Luís Miguel Sá Neiva Ferros

Extracção e concentração de metainformação distribuída por vários repositórios

Novembro de 2009

Universidade do Minho
Escola de Engenharia

Luís Miguel Sá Neiva Ferros **Extracção e concentração de metainformação distribuída por vários repositórios**

UMinho | 2009

Universidade do Minho
Escola de Engenharia

Luís Miguel Sá Neiva Ferros

Extracção e concentração de metainformação distribuída por vários repositórios

Dissertação de Mestrado em Sistemas de Dados e Processamento Analítico

Trabalho efectuado sob a orientação do
Professor Doutor José Carlos Leite Ramalho

Novembro de 2009

A todos aqueles que ao longo do meu percurso me têm apoiado e enriquecido.

Agradecimentos

A dedicação, motivação e persistência necessários para o desenvolvimento desta dissertação foram conseguidos com a contribuição directa ou indirecta de várias pessoas. A todas estas pessoas gostaria de exprimir a minha gratidão, agradecendo-lhes todo o apoio concedido ao longo deste meu percurso.

Começo por agradecer ao meu orientador, o Professor Doutor José Carlos Ramalho. Ao longo dos últimos anos ajudou-me a traçar alguns dos meus percursos académicos e profissionais. Este meu percurso, em particular durante a elaboração desta dissertação foi conseguido graças à sua orientação, partilha de ideias, estímulo e disponibilidade.

Agradeço também de forma especial, ao Doutor Miguel Ferreira. A partilha incondicional das suas experiências e opiniões ao longo dos últimos anos permitiram encarar e resolver muitos dos problemas que tive que ultrapassar, alguns deles relacionados com esta dissertação.

Agradeço ao corpo docente do Curso de Mestrado em Sistemas de Dados e Processamento Analítico, pelos conhecimentos que me transmitiram durante a exposição das matérias. Algum do conhecimento aqui assimilado influenciou positivamente o resultado final desta dissertação.

Agradeço a todos os colegas de mestrado, que pelas suas diferentes experiências profissionais, permitiram ver e resolver alguns problemas segundo diferentes perspectivas.

Finalmente, agradeço aos meus pais que sempre me apoiaram e incentivaram nos momentos mais difíceis.

Resumo

Extracção e concentração de metainformação distribuída por vários repositórios

Ao longo dos últimos anos, tem-se assistido à proliferação de entidades detentoras de arquivo que disponibilizam através da Internet partes do seu acervo em formato digital. O âmbito principal desta actividade é permitir aos seus utentes aceder ao catálogo do arquivo a partir de qualquer parte do mundo. No entanto, os repositórios dessas entidades detentoras estão geograficamente separados e apresentam diferentes interfaces de pesquisa com estruturas de dados diversas. Estas características implicam um dispendioso esforço por parte do utilizador para a pesquisa e recuperação de informação. Do ponto de vista do utilizador, seria vantajoso e confortável poder interagir com uma única interface de pesquisa, para a partir dela, submeter os parâmetros a pesquisar e receber a informação de forma consolidada, vinda de diferentes fontes de informação.

Visando atender a este propósito, os projectos em torno dos repositórios digitais deixam de ser idealizados somente a partir da concepção de sistemas de informação simples e isolados, tornando-se uma necessidade o desenvolvimento de sistemas de informação estruturados, baseados na integração de dados provenientes de diferentes sistemas.

A construção destes sistemas visa promover uma melhor articulação e inter-relacionamento entre entidades detentoras de um sistema de informação. Estes sistemas designam-se por sistemas de informação federados ou rede de repositórios digitais. No contexto desta dissertação é abordada a implementação da Rede Portuguesa de Arquivos ou, dito de outra forma, uma rede dos conteúdos detidos e disponibilizados pelas entidades de arquivo. Uma rede de arquivos 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 dos conteúdos de arquivo.

Nesta dissertação será proposto um modelo de arquitectura que permite agregar e disponibilizar toda a informação da rede a partir de um único ponto de acesso - o Portal Português de Arquivos. Trata-se de uma iniciativa da Direcção-Geral de Arquivos que visa a criação de um portal de pesquisa e ponto de acesso privilegiado a toda a informação de arquivo custodiada em território nacional.

O Portal Português de Arquivos também será a ligação nacional a projectos de carácter internacional como a Europeana, um portal de acesso à produção histórica e cultural da comunidade europeia. A adesão ao Portal Português de Arquivos e consequentemente à Europeana trará inúmeros benefícios para as entidades detentoras aderentes, nomeadamente, uma maior visibilidade dos seus acervos potenciando a sua procura e eventualmente um aumento da receita através dos serviços prestados pelo arquivo.

Abstract

Extraction and aggregation of metadata distributed across multiple repositories

Over the last few years, we have seen the proliferation of archives that provide access to archival resources in digital formats over the Internet. The main goal of this activity is to enable archive catalog access to users from anywhere in the world. But repositories that hold such content are autonomous, geographically dispersed and implement different search interfaces with a variety of data structures. These characteristics demand a greater user effort for information search and retrieval. From the user point view, it would be beneficial and comfortable to interact with a single search interface, submit search parameters, and receive information in a consolidated form.

In order to achieve such a goal, the projects around digital repositories should not be idealized as simple and isolated information systems. There is a need for structured information systems that fully support the integration of data originating from different systems.

The development of these systems promotes a better coordination and inter-relationship between distinct entities. This kind of system are designated a federated information systems or a digital repository networks. In this dissertation is exploited the implementation of a Portuguese Archives Network, i.e. a network of content, owned and provided by a set of archival entities that work together in an integrated and coordinated mode on pursuit of common goals. These goals consist in providing, harvesting and sharing archival metadata and content.

The current dissertation proposes an architecture to provide access to network harvested information from a single access point - the Portuguese Archives Portal - an initiative of the

Directorate-General of Portuguese Archives that seeks to create a search portal and a privileged access point to all archival resources available on national territory.

The Portuguese Archives Portal will also be the national gateway for international projects such as the Europeana, a web site that collects and provides access to the historical and cultural production of the European community. Adherence to Portal of Portuguese Archives and consequently to the Europeana will bring numerous benefits to the holding member entity, specifically, greater visibility to its collection, search enhancing and eventually an increase in revenue through the archive provided services.

Conteúdo

Lista de tabelas

Lista de figuras

Glossário

Arquivo – Organização responsável por gerir, descrever, armazenar e garantir o acesso à informação.

Base de dados – É um sistema cujo objectivo é registar, actualizar, manter e disponibilizar informação relevante para a actividade de uma organização.

Custódia – A responsabilidade pela conservação de documentos de arquivo, baseada na sua guarda física. A custódia nem sempre implica a propriedade legal ou o direito de controlar o acesso aos documentos.

Descrição arquivística – Processo que consiste em identificar e explicar o contexto e o conteúdo da documentação de arquivo.

Digitalização – Processo responsável pela transformação de informação analógica em informação digital.

Documento de arquivo – Informação de qualquer tipo, registada em qualquer suporte, produzida ou recebida e conservada por uma instituição ou pessoa no exercício das suas competências, ou actividades.

Fundo – Conjunto de documentos de arquivo, independentemente da sua forma ou suporte, organicamente produzido e/ou acumulado e utilizado por uma pessoa singular, família ou pessoa colectiva, no decurso das suas actividades e funções.

Internet – Rede global de comunicação baseada no protocolo TCP/IP.

Metainformação – Informação utilizada para descrever um determinado objecto ou recurso.

Nível de descrição – Posição de uma unidade de descrição na hierarquia de um fundo (ver fundo).

Norma de descrição arquivística – Estabelece orientações gerais para a descrição arquivística (ver descrição arquivística).

Pesquisa federada – Consiste na submissão de uma consulta a várias bases de dados em simultâneo.

Protocolo – Descreve orientações gerais para estabelecer acordos entre entidades ou serviços.

Rede – Estrutura de entidades ligadas por interesses comuns, com o objectivo de partilhar recursos e de realizar acções comuns das quais advenha proveito mútuo.

Repositório digital – Sistema de informação responsável por gerir e armazenar informação digital.

Suporte – Material sobre o qual a informação é registada.

Unidade de descrição – Documento ou conjunto de documentos, sob qualquer forma física, tratado como um todo e que, como tal, serve de base a uma descrição singular.

Siglas e acrónimos

BD	Base de Dados
DC	*Dubin Core*
DCMES	*Dublin Core Metadata Element Set*
DCMI	*Dublin Core Metadata Initiative*
DW	*Data Warehouse*
EAD	*Encoded Archival Description*
ESE	*Europeana Semantic Elements*
HTTP	*Hypertext Transfer Protocol*
ICA	*International Council on Archives*
ISAD(G)	*General International Standard Archival Description*
ISO	*International Organization for Standardization*
OAI	*Open Archives Initiative*
OAI-PMH	*Open Archives Initiative – Protocol Metadata Harvesting*
OCLC	*Online Computer Library Center*
PDF	*Portable Document Format*
PPA	Portal Português de Arquivos

RPA	Rede Portuguesa de Arquivos
SOA	*Service Oriented Architecture*
SOAP	*Simple Object Access Protocol*
SRU	*Search/Retrieve via URL*
SRW	*Search/Retrieve Web Service*
URL	*Uniform Resource Locator*
XML	*eXtensible Markup Language*

Capítulo 1

Introdução

Nesta secção faz-se a contextualização, motivação e objectivos da dissertação. É aqui que são definidos os objectivos da dissertação, tendo em conta a problemática que é descrita na contextualização e motivação.

1.1 Contextualização

Tem-se assistido à proliferação de entidades detentoras de arquivo que disponibilizam na Internet os seus conteúdos digitais. O principal objectivo na disponibilização destes conteúdos digitais é permitir ao utilizador o acesso ao catálogo do arquivo a partir de qualquer parte do mundo. Contudo, este acesso, muitas das vezes pode ser precedido por uma pesquisa onerosa por parte do utilizador. Isto porque os conteúdos estão, normalmente, armazenados em sistemas de dados com as seguintes características: são autónomos, possuem estruturas de dados heterogéneas, estão geograficamente separados, possuem interfaces de pesquisa diferentes.

No sentido de promover uma melhor articulação e inter-relacionamento entre entidades detentoras de uma infra-estrutura de informação, sente-se a necessidade de construir uma rede. Neste contexto particular trata-se de uma Rede Portuguesa de Arquivos (RPA) (DGARQ, 2008a, 2008b) ou, dito de outra forma, uma rede dos conteúdos detidos e

disponibilizados pelas entidades detentoras de arquivos. Uma rede de arquivos 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 dos conteúdos de arquivo. Realça-se aqui a importância da colaboração entre as entidades aderentes, motivada por expectativas e interesses mútuos.

A adesão à Rede Portuguesa de Arquivos pressupõe o cumprimento de um conjunto mínimo de requisitos (DGARQ, 2008a):

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 ou das unidades orgânicas dela dependentes.

b) *Requisitos de acesso*: a presença 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.

Nesta dissertação apenas vão ser estudados assuntos relativos aos dois últimos conjuntos de requisitos, isto é, aos requisitos técnicos e funcionais.

1.2 Motivação

Os princípios que motivam o estabelecimento da rede portuguesa de arquivos são os seguintes: integração estrutural, neutralidade, interoperabilidade, pesquisa inter-repositórios, acessibilidade e qualidade (DGARQ, 2008a).

A principal motivação para a concepção desta rede, tendo em conta a contextualização apresentada, prende-se com a capacidade de reunir informação proveniente de diferentes entidades detentoras de conteúdos, de forma a disponibilizar serviços consolidados de consulta e localização nesse universo de informação, a partir de um único ponto de acesso. Desta forma o utilizador poderá interagir com uma única interface de pesquisa, e a partir dela, submeter os parâmetros a pesquisar e receber a informação de forma consolidada vinda de diferentes fontes de informação, evitando percorrer *site* a *site* em busca da informação que deseja recuperar.

Tal interface única de pesquisa será o Portal Português de Arquivos (PPA) (DGARQ, 2009a). De uma forma simples, esse portal terá um formulário onde o utilizador insere as expressões de pesquisa e recebe como resultado uma lista de resultados. Esta lista de resultados contém informações mínimas sobre os registos encontrados, e apontadores para o registo completo. Ao clicar num desses apontadores, o utilizador será direccionado para o contexto e descrição detalhada do registo, a qual estará na página da instituição detentora.

