

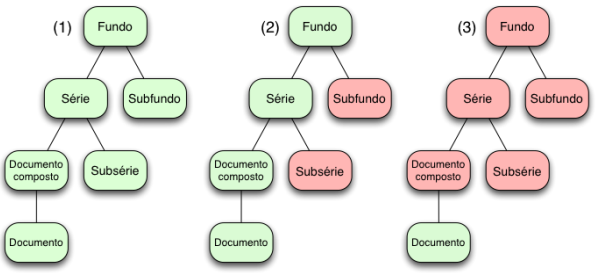

Figura 2 - Tipos de recolha de registos

Recolher o fundo completo ao qual pertence esse registo

Esta é a solução mais simples de implementar uma vez que para dar resposta a um pedido OAI-PMH o *data provider* apenas tem de enviar indiscriminadamente todos os registos de um fundo. O *service provider* tem de integrar o fundo recebido no repositório central, algo que se resume à substituição da versão anterior do mesmo pela versão mais recente.

Como se depreende facilmente, esta é uma solução extremamente ineficiente, pois uma simples alteração ou inserção de um registo no repositório iria desencadear uma posterior recolha de um fundo completo. Desta forma, alterações efectuadas a registos de diversos fundos, os quais podem ter um número elevado de registos (na ordem das centenas de milhar), iria desencadear uma transferência de dados que representaria um substancial desperdício de largura de banda.

Recolher o ramo da árvore do fundo até ao registo alterado

Este método desencadeia uma transferência de dados bastante inferior ao descrito anteriormente. No entanto as operações de extracção dos registos a partir dos *data providers* e a integração dos mesmos nos *service providers* é consideravelmente mais complexa. A extracção passa por seleccionar os registos ascendentes do registo alterado. A integração no *service provider* consiste na substituição dos registos antigos pelos novos registos recebidos.

Recolher apenas o registo alterado

Este será certamente o método mais eficiente, uma vez que apenas os registos novos ou alterados são recolhidos independentemente da sua posição na estrutura do fundo. O problema deste método reside no facto de um ficheiro EAD não ser considerado válido quando não existe o nível de descrição raiz, i.e. o fundo. O *schema* do EAD, contudo, não verifica este invariante durante a validação de uma instância EAD. Assim, pelo uso de identificadores persistentes, ou garantindo-se a unicidade dos identificadores dos registos, a recolha de registos isolados do fundo respectivo pode ser concretizável. Pela análise do EAD, verifica-se a existência dos campos *CountryCode* (código do país) e *RepositoryCode* (código do repositório) os quais fazem parte do código de referência de um registo. Desta forma, a existência destes campos garante a unicidade dos códigos de referência que identificam um registo no contexto de um repositório.

Assim, a tarefa de agregação de apenas um registo (ou um conjunto de registos), independentemente da organização hierárquica pode ser possível, atendendo a que, como já foi referido, poderem ser enviados EAD não válidos do ponto de vista semântico, apesar de sintacticamente válidos de acordo com o *schema* do EAD.

Recolha de metainformação em formato Dublin Core

Para potenciar a interoperabilidade entre repositórios, estes devem disponibilizar informação noutros formatos para além do EAD, e.g. em formato Dublin Core (DCMI, 2009c). O Dublin Core é actualmente o formato de metainformação mais utilizado em contextos onde a interoperabilidade de metainformação descritiva é um requisito essencial.

Assim, apesar dos intervenientes na Rede Portuguesa de

[2] Segundo a norma Dublin Core, um recurso define-se por algo digital ou material que possui um identificador único.

Arquivos corresponderem a entidades detentoras de arquivo que favorecem naturalmente o EAD, estes terão obrigatoriamente de disponibilizar os seus registos também em formato Dublin Core. Torna-se portanto necessário definir e implementar um mapeamento entre EAD e Dublin Core (i.e. *crosswalk*). O Portal Português de Arquivos irá oferecer serviços de localização e pesquisa a partir dos dados recolhidos dos vários *data providers* aderentes à rede, desde que estes sejam fornecidos em formato Dublin Core. Contudo a utilização deste formato pode revelar dificuldades ao nível da inexistência de elementos que são relevantes para a descrição arquivística. Além disso não é totalmente congruente com a realidade da descrição arquivística pois não possibilita uma representação hierárquica de descrições de arquivo.

De forma a garantir uma correcta implementação do protocolo e a compatibilidade semântica entre as descrições mantidas pelas entidades aderentes ao Portal Português de Arquivos e as funcionalidades do mesmo, apresenta-se na Tabela 2 o mapeamento entre o formato de metainformação EAD e o formato DC.

Descrição	EAD	DC
Resumo	Abstract	dc.description
Condições de acesso	AccessRestrict	dc.rights
Ingressos adicionais	Accruals	-
Modalidades de aquisição	AcqInfo	-
Existência de cópias	AltFormAvail	dc.relation.isformatof
Avaliação, selecção e eliminação	Appraisal	-
Organização e ordenação	Arrangement	-
Bibliografia	Bibliography	dc.source
História administrativa/biográfica	BiogHist	dc.description
Código do país	CountryCode	europeana.country
História custodial	CustodHist	dc.provenance
Dimensão e suporte	Dimensions	dc.format.extent
Tipologia	GenreForm	dc.format.medium
Localidade	GeogName	-
Idioma/Escrita	LangMaterial	dc.language
Estatuto legal	LegalStatus	-
Detalhes específicos dos materiais	MaterialSpec	-
Notas/observações	Note	-
Localização de originais	OriginalsLoc	dc.relation.isformatof
Autores/produtores	Origination	dc.creator
Instrumentos de pesquisa	OtherFindAid	-
Nível de descrição	OtherLevel	dc.type
Aspecto físico	PhysFacet	-
Localização física	PhysLoc	-
Características físicas e requisitos técnicos	PhysTech	-
Citação	PreferCite	-
Informações do processo	ProcessInfo	-
Materiais relacionados	RelatedMaterial	dc.relation.refences
Entidade detentora	Repository	dc.publisher
Código do repositório	RepositoryCode	-
Âmbito e conteúdo	ScopeContent	dc.subject
Material separado	SeparatedMaterial	dc.relation.hasPart
Datas	UnitDate	dc.date
Referência	Unitid	dc.identifier
Título do document	UnitTitle	dc.title
Tipo título	UnitTitleType	-
Condições de reprodução	UseRestrict	dc.rights
Plano de classificação	FilePlan	-

Tabela 2 – Mapeamento de EAD para DC

O mapeamento descrito na Tabela 2 revelou dificuldades ao nível da inexistência por parte do DC de elementos que são relevantes para a descrição arquivística. Para além desta constatação, verificou-se que em determinados casos seria necessário utilizar um elemento DC para representar dois ou mais elementos EAD, o que comprometia a coerência e clareza da descrição arquivística. Além disso, a necessidade de usar elementos do DC qualificado, os quais não são normalmente utilizados pelo protocolo OAI-PMH, torna o problema ainda mais crítico.

Após análise dos objectivos do PPA, concluiu-se que não havia necessidade de recolher todos os elementos descritivos existentes no EAD ou no Dublin Core qualificado, pois esse processo iria dificultar a adesão à Rede Portuguesa de Arquivos por parte de entidades detentoras que não implementassem na sua totalidade qualquer um destes esquemas.

Adicionalmente, uma vez que o Portal Português de Arquivos não tem como objectivo substituir-se aos repositórios das entidades detentoras (muito pelo contrário), mas sim canalizar potenciais utilizadores para os repositórios que detêm a informação desejada, concluiu-se que a melhor estratégia seria identificar o conjunto mínimo de atributos que fosse suficiente para garantir este pressuposto.