De um modo geral, este funcionamento é idêntico ao comportamento dos motores de busca normais da Internet (e.g. Google, Yahoo).

Para implementar este processo, é preciso ter acesso aos registos de metainformação disponibilizados pelas entidades aderentes. Um registo corresponde à descrição de um

elemento arquivístico (fundo, série, documento, etc.), obedecendo às normas de descrição arquivística. Parte desses metadados (e.g. código de referência, título, datas, nível de descrição) serão mostrados ao utilizador, na forma de uma página de resultados.

Um dos modos de obter tais metadados é através de processos de recolha e colecta automática, ou *harvesting*. Este modelo impõe que exista um sistema com uma base de dados centralizada, designada por *service provider*, que recolhe, através de processos automáticos, metadados disponibilizados por *data providers*. O *service provider* é responsável por sistematizar os metadados e por oferecer, com base neles, serviços de valor acrescentado, tais como a consulta e referência na informação armazenada. Periodicamente, por meio de um protocolo normalizado, o *service provider* consulta os *data providers*, para determinar se houve actualizações à base de dados. Em caso positivo, uma cópia dos metadados actualizados é importada para o *service provider*. Este modelo impõe a existência de dois elementos: um protocolo de comunicação e esquemas de metadados normalizados para a transferência de informação.
Um dos protocolos que segue estes requisitos é o protocolo OAI-PMH (*Open Archives Initiative – Protocol Metadata Harvesting*) (Open Archives Initiative, 2002).

1.3 Objectivos

O principal objectivo desta dissertação é modelar e prototipar um sistema que permita a federação de repositórios digitais de entidades detentoras de arquivo.

Para modelar o sistema será necessário fazer um estudo dos principais tipos de arquitectura capazes de satisfazer os requisitos impostos pela problemática enunciada. Após análise e selecção da arquitectura mais adequada, serão estudados os principais elementos que intervêm na arquitectura: protocolos de comunicação e esquemas de metadados para transferência de informação entre as entidades da rede.

A selecção da arquitectura deve ter em consideração alguns critérios que tornem facilitada a adesão de novas entidades à rede. Uma vez que a rede portuguesa de arquivos será aberta à adesão de qualquer entidade detentora de arquivo, a arquitectura da rede não deve dificultar

este processo, o qual deve ser expedito e regido por directrizes definidas no âmbito do projecto.

Os formatos de metainformação utilizados para a partilha de informação entre as entidades que participam na rede devem ser seleccionados tendo em consideração as características de simplicidade, interoperabilidade, e consenso. Isto para que haja uma normalização que torne clara e possível a comunicação entre as entidades que participam na rede.

A rede deve ser modelada de forma a ir ao encontro dos seus pressupostos (DGARQ, 2009b):

- O cliente fundamental é o cidadão, o qual tirará partido dela a partir do acesso à informação disponibilizada pelo Portal Português de Arquivos;
- A rede deverá basear-se na partilha de serviços e disponibilização dos conteúdos de arquivo de todas as entidades aderentes;
- A rede deverá ter a capacidade de se articular com outras redes de serviços, de forma a proporcionar serviços integrados ao cidadão;
- Também deverá articular-se com outras estruturas semelhantes a nível internacional, como a Europena (Europeana, 2009b), ou o APEnet (APEnet, 2009).

Desta forma, a análise, avaliação e selecção do modelo de arquitectura, protocolos de comunicação e formatos de metainformação devem ser realizados à luz dos pressupostos enumerados.

1.4 Estrutura do documento

Este documento é composto por 5 capítulos. No primeiro capítulo faz-se a introdução, onde é descrita a contextualização, motivação e objectivos da dissertação. No segundo capítulo apresentam-se os diversos modelos de arquitecturas que foram estudados e que apresentam soluções para os requisitos impostos na problemática enunciada. No terceiro capítulo apresentam-se as descrições de alguns formatos de metainformação que servem para partilhar metadados entre as entidades aderentes da rede. No quarto capítulo descreve-se a arquitectura e funcionamento do sistema que será implementado no contexto da rede portuguesa de arquivos. Por fim, no quinto capítulo tecem-se as principais conclusões, discussões e

considerações finais. Ainda neste capítulo, apresenta-se um conjunto de pontos que se podem realizar como trabalho futuro.

Capítulo 2

Modelos de arquitecturas

Neste capítulo apresentam-se os diversos modelos de arquitecturas que foram estudados e que apresentam soluções para os requisitos impostos pela problemática enunciada. Para cada um dos modelos apresentam-se as suas características, arquitectura e modo geral de funcionamento. Também é efectuada uma descrição do modelo quando é instanciado com o problema em estudo, isto é, a sua aplicabilidade ao contexto da Rede Portuguesa de Arquivos (RPA). Os modelos apresentados são: o sistema com base de dados centralizada, o sistema com centralização de metainformação através de processos de ETL (*Extract, Transform and Load*), o sistema com pesquisa federada e o sistema de recolha de metadados baseado no protocolo OAI-PMH.

No final deste capítulo é elaborada uma análise comparativa entre os modelos enumerados. Essa análise é feita com base em algumas características, ou métricas, consideradas relevantes para a avaliação do sistema. Tais características são: a taxa de revogação, a precisão, a taxa de actualização, a rapidez, a independência, a preservação e a facilidade de adesão.

2.1 Sistema com base de dados centralizada

Nesta secção, apresenta-se o modelo de arquitectura baseado numa base de dados centralizada. Começa-se por apresentar os conceitos e características gerais desta arquitectura e posteriormente descreve-se a sua aplicabilidade, caso o modelo seja instanciado com o contexto da rede portuguesa de arquivos.

2.1.1 Características gerais

Numa arquitectura com base de dados centralizada existe uma única base de dados que suporta todo o sistema operacional, isto é, todas as transacções de actualização (operações de inserção, alteração e eliminação) e consulta de dados (operações de selecção) são efectuadas sobre uma única base de dados. Desta forma, todas as aplicações do sistema quando necessitam de efectuar algum tipo de operação sobre os dados acedem a essa base de dados, a qual é normalmente gerida por um SGBD - sistema de gestão de base de dados (e.g. MySQL, SQL Server, Oracle).

Uma base de dados centralizada tem como propósito centralizar todo o armazenamento, actualizações, pesquisas e outras alterações necessárias num único local, tendo em vista vantagens como:

- Agilidade na pesquisa de informações;
- Padronização na inserção dos dados;
- Maior facilidade de manutenção;
- Redução de custos.

Contudo, dependendo do cenário onde o sistema está implementado, podem surgir problemas que tornem inviável a adopção deste tipo de arquitectura. Tais problemas podem ser:

- Baixa disponibilidade;
- Comprometimento do desempenho;
- Interdição de acesso caso ocorra uma falha no sistema central ou à sua ligação.

2.1.2 Aplicabilidade ao contexto da RPA

No contexto da rede portuguesa de arquivos, com este tipo de arquitectura, todas as entidades aderentes iriam partilhar a mesma base de dados. Sem dúvida que os principais objectivos do Portal Português de Arquivos seriam facilmente satisfeitos, pois os processos de disponibilização, recolha e partilha dos conteúdos de arquivo vindos de uma única fonte de dados seriam facilmente implementados. Contudo este modelo seria pouco viável devido aos seguintes factores:

- A indisponibilidade da base de dados reflectir-se-ia imediatamente em todas as entidades aderentes. Falhas no repositório central ou no acesso ao repositório central impediriam o acesso de qualquer transacção das entidades aderentes sobre o repositório. Note-se que o repositório central além de disponibilizar o conteúdo, isto é, suportar as operações de selecção do portal, também suportaria todo o processo transaccional de actualização (operações de inserção, alteração e remoção) de informação;
- Todas as entidades aderentes teriam que utilizar o mesmo sistema de interface com a base de dados. Isto obrigaria as entidades aderentes a adaptarem os seus sistemas de informação para poderem comunicar com o repositório central, o que coloca em causa a independência de plataformas;
- O sistema central teria que implementar um elevado controlo de acesso, impedindo que os dados de uma entidade não estejam acessíveis de modo a poderem ser alterados ou corrompidos por outros;
- O desempenho do sistema poderia estar comprometido, devido à concorrência de acessos e ao facto de todas as operações serem efectuadas sobre um sistema de dados alojado remotamente.

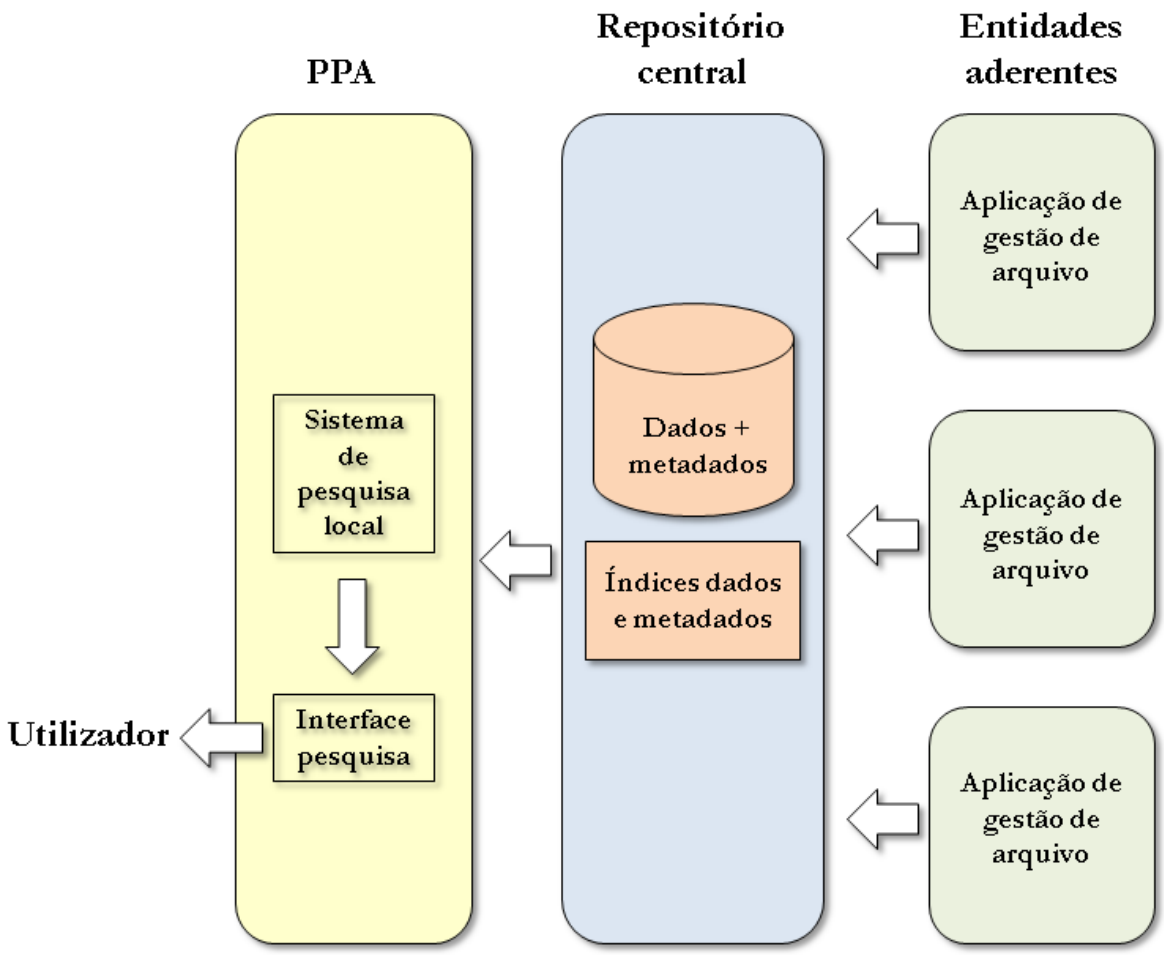

Figura 1 – Sistema com base de dados centralizada

2.2 Sistema com centralização de metainformação através de processos de ETL

Nesta secção apresenta-se o modelo de arquitectura que centraliza num único repositório, através de processos de ETL, a metainformação das várias fontes de informação. Começa-se por apresentar os conceitos e características gerais desta arquitectura e depois descreve-se a sua aplicabilidade, caso o modelo seja instanciado com o cenário da rede portuguesa de arquivos.

2.2.1 Características gerais

O processo de Extracção, Transformação e Integração (ETL do inglês *Extract, Transform and Load*) é um processo que envolve:

- A extracção de dados, a partir de fontes de dados externas;
- A transformação dos mesmos para atender às necessidades de negócio;
- A integração dos dados no sistema de dados destino.

Os processos de ETL são normalmente utilizados para o povoamento de *data warehouses* (DW), contudo, estes processos podem ser aplicados ao carregamento e povoamento periódico de qualquer base de dados.

As 3 fases que compõem o processo de ETL são descritas a seguir.

2.2.1.1 Selecção e extracção

A primeira parte do processo de ETL é a extracção de dados a partir dos sistemas fonte. Cada um destes sistemas pode apresentar formatos diferentes para a organização dos dados, o que implica a existência de diferentes selectores para a selecção e extracção de informação.

2.2.1.2 Transformação

O processo de transformação aplica uma série de regras ou funções aos dados extraídos para derivar os dados a serem integrados no sistema destino. Algumas fontes de dados necessitam de pouca manipulação, em contrapartida, noutros casos, pode ser necessário aplicar vários tipos de transformação, tais como:

- Selecção de determinadas colunas para integração numa única coluna;
- Tradução de valores codificados (e.g. se o sistema de origem armazena 1 para sexo masculino e 2 para feminino, mas o sistema destino armazena M para masculino e F para feminino), o que é conhecido como limpeza de dados;
- Tradução e codificação de valores armazenados de forma livre (e.g. mapear "Masculino","1" e "Sr." para M);
- Derivação de um novo valor calculado (e.g. montante_venda = quantidade x preco_unitario);
- Junção de dados provenientes de diversas fontes;
- Resumo de várias linhas de dados (e.g. total de vendas para cada loja e para cada região);
- Geração de valores de chaves de substituição (*surrogate keys*);
- Transposição ou rotação (transformando múltiplas colunas em múltiplas linhas ou vice-versa);
- Quebra de uma coluna em diversas colunas (e.g. transformar uma lista separada por vírgulas de uma coluna em valores individuais para diferentes colunas).

2.2.1.3 Integração

Após a selecção e as transformações de dados necessárias, esta fase contempla a integração dos dados no sistema destino. O processo de integração pode variar consoante as necessidades e requisitos da organização. Por exemplo, a periodicidade de actualização dos dados em alguns sistemas pode ser efectuado semanalmente, ao passo que outros necessitam de efectuar actualizações a cada hora. Estes critérios constituem opções estratégicas do projecto, que dependem do tempo disponível e das necessidades do negócio. Em alguns sistemas mantém-se um histórico que contempla todas as mudanças sofridas pelos dados.

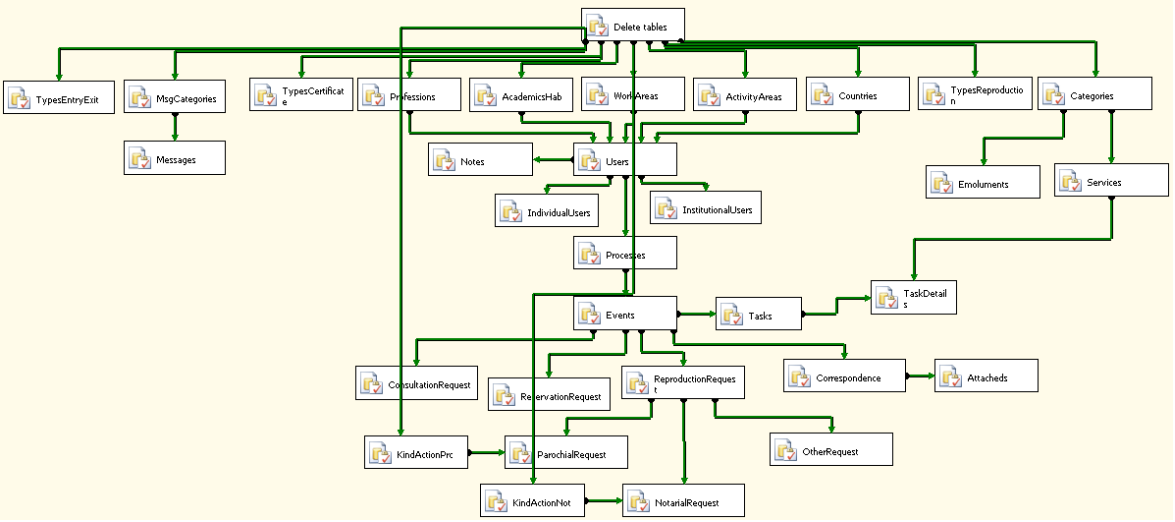

Figura 2 – Exemplo de uma implementação de um sistema de ETL

O desenvolvimento de sistemas de ETL implica um estudo em torno dos seguintes contextos:

- O reconhecimento e a avaliação dos sistemas operacionais que actuem como fontes de informação;
- A análise dos dados a extrair das fontes de informação;
- O conhecimento da estrutura da fonte de dados de destino;
- A análise, a especificação, o desenvolvimento, a instalação, a monitorização e a manutenção dos sistemas de povoamento.

2.2.2 Aplicabilidade ao contexto da RPA

No contexto da RPA, as fontes de informação corresponderiam aos repositórios das entidades aderentes. Os processos de ETL seriam responsáveis pela selecção e extracção de metadados a partir dos repositórios das entidades aderentes, os quais sofreriam as transformações necessárias para poderem ser integrados na estrutura de metadados do repositório central que alimenta o PPA. Os processos teriam que estar bem escalonados e implementados para manter o repositório central actualizado.

A maior desvantagem deste tipo de arquitectura prende-se com o facto de terem que ser implementados tantos processos de ETL quantas as diferentes estruturas de dados das

entidades aderentes. Isto porque os processos de selecção, extracção e transformação de dados inerentes ao sistema de ETL dependem da fonte de dados.

Desta forma, a adesão de uma entidade à RPA implica a implementação de um processo de ETL adaptado à estrutura de dados do seu repositório.

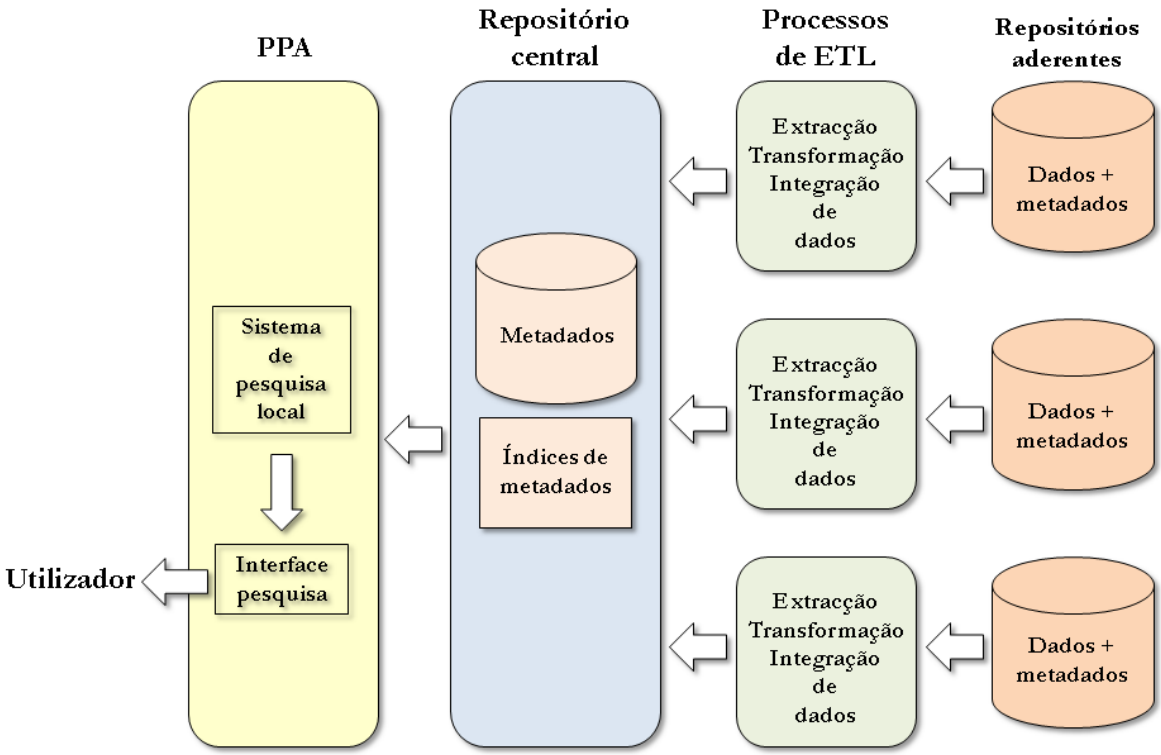

Figura 3 – Sistema com centralização de metainformação através de processos de ETL

2.3 Sistema com pesquisa federada

Nesta secção apresenta-se o modelo de arquitectura de um sistema com pesquisa federada. Começa-se por apresentar os conceitos e as características gerais desta arquitectura, posteriormente são descritos alguns dos protocolos de comunicação para possibilitar a interacção do sistema de pesquisa com as fontes de informação. Finalmente descreve-se a aplicabilidade desta arquitectura, caso este modelo seja instanciado com o contexto da rede portuguesa de arquivos.

2.3.1 Características gerais

O termo "pesquisa federada" possui vários sinónimos, os quais dependem da adopção tomada pelo autor. Entre estes sinónimos podem-se encontrar termos, como: "metapesquisa", "pesquisa distribuída", "pesquisa paralela", "pesquisa cruzada", "*broadcast search*", "*cross-database search*". Neste documento adoptou-se o termo pesquisa federada.

A pesquisa federada consiste basicamente na submissão de uma consulta a várias bases de dados em simultâneo. A consulta é enviada a cada uma das bases de dados, as quais efectuam o seu processamento e remetem a lista de resultados. Posteriormente, os resultados das consultas às várias bases de dados são consolidados e mostrados numa lista combinada de resultados, a qual pode estar ordenada pelo respectivo peso ou relevância. Além disso, poderá ser possível seleccionar bases de dados específicas e consultar apenas os seus resultados, permitindo desta forma um total controlo sobre as bases de dados a serem pesquisadas.

Como exemplo, o Dogpile (http://www.dogpile.com) permite a pesquisa simultânea nos motores de busca Google, Yahoo, Bing e Ask. Neste caso, os resultados orgânicos e patrocinados surgem indistintamente, com a referência do motor de busca fonte.

Existem outros sistemas de pesquisa federada internacional, os quais merecem destaque pela sua completude:

- pesquisa.b-on.pt – pesquisa da Biblioteca do Conhecimento Online (b-on), a qual disponibiliza o acesso ilimitado e permanente a publicações científicas internacionais de 16 editoras (FCCN, 2009b);
- O projecto MERLOT - *Multimedia Educational Resource for Learning and Online Teaching* (http://www.merlot.org/) - que se refere ao desenvolvimento cooperativo e gratuito de recursos baseados na web para que professores, alunos e profissionais em geral possam facilmente encontrar e disponibilizar materiais digitais de aprendizagem com as respectivas avaliações e indicações de usos mais apropriados (MERLOT, 2009). Pode-se aceder à pesquisa federada a partir de http://fedsearch.merlot.org/fedsearch;
- www.scitopia.org – pesquisa federada em bibliotecas digitais no âmbito de disciplinas de ciência e tecnologia para pesquisa de trabalhos académicos e patentes (Deep Web Technologies, 2009);
- www.science.gov – portal para pesquisa de informação científica do governo dos Estados Unidos e resultados de investigação (U.S. Department of Energy, 2009b);
- worldwidescience.org – portal de busca em bases de dados e portais científicos internacionais (U.S. Department of Energy, 2009c);
- www.scienceaccelerator.gov – pesquisa informação nos recursos científicos e técnicos do Departamento de Energias dos Estados Unidos (U.S. Department of Energy, 2009a);
- metalib.bris.ac.uk – pesquisa de conteúdos digitais em repositórios da Universidade de Bristol, Inglaterra (University of Bristol, 2008).

2.3.2 Protocolo Z39.50

O protocolo Z39.50 é um dos protocolos de comunicação que pode ser usado para permitir a pesquisa e a recuperação de informação em redes de computadores distribuídos.