Esses elementos foram:

- Código de referência
- Título
- Datas extremas
- Nível de descrição
- Entidade detentora
- Dimensão e suporte
- Âmbito e conteúdo

Este conjunto de elementos são suficientes para dar ao utilizador uma percepção dos registos que lhe interessam, podendo navegar directamente para o registo existente no repositório da entidade detentora e visualizar toda a restante informação existente, bem como compreender o contexto no qual o registo se situa.

Como todos estes elementos são directamente transponíveis para campos do DC simplificado, não foi necessário mapear todos os campos do EAD para DC (Tabela 3).

Descrição	EAD	DC	Obrigatório
Código de referência	Unitid	dc.identifier	Sim
Título	UnitTitle	dc.title	Sim
Datas extremas	UnitDate	dc.date	Sim
Nível de descrição	OtherLevel	dc.type	Sim
Entidade detentora	Repository	dc.publisher	Sim
Dimensão e suporte	Dimensions	dc.format	Não
Âmbito e conteúdo	ScopeContent	dc.subject	Não
-	-	dc.identifier[3]	Sim
-	-	dc.relation[4]	Não

Tabela 3 – Mapeamento entre EAD para DC simplificado

REGRAS DE CONFORMIDADE

No âmbito do projecto Portal Português de Arquivos foi desenvolvida uma ferramenta[5] que tem como missão aferir o nível de conformidade de um *data provider* com as directrizes e requisitos definidos pelo portal.

Durante o processo de verificação de conformidade são

3 URL do registo de descrição no *data provider*.

4 URL da miniatura do objecto digital associado ao registo de descrição (caso exista).

5 A ferramenta de verificação de conformidade encontra-se em http://www.arquivos.pt

analisadas as regras apresentadas na Tabela 4 (problemas graves) e na Tabela 5 (problemas ligeiros). A não conformidade com alguma destas regras irá ser devidamente assinalada no relatório de conformidade que é produzido por esta ferramenta. Um repositório para poder aderir à Rede Portuguesa de Arquivos deverá apresentar zero problemas graves sob pena de ver o seu pedido de adesão rejeitado.

Regras	Elemento	Notas
Registo possui identificador	dc.identifier	O identificador é um elemento obrigatório
Registo possui identificador válido	dc.identifier	O identificador deverá iniciar-se com o código do país, seguido do código da entidade detentora, e.g. PT/TT, PT/ADPRT. No caso de identificadores baseados em URL, então estes têm de ser URL válidos.
Registo possui título	dc.title	O título é um elemento obrigatório
Registo possui datas extremas	dc.date	As datas extremas são elementos obrigatórios
Registo possui datas extremas válidas	dc.date	As datas extremas deverão apresentar-se em dois elementos dc.date ou então num único elemento dc.date sob a forma YYYY-MM-DD/YYYY-MM-DD. Todas as datas deverão estar formatadas na norma ISO 8601 sem as componentes relativas à hora. Exemplos de datas válidas são: 1900, 1900-12, 1960-12-01, 1930-03-21/2001-01-03, 1900/2001-02-03. Elementos de data desconhecidos deverão ser assinalados com zeros, e.g. 1900-00-23, para uma data cujo mês é desconhecido.
Registo possui nível de descrição	dc.type	O elemento dc.type é obrigatório.

Tabela 4 – Regras de verificação de problemas graves

Para além dos problemas graves anteriormente descritos, a ferramenta de verificação de conformidade analisa alguns problemas nos metadados do repositório que, apesar de não serem impeditivos para o funcionamento do Portal Português de Arquivos, diminuem a experiência do utilizador durante a sua utilização.

Regras	Elemento	Notas
Registo apresenta nível de descrição, porém este não é conhecido pelo portal	dc.type	Os valores do elemento dc.type deverão pertencer ao vocabulário controlado: Fundo, Subfundo, Subsubfundo, Secção, Subsecção, Subsubsecção, Série, Subsérie, Subsubsérie, Unidade de instalação, Documento composto, Documento simples
Registo apresenta o elemento idioma, porém este não está de acordo com a norma ISO 639-3:2007	dc.language	Os valores do elemento dc.language devem respeitar a norma ISO 639-3:2007.