A forte necessidade de haver um mecanismo que normalizasse a comunicação entre sistemas de computadores levou ao surgimento deste protocolo. Isto levou a NISO (*The National Information Standards Organization*) (NISO, 2009a) a estabelecer uma comissão para elaborar um protocolo para recuperação de informação. Após estudos iniciados nos anos 70, foi lançada

em 1988 a primeira versão do protocolo Z39.50. Em 1992 foi lançada a segunda versão, já revista e em conformidade com as normas da ANSI (*American National Standards Institute*) (ANSI, 2009).

A norma ANSI/NISO Z39.50 (NISO, 2009b) é independente de plataformas, permitindo assim, a interoperabilidade entre diferentes sistemas de informação com diferentes sistemas operativos, equipamentos, interfaces de pesquisa e sistemas de gestão de bases de dados. Através de uma implementação Z39.50 permite-se a conexão e acesso a múltiplos sistemas de informação de uma forma quase transparente para o utilizador. Desta forma, é possível a pesquisa de informação a partir de um único ponto de acesso, evitando que o utilizador percorra vários pontos de pesquisa, onde terá que se adaptar às várias *interfaces*, comandos e técnicas de pesquisa (Figura 4).

Este tipo de sistema é vantajoso para repositórios digitais que pretendam oferecer ao utilizador um único ponto de acesso para a realização de pesquisas no seu catálogo local e em bases de dados referenciais e remotas.

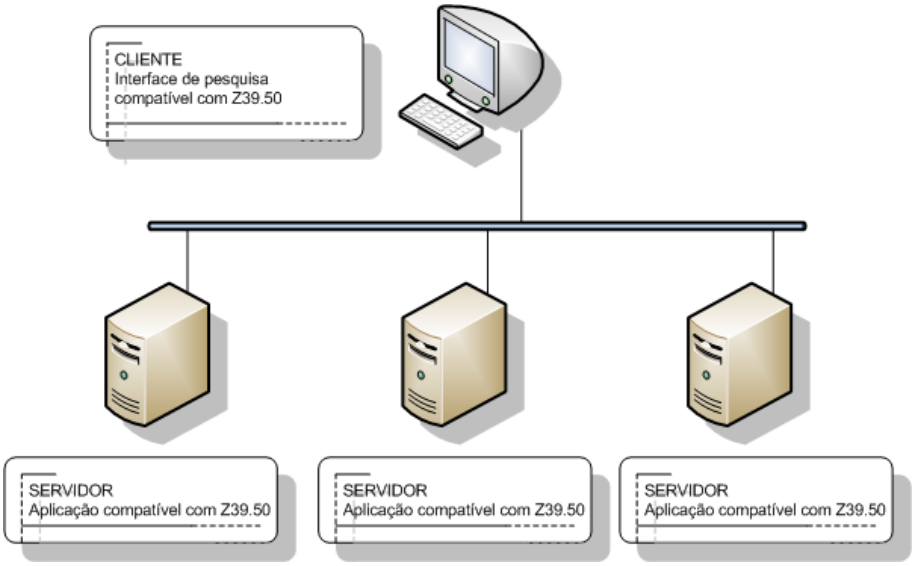

Figura 4 – Arquitectura com protocolo Z39.50

2.3.3 SRU

O protocolo SRU surgiu de um projecto de reestruturação do protocolo Z39.50, o qual foi inicialmente denominado pelo acrónimo ZIG/ZING (Z39.50 *next generation*), posteriormente modificado para SRU (*Search/Retrieval via URL*). Seguindo o âmbito patente no protocolo Z39.50, surgiram os protocolos desenhados para aplicações digitais, o SRU e, posteriormente, o SRW (*Search/Retrieve Web Service*), ambos desenvolvidos pela *Library of Congress*, nos Estados Unidos. Os protocolos SRU/SRW trouxeram as facilidades e características do protocolo Z39.50 para o contexto da internet, em ambientes de URL e de serviços web.

O protocolo SRU, pelo uso do protocolo HTTP introduz uma maior abrangência e normalização na comunicação entre sistemas, tornando o desenvolvimento de *software* para estes fins mais facilitado.

A utilização deste protocolo também é importante para uma conciliação na catalogação de dados. Isto é possível porque o protocolo SRU recupera registos no formato MARC, utilizado para catalogar registos de bibliotecas, e Dublin Core, utilizado para catalogar arquivos digitais, os quais estão entre os principais formatos utilizados na catalogação de informações documentais.

O SRU utiliza o serviço Web RESTful (*Representational State Transfer*), o qual codifica comandos do cliente para o servidor numa *string* (sequência de caracteres), na forma de um URL. Cada um desses valores é especificado no formato `nome=valor`, e a cada nova especificação é atribuído um novo parâmetro para o servidor.

O servidor processa estes parâmetros pré-estabelecidos e retorna os valores no formato XML (*eXtensible Markup Language*) (Figura 5).

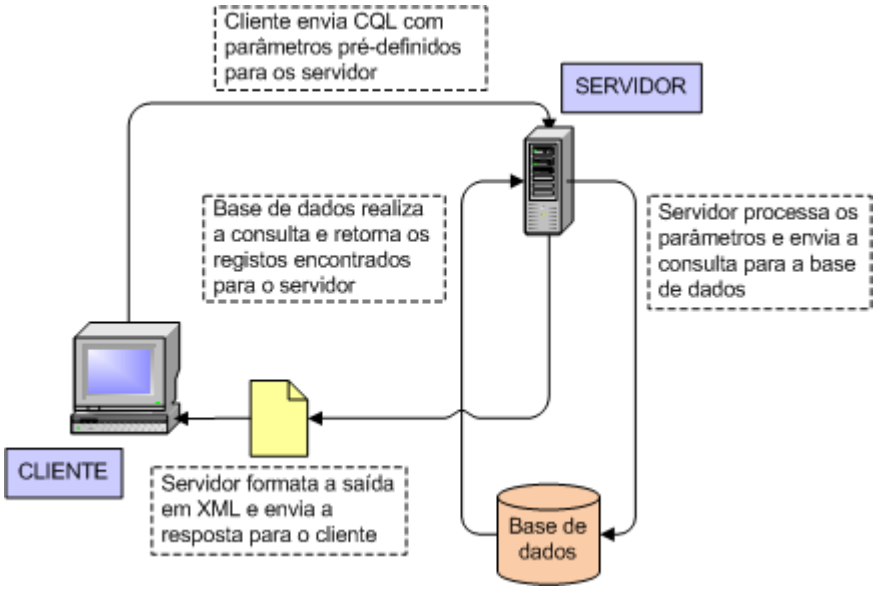

Figura 5 – Arquitectura baseada em SRU

2.3.3.1 Estrutura de um URL SRU

```
Protocolo://endereço-do-servidor:porta/nome-base-dados?versao=numero-
versao&operacao=tipo-operacao&consulta=consulta&opcoesRetorno=opcoes&
parametrosRequisicao=parametros
```

Exemplo:
http://z3950.loc.gov:7090/voyager?version=1.1&operation=searchRetrieve&query=dinosaur&maximumRecords=1&recordSchema=dc

Onde:

- Protocolo: http
- Endereço do servidor: z3950.loc.gov
- Porta: 7090
- Nome da base de dados: voyager
- Versão do protocolo: 1.1
- Operação: searchRetrieve (ver na secção seguinte os tipos de operações)

2.3.3.2 Operações SRU

Existem três tipos de operações que podem ser utilizadas no serviço SRU/SRW: *explain*, *scan* e *searchRetrieve*.

- *explain*: esta operação é utilizada pelo cliente para obter mais informações sobre a situação do servidor. Algumas opções que são listadas pelo *explain* são: versão, bases de dados, localização das bases de dados, serviços oferecidos, etc.
  http://z3950.loc.gov:7090/voyager?operation=explain
- *scan*: utilizada para listar e numerar os termos encontrados numa pesquisa na base de dados. Deve ser utilizada com a opção *scanClause* para se fazer a pesquisa do termo desejado. A pesquisa pelo termo "mundo", por exemplo, na base de dados *voyager*, seria representada assim:
  http://z3950.loc.gov:7090/voyager?operation=scan&scanClause=mundo
- *searchRetrive:* a operação principal do serviço SRU/SRW. Identifica a requisição solicitada, realiza a pesquisa na base de dados e retorna os resultados encontrados. Esta opção deve ser utilizada com o parâmetro *query* para se fazer a pesquisa do termo desejado. Os resultados podem ser requeridos num determinado formato de acordo com a definição do cliente. Para efectuar a pesquisa pelo termo "portugal", no formato "dc":
  http://z3950.loc.gov:7090/voyager?version=1.1&operation=searchRetrieve&query=portugal&maximumRecords=10&recordSchema=dc

2.3.3.3 Linguagem de Query

As consultas são expressas em *Common Query Language* (CQL) (Congress, 2008) para a versão 1.1 e para versões superiores em *Contextual Query Language* (CQL) (Congress, 2008). Trata-se de uma linguagem formal utilizada para representar pesquisas em sistemas de informação. Como mostrado em exemplos anteriores, utiliza-se o parâmetro *scanClause* para a operação *scan* e o parâmetro *query* para *searchRetrive*.

2.3.4 SRW

O SRW foi desenvolvido com os mesmos propósitos do SRU, diferenciando-se pelo uso de um serviço pré-estabelecido, o *Simple Object Access Protocol* (SOAP), e não de um URL.

2.3.4.1 Protocolo SOAP

O SOAP (*Simple Object Access Protocol*) é um protocolo para troca de informações estruturadas em plataformas descentralizadas e distribuídas, utilizando tecnologias baseadas em XML. O SOAP encarrega-se de encapsular e transportar as chamadas de procedimento remoto (*Remote Procedure Calls*, ou RPCs), criando mensagens estruturadas no formato XML para a troca de informação em ambientes remotos. Definido de modo mais simples, pode-se compreendê-lo como um protocolo para aceder a um serviço web e possibilitar a interoperabilidade entre aplicações (W3C, 2000).

O uso do SOAP tem como vantagem a sua independência em relação às linguagens de programação, a sua simplicidade e o facto de ser extensível, o que permite a sua utilização em qualquer aplicação. Pode-se citar ainda como vantagem o facto do SOAP poder realizar as suas chamadas sobre o protocolo HTTP e ser estruturado em XML, ou seja, duas tecnologias instituídas como padrão, neutras, e usadas em larga escala.

Na Figura 6 está representada uma arquitectura cliente/servidor onde é utilizado o protocolo HTTP para troca de mensagens SOAP.

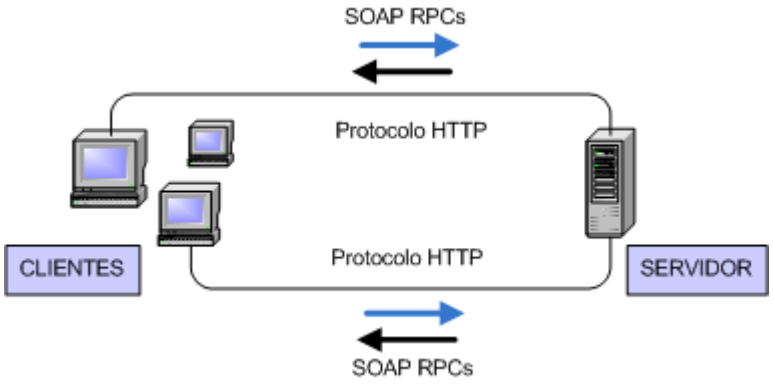

Figura 6 – Arquitectura baseada em SRW

2.3.4.2 Estrutura de uma mensagem SOAP

Uma mensagem SOAP é composta por três elementos: envelope, cabeçalho e corpo (Figura 7). O envelope contém o cabeçalho e o corpo, sendo o elemento de uma mensagem SOAP indispensável para a utilização do protocolo. O cabeçalho, ou *header*, é um elemento opcional

no envelope e é utilizado quando a mensagem é enviada para um determinado nó. Caso seja necessária a sua utilização, esta deverá ser a primeira informação no conteúdo do envelope. O corpo, ou *body*, é um elemento obrigatório no envelope. Dentro do corpo encontra-se o *payload*, que é a informação que se quer transportar para o destino. O corpo também pode transmitir e receber mensagens de erro e alerta através do elemento opcional *fault*.

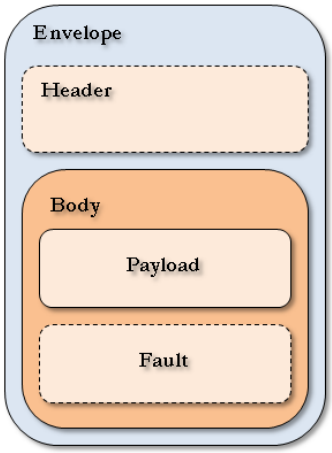

Figura 7 – Estrutura de uma mensagem SOAP

Como se pode perceber pelas definições descritas acima, a diferença fundamental entre os protocolos SRU e SRW é formal, reside apenas na forma como a solicitação é realizada. A simplicidade do protocolo SRU tem-lhe dado vantagem de utilização em comparação com o protocolo SRW, onde é necessário proceder à geração de um objecto XML.

2.3.5 Aplicabilidade no contexto da RPA

A aplicação de um sistema baseado em pesquisa federada no contexto da RPA consistiria basicamente na interacção entre os dois tipos de entidades constituintes da rede: o servidor cliente onde seria implemento o sistema de pesquisa (PPA) e as fontes de informação, que corresponderiam aos repositórios das entidades aderentes. Esta interacção, baseada num protocolo de comunicação (e.g. Z39.50, SRU), compreenderia o envio de pedidos e respostas necessários para satisfazer as pesquisas efectuadas a partir do PPA.

Uma pesquisa efectuada por um utilizador a partir do PPA consistiria no envio dessa consulta a todas as bases de dados das entidades aderentes em simultâneo. Posteriormente, as respostas

às consultas enviadas a cada base de dados seriam consolidadas e mostradas numa lista combinada de resultados. Esta lista de resultados mostraria apenas alguma metainformação de descrição de arquivo (e.g. código de referência, título, datas, nível de descrição) dos registos recuperados. Nessa lista, para cada registo, existiria um apontador para a informação detalhada do registo, a qual seria da responsabilidade da entidade aderente correspondente.

A implementação deste tipo de arquitectura de forma a satisfazer os requisitos impostos pela RPA implicaria que as entidades aderentes respondessem a pedidos de acordo com um dos protocolos de comunicação suportados e definidos no âmbito do projecto.

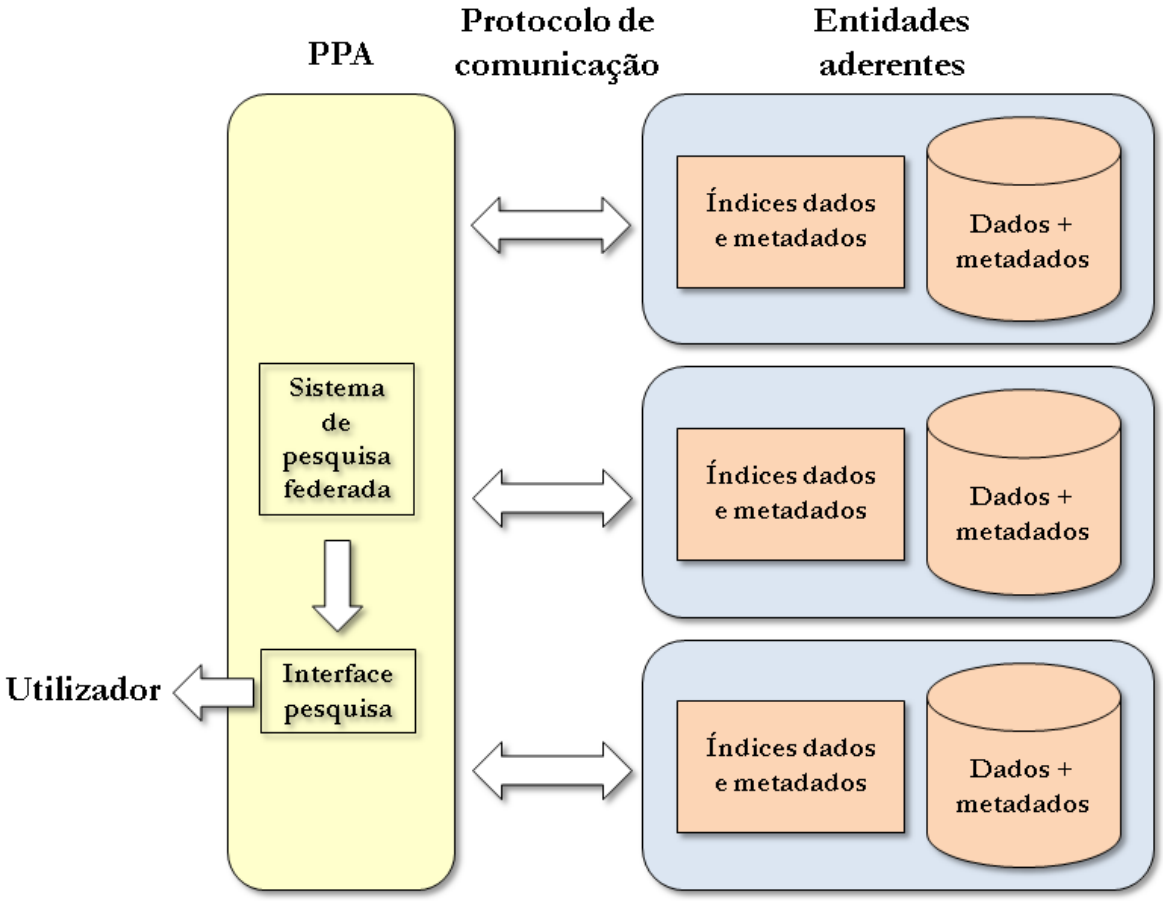

Figura 8 – Sistema com pesquisa federada

2.4 Sistema de recolha de metadados baseado no protocolo OAI-PMH

Nesta secção apresenta-se o modelo de arquitectura de um sistema de recolha (*harvesting*) de metadados baseado no protocolo OAI-PMH (Open Archives Initiative, 2002). Numa primeira secção apresentam-se os conceitos e as características gerais deste modelo de arquitectura, noutra secção descrevem-se os conceitos do protocolo OAI-PMH e finalmente descreve-se a aplicabilidade do modelo caso seja instanciado com o contexto da RPA.

2.4.1 Características gerais

Num sistema de recolha de metadados, ou sistema de *harvesting*, existe uma entidade designada por *harvester* ou agregador que recolhe, a partir dos repositórios aderentes, os metadados disponíveis, os quais são colectados num repositório central para disponibilizar uma interface única de pesquisa. A partir dos resultados da pesquisa, onde é mostrada metainformação do registo, os utilizadores são direccionados directamente para o registo original e/ou documento completo localizado no correspondente repositório local da entidade parceira.

A vantagem dos sistemas de recolha de metadados em relação à pesquisa federada é a persistência dos dados, pois a existência de uma base de metadados assegura que seja sempre fornecida uma resposta ao utilizador final, além de garantir a preservação dos metadados ao longo do tempo. Na pesquisa federada, problemas no servidor ou alteração de endereço levam à perda de ligação à informação.

A recolha de metadados ganhou maior ênfase após o surgimento do OAI – *Open Archive Iniciative* – em 1999, que pressupõe uma estrutura técnica e organizacional para recolha de metadados normalizados de modo a facilitar a recuperação do conteúdo armazenado em repositórios digitais. Esta iniciativa consiste num esforço da comunidade científica para garantir a interoperabilidade entre arquivos digitais, o que é possível pelo uso do protocolo OAI-PMH (*Open Archive Iniciative - Protocol for Metadata Harvesting*).

Neste protocolo participam dois tipos de entidades que comunicam entre si: os *data providers* e os *service providers*. Os *data providers* que implementam e gerem repositórios digitais e que

utilizam o protocolo OAI-PMH para expor os seus metadados que podem assim ser recolhidos e armazenados pelo *service provider* ou agregador. Este, além de recolher e armazenar automaticamente os metadados expostos pelos *data providers*, organiza-os e oferece ao utilizador final produtos e serviços de valor agregado por via de uma interface única de acesso.

2.4.2 Protocolo OAI-PMH

A *Open Archives Initiative* (Open Archives Initiative, 2002) tem um papel muito importante para permitir a interoperabilidade entre repositórios. O seu principal objectivo é permitir que diferentes repositórios geograficamente separados possam inter-operar formando uma federação de repositórios. Para isso foi criado um protocolo de comunicação entre repositórios, o OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*) que define como deve ser realizada a transferência de metadados entre duas entidades: *data providers* e *service providers*. Os *data providers* suportam o OAI-PMH como meio de disponibilizar os seus metadados. Os *service providers* enviam pedidos OAI-PMH a *data providers* e utilizam os metadados recolhidos como base para fornecer serviços de valor acrescentado, tais como serviços de pesquisa e referência sobre a informação armazenada.

A interacção entre as duas entidades básicas do OAI-PMH pode ser vista na Figura 9. Pode-se observar que um *service provider* que deseja realizar uma colheita de metadados envia um pedido HTTP para um *data provider* que, de acordo com a requisição solicitada, envia como resposta os metadados solicitados em formato XML, segundo o *schema* do OAI-PMH. Com base nos metadados colectados, o *service provider* pode, então, oferecer um determinado serviço como, por exemplo, um sistema de pesquisa.

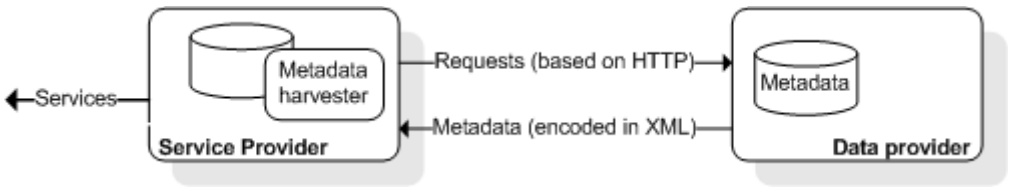

Figura 9 – Interacção entre as entidades do OAI-PMH

Para que seja possível a tarefa de *harvesting* de metadados a partir dos *data providers* são definidos seis tipos de requisições: *Identify, ListMetadataFormats, ListSets, ListIdentifiers, ListRecords, e GetRecord*. Para mais detalhes ver (OAI-PMH).

2.4.2.1 Definições e conceitos

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

Harvester

Um *harvester* é uma aplicação cliente que envia pedidos OAI-PMH. O *harvester* é operado por um *service provider* com o objectivo de colectar metadados a partir dos repositórios.

Repository

Um *repository* ou repositório é um servidor acessível na rede que pode processar os 6 pedidos OAI-PMH. Um repositório é gerido por um data provider que expõe os metadados aos *harvesters*. Podem ser distinguidas 3 entidades distintas relacionadas com os metadados que estão acessíveis pelo OAI-PMH:

- *Resource* – Um *resource* pode ser um objecto físico ou digital, armazenado num repositório ou noutra base de dados. A sua natureza está fora do âmbito do OAI-PMH.
- *Item* – Um *item* é um componente de um repositório de metadados acerca de um recurso que pode ser divulgado.
- *Record* – Um *record* é um registo de metadados num formato de metadados específico. Um *record* é retornado na resposta (codificada em XML) a um pedido de recolha de um item num determinado formato de metadados.

Item

Um *item* é um componente de um repositório de metadados acerca de um recurso que pode ser divulgado. Um *item* é conceptualmente um recipiente que armazena ou gera dinamicamente metadados sobre um único recurso em vários formatos, cada um dos quais podem ser recolhidos como registos através do OAI-PMH. Cada *item* tem um identificador que é único no âmbito do repositório do qual é um constituinte.

Unique Identifier

Um identificador único identifica de forma não ambígua um *item* dentro de um repositório. O identificador único é usado nos pedidos OAI-PMH para extrair metainformação de um *item*.