Tabela 5 – Regras de verificação de problemas ligeiros.

TRABALHO FUTURO

Um dos objectivos da RPA é a sua integração com outros *service providers* de âmbito internacional, e.g. Europeana (Europeana, 2009), APEnet (APEnet, 2009), entre outros. Esta integração dará ainda mais visibilidade às entidades aderentes, uma vez que os seus metadados passarão a estar acessíveis a uma comunidade de utilizadores que se estende para além das fronteiras nacionais.

Para concretizar esse objectivo, o PPA, para além de assumir a qualidade de *service provider*, terá também de servir a função de *data provider,* disponibilizando a informação que recolheu das várias entidades aderentes a *service providers* geridos por terceiros. A Figura 3 exemplifica a integração do PPA com a Europeana. A recolha de dados do por parte desta é também realizada através do protocolo OAI-PMH.

Para proteger os direitos associados à informação, as entidades aderentes terão a possibilidade de controlar que informações desejam disponibilizar a terceiros, mesmo no caso de a informação já estar a ser partilhada com o Portal Português de Arquivos.

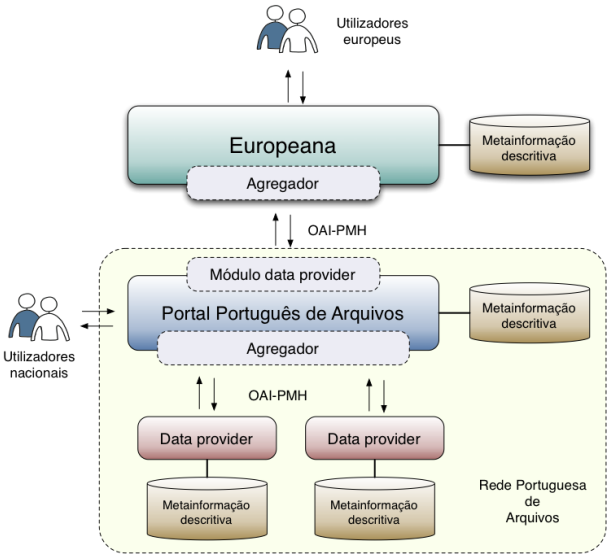

Figura 3 – Integração do PPA com a Europeana

CONCLUSÕES

Ao longo deste artigo foram abordados os principais aspectos técnicos e políticos que estiveram subjacentes ao desenvolvimento do Módulo de interoperabilidade OAI-PMH do DigitArq. Começou-se por contextualizar o seu principal cenário de actuação, a Rede Portuguesa de Arquivos, onde se descreveram os principais objectivos e requisitos de adesão.

Seguiu-se uma breve descrição do projecto DigitArq, do protocolo de comunicação OAI-PMH e dos formatos de metainformação EAD e DC para transferência de informação entre as entidades que participam na rede.

Dentro destes formatos de metainformação, o EAD é, normalmente, o formato utilizado para representar a informação de arquivo. Este formato permite descrever informação segundo uma estrutura hierárquica que, apesar de ser a mais adequada para representar a informação de arquivo, apresenta uma complexidade e variações na sua utilização que podem impossibilitar a transferência e utilização de informação por parte de um *service provider*.

Por outro lado, o Dublin Core é mais adequado para

transferir metadados entre duas entidades por ser simples, semanticamente interoperável e por ser considerado internacionalmente como a língua-franca entre sistemas que trocam metainformação descritiva. Contudo, a sua estrutura não é suficientemente rica para captar integralmente a metainformação descritiva de um arquivo.

Apesar do estudo aprofundado que foi realizado sobre as diferentes alternativas de recolha de metainformação em formato EAD utilizando o protocolo OAI-PMH, concluiu-se da não necessidade de recolher todos os elementos descritivos presentes neste formato, uma vez que a eficácia e os propósitos do PPA não seriam comprometidos. O portal não pretende substituir-se aos repositórios aderentes, funcionando apenas como uma plataforma que direcciona os utilizadores para os vários repositórios aderentes.