Record

Um *record* é um registo de metainformação expressa num único formato. Um *record* é retornado em XML como resposta a um pedido OAI-PMH da metainformação de um item. Um *record* é identificado por um identificador único do *item* a partir do qual o *record* está disponível, um *metadataPrefix* que identifica o formato dos metadados do registo e um *datestamp* do registo. A codificação XML dos registos é organizada da seguinte forma:

```
<header>
       <identifier></identifier>
       <dateStamp></dateStamp>
       <setSpec></setSpec>
</header>
<metadata>
       <oai_dc:ead xmlns...>
       <oai_dc>
       …
       </oai_dc>
</metadata>
<about>
       <provenance></provenance>
       …
</about>
```

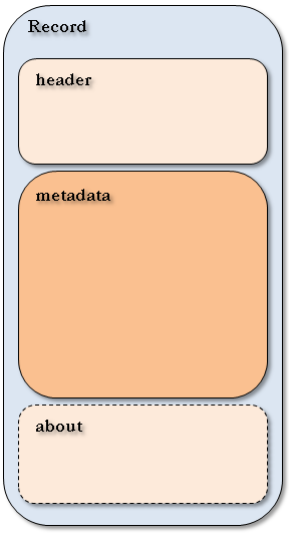

Figura 10 – Organização do *Record*

Deleted records

Existem **3 níveis de suporte** para registos eliminados:

1. **no** - o repositório não mantém informação sobre as remoções. Em nenhuma resposta é revelado o *deleted status*;
2. **persistent** - o repositório mantém informação acerca dos registos eliminados sem limites de tempo. O repositório que apresenta este nível de suporte guarda o histórico das remoções e consistentemente revela o estado do registo apagado.
3. **transient** - o repositório não garante que a lista de remoções é mantida de forma persistente e consistente. Um repositório que revela este nível de suporte pode revelar um *deleted status* para os registos.

A resposta a um pedido **GetRecord** de um registo eliminado deve incluir um *header* com o atributo **status="deleted"** e não deve incluir as partes **metadata** ou **about**.

Set

Um *set* é um elemento opcional para agrupar *items*, e para possibilitar recolhas selectivas. Desta forma os repositórios podem organizar *items* em *sets*. A organização dos *sets* pode ser plana, isto é, uma simples lista, ou hierárquica. Quando um repositório apresenta uma organização em *sets*, esta informação deve ser incluída nos cabeçalhos dos *items* devolvidos em resposta aos pedidos *ListIdentifiers*, *ListRecords* e *GetRecord*.

Selective Harvesting

A recolha selectiva permite aos *harvesters* limitar os pedidos de recolha a porções de metainformação disponível no repositório. O OAI-PMH suporta recolhas selectivas através de dois tipos de critérios que podem ser combinados no pedido OAI-PMH: *datestamps* e *sets*.

a) *Datestamps*

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

- *from* – maior ou igual que
- *until* – 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 enviar um erro *badArgument*.

Os intervalos *datestamp* para recolha selectiva são expressos nos argumentos *from* e *until* que podem ser submetidos nos pedidos *ListRecords* e *ListIdentifiers*. Os repositórios devem seguir as seguintes regras para criar a resposta com a lista de registos/identificadores que correspondem ao intervalo indicado no pedido de acordo com o tipo de alteração que ocorreu dentro do repositório.

- criação – a resposta deve incluir os registos, que correspondem ao argumento *metadataPrefix* e que ficaram disponíveis no repositório no intervalo de datas definido pelos argumentos *from* e *until*;
- alteração – a resposta deve incluir os registos, que correspondem ao argumento *metadataPrefix* e que foram alterados no intervalo de datas definido pelos argumentos *from* e *until*;
- eliminação – dependendo do nível que o repositório mantém para os registos eliminados, a resposta deve incluir os cabeçalhos dos registos que correspondem ao argumento *metadataPrefix* e que foram eliminados no intervalo de datas definido pelos argumentos *from* e *until*;

b) Sets

Os *harvesters* podem especificar um *set* como critério para recolha selectiva. Para especificar um *set*, baseado na recolha selectiva, o *setSpec* é incluído como um valor do argumento opcional *set* nos pedidos *ListRecords* e *ListIdentifiers*.

2.4.2.2 Características do protocolo

Os pedidos OAI-PMH são expressos da mesma forma que os pedidos HTTP.

a) Formato do pedido

Os pedidos OAI-PMH devem ser submetidos usando um dos métodos HTTP: GET ou POST. O método POST tem a vantagem de não impor limitações no tamanho dos

argumentos. Os repositórios devem suportar ambos os métodos. Existe um único URL base para todos os pedidos. URL base: host:port/path

Adicionalmente ao URL base, todos os pedidos consistem numa lista de argumentos na forma `key=value`. Os argumentos podem aparecer por qualquer ordem e múltiplos argumentos são separados por &. Cada pedido OAI-PMH deve ter pelo menos um par `key=value` que especifica o pedido OAI-PMH enviado pelo *harvester*:

- key – verb
- value – um dos pedidos OAI-PMH definidos

b) Formato da resposta

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

Content-type

O content-type retornado deve ser text/xml para todos os pedidos.

2.4.2.3 Pedidos e respostas do Protocolo

Nesta secção descrevem-se os pedidos, ou *verbs*, definidos no protocolo OAI-PMH.

Identify

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

ListMetadataFormats (identifier)

Este parâmetro é utilizado para recuperar os formatos de metainformação disponíveis no repositório. É obrigatória a implementação de pelo menos o formato de metainfrmação Dublin Core.

ListSets

Utilizado para recuperar a estrutura do conjunto do repositório, isto é, traz a árvore de assuntos que classificam os documentos no repositório ou outro conjunto de classificação.

GetRecord (identifier, metadataPrefix)

Este *verb* é utilizado para recuperar um registo individual de metainformação existente num repositório. Os argumentos devem especificar o identificador do *item* de onde o registo é solicitado e o formato da metainformação que deve ser incluído no registo.

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.

A Figura 11 mostra um exemplo de um pedido OAI-PMH. Trata-se de um pedido que lista os formatos de metainformação que podem ser disseminados a partir do repositório `http://www.perseus.tufts.edu/cgi-bin/pdataprov` para o *item* com identificador `oai:perseus.tufts.edu:Perseus:text:1999.02.0119`.

A resposta a este pedido (Figura 12) mostra que 3 formatos de metadados são suportados para o identificador dado: oai_dc, olac e perseus.

```
http://www.perseus.tufts.edu/cgi-bin/pdataprov?
      verb=ListMetadataFormats&identifier=oai:perseus.tufts.edu:Perseus:text:1999.02.0119
```

Figura 11 – Pedido OAI-PMH com o *verb ListMetadataFormats*

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
         http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-02-08T14:27:19Z</responseDate>
  <request verb="ListMetadataFormats"
    identifier="oai:perseus.tufts.edu:Perseus:text:1999.02.0119">
    http://www.perseus.tufts.edu/cgi-bin/pdataprov</request>
  <ListMetadataFormats>
   <metadataFormat>
     <metadataPrefix>oai_dc</metadataPrefix>
     <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd
       </schema>
     <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/
       </metadataNamespace>
   </metadataFormat>
   <metadataFormat>
     <metadataPrefix>olac</metadataPrefix>
     <schema>http://www.language-archives.org/OLAC/olac-0.2.xsd</schema>
     <metadataNamespace>http://www.language-archives.org/OLAC/0.2/
      </metadataNamespace>
   </metadataFormat>
   <metadataFormat>
     <metadataPrefix>perseus</metadataPrefix>
     <schema>http://www.perseus.tufts.edu/persmeta.xsd</schema>
     <metadataNamespace>http://www.perseus.tufts.edu/persmeta.dtd
       </metadataNamespace>
   </metadataFormat>
 </ListMetadataFormats>
</OAI-PMH>
```

Figura 12 – Resposta OAI-PMH

2.4.3 Aplicabilidade ao contexto da RPA

A adopção desta arquitectura no contexto da RPA exige que as entidades que pretendam aderir à rede materializada pelo PPA implementem um conjunto de directrizes e requisitos técnicos. A arquitectura apresentada nesta secção tem a vantagem de se reger por protocolos e normas, exigindo desta forma que os repositórios de dados (i.e. *data providers*) agregados pelo PPA usem um conjunto de directrizes e normas comuns, no sentido de garantir a interoperabilidade e qualidade dos resultados das pesquisas.

A imposição destes requisitos é essencial para se conseguir um bom funcionamento do PPA e a sua credibilidade junto dos seus potenciais utilizadores, em primeiro lugar a comunidade nacional, mas também a comunidade internacional aquando da integração com a Europena e Portal Europeu de Arquivos.

Os repositórios de dados individuais (*data providers*) para poderem participar na rede e consequentemente serem agregados pelo PPA, devem seguir os seguintes critérios e requisitos técnicos:

- Ser compatíveis com o protocolo OAI-PMH versão 2 (Open Archives Initiative, 2002);
- Disponibilizar os metadados seguindo nos formatos definidos no âmbito do projecto.
- Fornecer o URL base (válido e operacional) OAI-PMH do repositório de dados;
- Implementar as directrizes de interoperabilidade definidas pela DGARQ, nomeadamente ao nível da descrição dos conteúdos, vocabulários controlados, elementos de metadados obrigatórios e formato das referências;
- Os registos de metainformação fornecidos pelos repositórios de dados deverão encontrar-se em acesso livre, sendo que a sua consulta integral online deverá sempre ser possível através de indicação no PPA de que se pretende recuperar documentos em acesso livre.

Todos estes requisitos técnicos serão validados por uma ferramenta que ficará disponível *online* para garantir a qualidade dos metadados fornecidos pelas entidades aderentes.

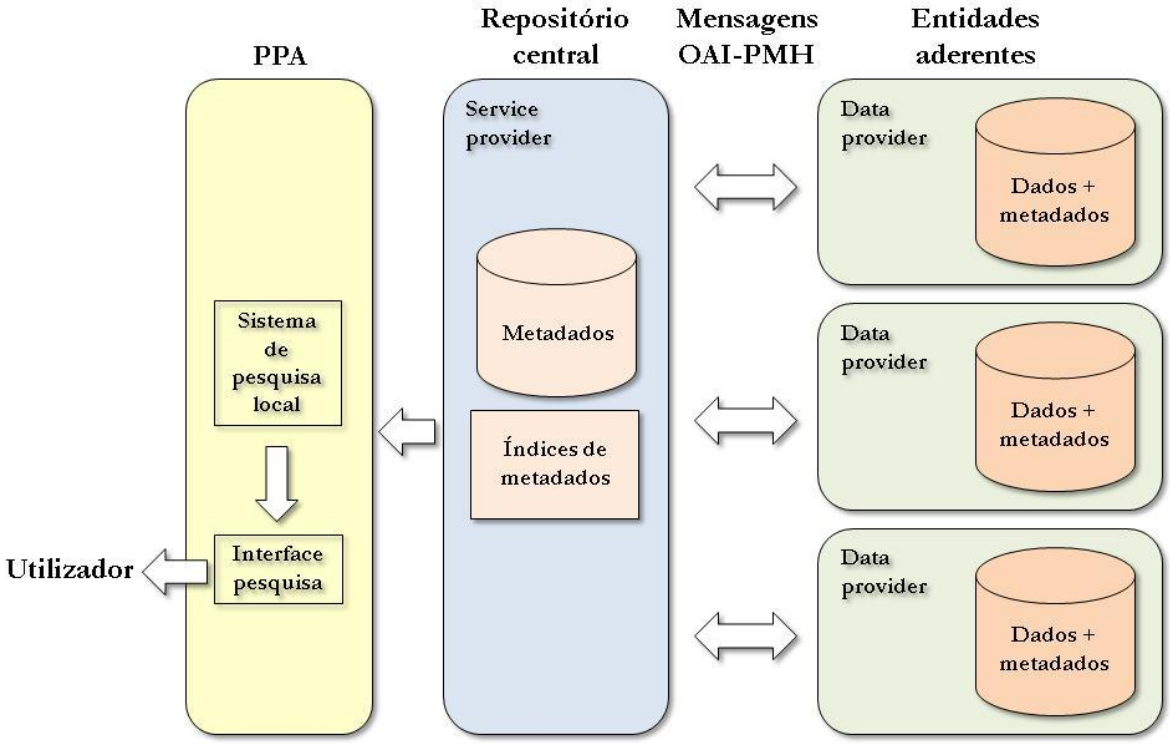

Figura 13 – Sistema com recolha de metadados baseado no protocolo OAI-PMH

2.5 Análise comparativa

Nesta secção é feita uma análise comparativa entre os modelos de arquitecturas que foram apresentados atrás. Esta análise tem como objectivo a escolha da arquitectura mais adequada para cumprir os requisitos desta problemática. Para esta análise utilizam-se algumas métricas, as quais são definidas a seguir.

Algumas das métricas utilizadas são as mesmas que a b-on - Biblioteca do Conhecimento Online (FCCN, 2009a) utilizou para estudar arquitecturas alternativas de pesquisa que pudessem contribuir para uma maior adequação do portal b-on às actividades de pesquisa e acesso à informação científica disponibilizada.

Nesta fase, a aplicação de uma destas métricas a um modelo ainda não produz resultados concretos, uma vez que não existe uma implementação desses modelos no cenário descrito.

Desta forma, apenas será feita uma previsão dos resultados com base nas características dos modelos, descritas nas secções anteriores.

Taxa de revocação (*recall*): mede a relação entre a informação relevante para uma pesquisa e a informação que é efectivamente retornada nos resultados.
Podem existir diversos documentos na base de dados que o utilizador do sistema considere relevantes, mas somente alguns deles serão recuperados pelo sistema. A taxa de revogação de uma consulta é dada pelo número de documentos relevantes recuperados pelo sistema dividido pelo número total de documentos relevantes existente na base de dados.

$$Taxa\ revogação = \frac{Número\ de\ registos\ relevantes\ recuperados}{Número\ total\ de\ registos\ relevantes\ na\ base\ de\ dados}$$

Precisão (*precision*): mede se os resultados retornados para uma pesquisa estão de acordo com as necessidades dos utilizadores, o que pode ser traduzido na quantidade de registos relevantes recuperados pelo sistema, divididos pelo número total de registos recuperados. Por exemplo, se para uma pesquisa realizada forem recuperados 6 registos e destes apenas 3 forem realmente relevantes, a precisão do sistema é 0,5 ou 50%. A polissemia[1] pode produzir baixas taxas de precisão, pois eventualmente documentos irrelevantes podem ser recuperados.

$$Precisão = \frac{Número\ de\ registos\ relevantes\ recuperados}{Número\ total\ de\ registos\ recuperados}$$

Actualização: mede a diferença de tempo entre a publicação de uma informação e a sua inclusão em resultados de pesquisa.

Rapidez: mede o intervalo de tempo entre a submissão de uma pesquisa por parte de um utilizador e a recepção dos resultados respectivos.

Independência: analisa a dependência da qualidade das pesquisas de entidades externas.

[1] Polissemia - É a propriedade que uma mesma palavra tem de apresentar vários significados (multiplicidade semântica).

Preservação: avalia o contributo do sistema no âmbito da preservação e acesso perpétuo aos documentos.

Facilidade de adesão: mede o esforço do processo de adesão de uma entidade detentora de arquivo à RPA, isto é, todo o esforço até se tornar uma entidade aderente da rede.

Na tabela seguinte, para cada arquitectura e métrica é atribuída uma previsão do resultado por aplicação dessa métrica a essa arquitectura. Tal previsão é feita com base em 4 valores por ordem crescente de cumprimento: (--, -, +, ++).

	Sistema de base de dados centralizada	Centralização de metainformação através de processos de ETL	Sistema com pesquisa federada	Sistema de recolha de metadados baseado no protocolo OAI-PMH
Taxa de revogação	++	+	-	+
Precisão	++	+	-	+
Actualização	++	+	++	+
Rapidez	-	+	--	+
Independência	++	+	--	+
Preservação	++	+	--	+
Facilidade de adesão	--	-	+	++

Tabela 1 – Análise comparativa dos modelos de arquitectura

A **taxa de revogação** e a **precisão** são influenciados pelo número de registos relevantes recuperados numa pesquisa. Dito de outra forma, estas duas métricas são tanto maiores quanto maior o número de registos relevantes recuperados na pesquisa. O número de registos relevantes recuperados depende da correcta relação entre o sistema de pesquisa e as bases de dados.

O valor destas métricas para um sistema de base de dados centralizado é alto, dado que as pesquisas incidem sobre uma única base de dados local, o que oferece um total conhecimento e controlo entre o sistema de pesquisa e a base de dados. Num sistema com centralização de metainformação através de processos de ETL e através do protocolo OAI-PMH a taxa de revogação também é elevada, pois a metainformação é periodicamente recolhida para uma base de dados única sobre a qual funciona o sistema de pesquisa. Num sistema com pesquisa

federada a taxa de revogação é baixa, uma vez que as estruturas de dados onde a pesquisa incide são diversas, o que pode influenciar a diferença entre a quantidade de informação relevante para uma pesquisa e a informação que é efectivamente retornada nos resultados.

A **actualização** num sistema de base de dados centralizada é máxima, pois os dados actualizados (através de operações de inserção, alteração ou eliminação) pelas aplicações cliente ficam imediatamente visíveis nos resultados de uma pesquisa. De forma análoga, num sistema com pesquisa federada qualquer alteração de dados efectuada em qualquer das entidades aderentes reflecte-se de forma imediata nas selecções efectuadas pelas pesquisas. Nos sistemas com centralização de metainformação através de processos de ETL ou através do protocolo OAI-PMH a actualização dos dados depende da configuração temporal das operações de povoamento, isto é, da periodicidade da actualização da base de dados central.

A **rapidez** na recuperação de dados pelo sistema de pesquisa será mais elevada para os sistemas que colectam informação (através de processos de ETL ou baseados no protocolo OAI-PMH) numa base de dados central, pois a pesquisa é efectuada apenas sobre uma única base de dados local com a informação de todas as entidades detentoras aderentes.
Num sistema com pesquisa federada o intervalo de tempo entre a submissão de uma pesquisa por parte de um utilizador e a recepção dos resultados respectivos diz respeito ao tempo necessário para enviar a pesquisa a todas as entidades detentoras, processamento da pesquisa em cada uma dessas entidades, envio dos resultados e consolidação e apresentação dos resultados. Este intervalo de tempo pode ser elevado e tende a aumentar com o aumento do número de entidades detentoras aderentes.
Se for utilizada uma arquitectura com base de dados centralizada o desempenho e velocidade de pesquisa podem estar comprometidos, uma vez que a base de dados além de suportar as pesquisas submetidas através do portal, suporta todas as transacções de actualização de dados das entidades aderentes.

A **preservação** e a **independência** estão relacionadas. Num sistema com base de dados central estas duas características estão bem salientes uma vez que os dados e metadados estão no mesmo repositório local. Nos sistemas com centralização de metainformação através de processos de ETL ou através do protocolo OAI-PMH a preservação e independência também

são elevadas, pois os metadados vão sendo colectados para um repositório central local. No sistema com pesquisa federada estas características não são controladas pelo sistema, umas vez que os dados estão remotamente nos repositórios das entidades aderentes.