O artigo abordou também as regras de conformidade que foram definidas no âmbito do projecto Portal Português de Arquivos. Tais regras são utilizadas para verificar o nível de conformidade de um *data provider* com as directrizes e requisitos definidos pelo portal.

REFERÊNCIAS

APEnet (2009). APEnet web site Retrieved 2009-10-19, from http://www.apenet.eu

Congress, L. o. (2009). EAD - Encoded Archival Description Retrieved 2009-10-12, from http://www.loc.gov/ead/

DCMI (2009a). About the Initiative Retrieved 2009-08-21, from http://dublincore.org/about/

DCMI (2009b). Dublin Core Metadata Element Set, Version 1.1 Retrieved 2009-09-02, from http://dublincore.org/documents/dces/

DCMI (2009c). The Dublin Core Metadata Initiative Retrieved 2009-08-21, from http://dublincore.org/

DGARQ (2008a). *Rede Portuguesa de Arquivos (RPA): fundamentos para o seu desenvolvimento e gestão. Módulo 1: Modelo Conceptual*.

DGARQ (2008b). *Rede Portuguesa de Arquivos (RPA): fundamentos para o seu desenvolvimento e gestão. Módulo 1: Modelo Lógico*.

DGARQ (2009). Direcção-Geral de Arquivos, from http://www.dgarq.gov.pt/

Encoded Archival Context Working Group (2003). Encoded Archival Context (EAC) Retrieved 2004-12-12, from http://www.library.yale.edu/eac/

Europeana (2009). Europena - Homepage Retrieved 2009-10-19, from http://www.europeana.eu/portal/

Ferreira, M., & Ramalho, J. C. (2004a). *DigitArq - Creating and Managing a Digital Archive.* Paper presented at the ICCC/IFIP International Conference on Electronic Publishing, Brasília, Brazil.

Ferreira, M., & Ramalho, J. C. (2004b). *DigitArq: Creating a Historical Digital Archive.* Paper presented at the 5ª Conferência da Associação Portuguesa de Sistemas de Informação, Lisboa.

Ferros, L., Ramalho, J. C., & Ferreira, M. (2008). *Creating a National Federation of Archives using OAI-PMH.* Paper presented at the XATA - XML, Aplicações e Tecnologias Associadas, Évora, Portugal.

ICA (2008). ISAD(G): General International Standard Archival Description, Second edition 2nd edition. Retrieved 2009-04-05, from http://www.ica.org/en/node/30000

International Council on Archives (1995). *ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, Persons, and Families*: International Council on Archives.

International Council on Archives (1999). *ISAD(G): General International Standard Archival Description, Second edition* (No. 0-9696035-6-8): International Council on Archives.

Library of Congress (1998). EAD - Encoded Archival Description Retrieved 2008-04-21, from http://www.loc.gov/ead/

OCLC (2009). Online Computer Library Center Retrieved 2009-09-01, from http://www.oclc.org

Open Archives Initiative (2002). The Open Archives Initiative Protocol for Metadata Harvesting Retrieved 2009-10-20, from http://www.openarchives.org/OAI/openarchivesprotocol.html

Ramalho, J. C., Ferreira, M., Ferros, L., Lima, M. J. P., & Sousa, A. (2006). *DigitArq 2 - Nova arquitectura aplicacional para gestão de Arquivos Definitivos.* Paper presented at the 2nd International Conference on Enterprise Archives, Seixal, Portugal.

RLG (2002). RLG Best Practices Guidelines for Encoded. Retrieved from http://www.oclc.org/programs/ourwork/past/ead/bpg.pdf

SAA (2009a). EAD Help Pages Retrieved 2009-09-01, from http://www.archivists.org/saagroups/ead/

SAA (2009b). Society of American Archivists Retrieved 2009-09-01, from http://www.archivists.org/