A **facilidade de adesão** de uma entidade detentora à RPA prende-se com o cumprimento dos requisitos de adesão impostos pela entidade gestora da rede. A facilidade no cumprimento dos requisitos técnicos é influenciada pelo tipo de arquitectura e tecnologia utilizada pela rede.
A adesão de uma entidade num sistema com base de dados centralizada implica que as suas aplicações de gestão de arquivo utilizem a base de dados central como suporte de armazenamento e gestão de dados. Desta forma todas as entidades aderentes teriam que utilizar o mesmo sistema de interface com a base de dados, o que obrigaria as entidades aderentes a adaptarem os seus sistemas de informação para poderem comunicar com o repositório central, o que coloca em causa a independência de plataformas.
Num sistema com centralização de metainformação através de processos de ETL a adesão de uma entidade à RPA implica a implementação de um processo de ETL adaptado à estrutura de dados do seu repositório, o que é um processo custoso e de difícil gestão.
A adesão de uma entidade à RPA com sistema de pesquisa federada implica que essa entidade implemente um dos protocolos de comunicação (e.g. Z39.50, SRU) para responder aos pedidos de pesquisa. Se a entidade utilizar aplicações que ainda não suportem estes protocolos, a adaptação do seu sistema para responder a este requisito pode não ser trivial.
Se a arquitectura utilizada pela RPA for baseada num sistema com centralização de metainformação através do protocolo OAI-PMH a adesão de uma entidade à rede obriga a implementação do protocolo OAI-PMH, isto é, o seu repositório terá que ser um *data provider* que responda a pedidos OAI-PMH.

Após a análise comparativa dos vários modelos de arquitectura, depreende-se que o modelo que melhor se adequa ao cenário em estudo e consequentemente o mais fácil de implementar é a arquitectura baseada num sistema de recolha de metadados baseado no protocolo OAI-PMH.

2.6 Considerações finais

Ao longo deste capítulo foram estudados 4 modelos de arquitecturas, nomeadamente, o sistema com base de dados centralizada, o sistema com centralização de metainformação através de processos de ETL, o sistema com pesquisa federada e o sistema de recolha de metadados baseado no protocolo OAI-PMH.

Todos os modelos apresentam características que dotam o modelo com capacidade de ser instanciado com o cenário em estudo. Contudo, as características de alguns podem condicionar o sucesso e viabilidade do futuro sistema.

Após a apresentação das principais características de cada um dos modelos e análise comparativa entre eles, verificou-se que o modelo que melhor se adequará aos requisitos da rede portuguesa de arquivos será o sistema de recolha de metadados baseado no protocolo OAI-PMH. Esta arquitectura tem a vantagem de se reger por protocolos e normas, exigindo desta forma que os repositórios de dados que queiram aderir à rede, e posteriormente agregados pelo PPA usem um conjunto de directrizes e normas comuns, no sentido de garantir a interoperabilidade e qualidade dos resultados das pesquisas. Outra característica que coloca em vantagem esta arquitectura em relação às outras é o facto de a sua implementação ser independente de plataformas, o que facilita as entidades que queiram aderir à rede.

Pela aplicação das métricas ao modelo baseado no protocolo OAI-PMH, verifica-se que este modelo não compromete o cumprimento dos princípios da rede portuguesa de arquivos, isto é a integração estrutural, a neutralidade, a interoperabilidade, a pesquisa inter-repositórios, a acessibilidade e a qualidade.

Capítulo 3

Formatos de metainformação

Michael Fox, em certo artigo (Fox, 2001), faz uma distinção entre esquemas de estrutura de dados e esquemas de comunicação. Os primeiros definem *"o que se pode dizer"* a respeito de um documento, isto é, definem o formato da metainformação para descrever um documento. Os últimos correspondem a formatos através dos quais os metadados podem ser partilhados entre instituições. Muitos esquemas encaixam-se em mais que uma categoria, como é o caso do *Dubin Core*. Por outro lado, pode-se ver a ISAD(G) - *General International Standard Archival Description* (ICA, 2008) como um esquema de estrutura de dados, e o EAD - *Encoded Archival Description* (Library of Congress, 1998) como um formato de comunicação.

Neste capítulo apresenta-se uma descrição de alguns formatos de comunicação que servem para partilhar metadados entre as instituições dentro do cenário em estudo. Os formatos que foram estudados são: o EAD - *Encoded Archival Description*, o DC – *Dublin Core* e o ESE - *Europeana Semantic Elements* (Europeana, 2009a).

3.1 EAD - Encoded Archival Description

O EAD (*Encoded Archival Description*) (Library of Congress, 1998) 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 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.

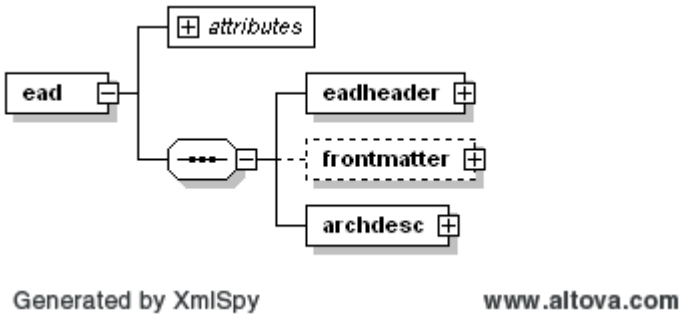

Figura 14 – Vista colapsada do *schema* do EAD

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.

Cada nível de descrição contém informação descritiva adequada, seguindo o modelo da ISAD(G) (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.

Na tabela seguinte (Tabela 2), são apresentados os elementos principais que constituem o EAD. Na primeira coluna apresenta-se o nome do elemento. Na segunda coluna apresenta-se a expressão XPath (W3C, 1999) a partir do segmento "ead/archdesc/". Este segmento deve preceder todas as expressões de forma a obter o caminho completo para obter o valor dos elementos ou atributos apresentados. Na terceira coluna faz-se a descrição de cada um dos elementos.

Nome do elemento	**Expressão XPath (ead/archdesc/…)**	**Descrição**
Abstract	did/abstract	Resumo
AccessRestrict	accessrestrict/p	Condições de acesso
Accruals	accruals/p	Ingressos adicionais
AcqInfo	acqinfo/p	Modalidades de aquisição
AltFormAvail	altformavail/p	Existência de cópias
Appraisal	appraisal/p	Avaliação, selecção e eliminação
Arrangement	arrangement/p	Organização e ordenação
BiogHist	bioghist/p	História administrativa/biográfica
CountryCode	did/unitid/@countrycode	Código do país
CustodHist	custodhist/p	História custodial
Dimensions	did/physdesc/dimensions	Dimensão e suporte
GenreForm	did/physdesc/genreform	Tipologia
GeogName	did/physdesc/geogname	Localidade
LangMaterial	did/langmaterial	Idioma/Escrita
LegalStatus	accessrestrict/legalstatus	Estatuto legal
MaterialSpec	did/materialspec	Detalhes específicos dos materiais
Note	note/p	Notas/observações
OriginalsLoc	originalsloc/p	Localização de originais
Origination	did/origination	Autores/produtores
OtherFindAid	otherfindaid/p	Instrumentos de pesquisa
OtherLevel	@otherlevel	Nível de descrição
PhysFacet	did/physdesc/physfacet	Aspecto físico
PhysLoc	did/physloc	Localização física
PhysTech	phystech/p	Características físicas e requisitos técnicos

PreferCite	prefercite/p	Citação
ProcessInfo	processinfo/p	Informações do processo
RelatedMaterial	relatedmaterial/p	Materiais relacionados
Repository	did/repository	Entidade detentora
RepositoryCode	did/unitid/@repositorycode	Código do repositório
ScopeContent	scopecontent/p	Âmbito e conteúdo
SeparatedMaterial	separatedmaterial/p	Material separado
UnitDate	did/unitdate	Datas
Unitid	did/unitid	Referência
UnitTitle	did/unittitle	Título do documento
UnitTitleType	did/unittitle/@type	Tipo título
UseRestrict	userestrict/p	Condições de reprodução
FilePlan	fileplan/p	Plano de classificação

Tabela 2 – Elementos do EAD

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)

A seguir são apresentadas duas imagens que ilustram partes do esquema do EAD. Na Figura 15 são mostrados os elementos de *archdesc* e na Figura 16 os elementos de *archdesc/did*.

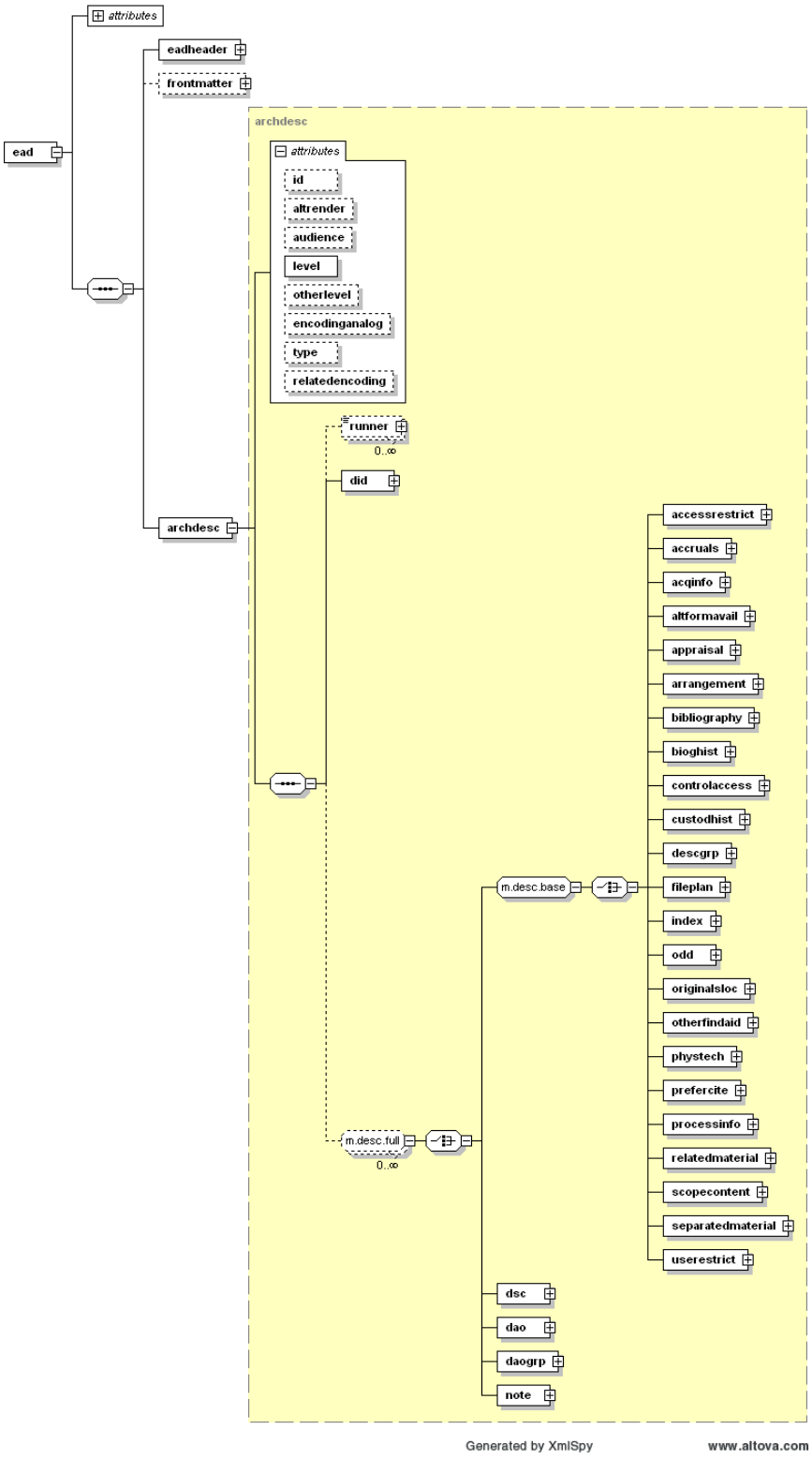

Figura 15 – Vista do *schema* do EAD com os elementos de *archdesc*

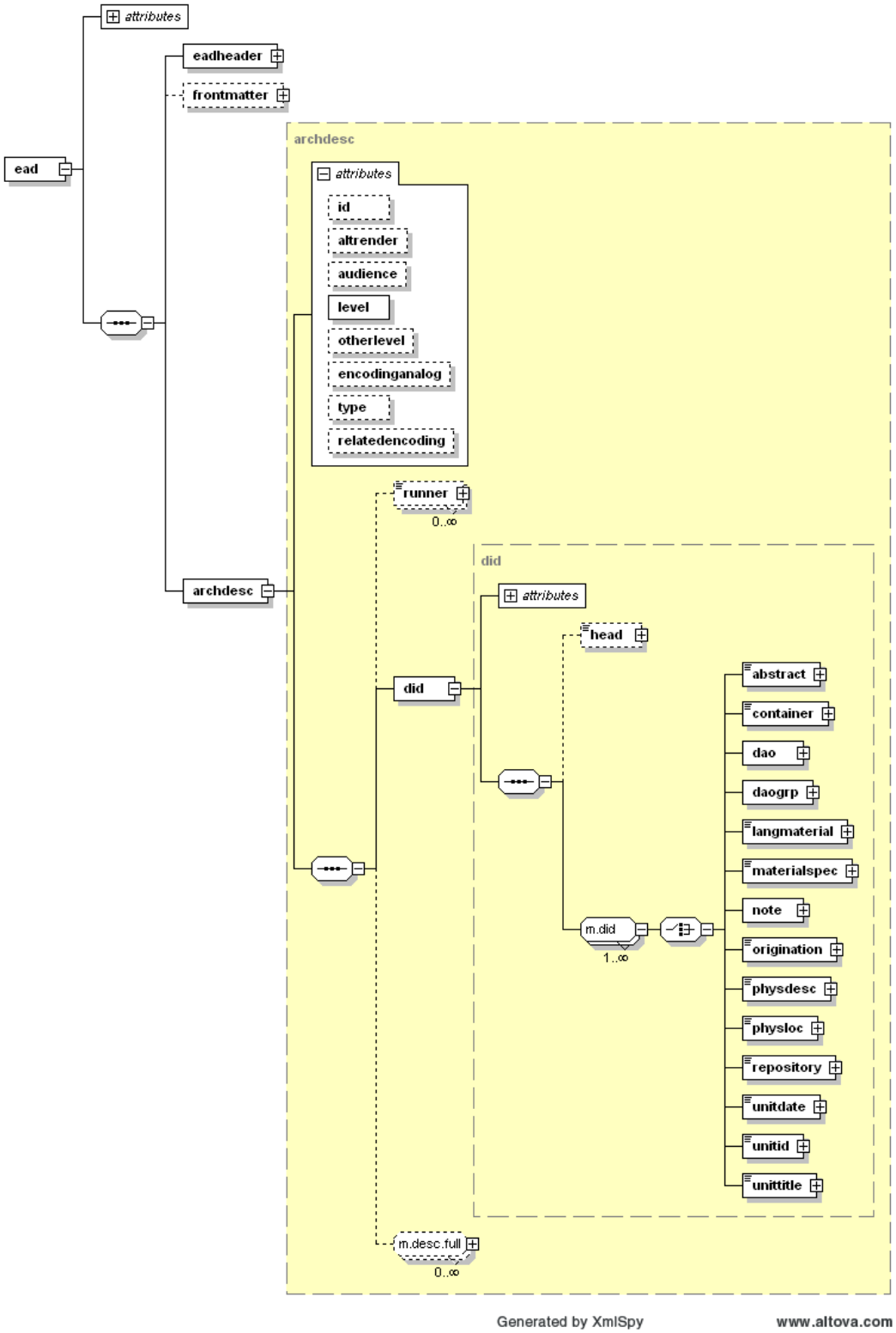

Figura 16 – Vista do *schema* do EAD com o elemento de *did* expandido

3.2 DC - Dublin Core

O Dublin Core (DC) (DCMI, 2009c) é um esquema de metadados que visa descrever recursos electrónicos, sejam eles textos, vídeos, imagens, sons, bases de dados ou websites. As características mais relevantes deste esquema são a simplicidade, a interoperabilidade semântica, o consenso internacional e a modularidade/extensibilidade de metadados na Web.

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 esquema. Esta reunião realizada em Dublin, Ohio, gerou a designação Dublin Core. Como resultado dos trabalhos levados a cabo, foram definidos um conjunto mínimo de elementos para a identificação dos objectos digitais.

Actualmente, o esquema 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).

O Dublin Core inclui dois níveis de especificação: o simples e o qualificado. O simples é constituído por um conjunto de 15 elementos (ver Tabela 3) conforme o DCMES – *Dublin Core Metadata Element Set* (DCMI, 2009b) e o qualificado incluí mais elementos e um conjunto de qualificadores que permitem refinar a semântica dos elementos e, assim, melhorar os níveis de recuperação da informação. Cada elemento é opcional e pode ser repetido.

Elemento	Descrição
Title	Nome pelo qual o recurso é conhecido
Creator	Entidade (individuo ou instituição) responsável pela existência do recurso
Subject	Tópicos do 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. Deve incluir a identificação das aplicações ou equipamento necessário para utilizar o recurso bem como as dimensões (tamanho e duração) do recurso
Identifier	Referência do recurso. Geralmente através de um sistema de identificação formal, como URL, DOI ou ISBN
Source	Referência ao recurso de onde o presente recurso 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 do recurso (direitos de autor, de propriedade intelectual ou outros direitos)

Tabela 3 – Elementos do Dublin Core simplificado

Existem uma série de recomendações de boas práticas associadas à utilização de alguns elementos. Por exemplo, no elemento Data recomenda-se a utilização do formato AAAA-MM-DD, de acordo com a norma ISO 8601 (W3C, 1997). Também é aconselhada a utilização de vocabulários controlados para o preenchimento dos elementos Tipo (e.g. *List of Resource Types* (DCMI, 1999)), Formato (e.g. *MIME Media Types* (IANA, 2007)) e Cobertura (e.g. *Thesaurus of Geographic Names* [TGN]). Relativamente ao elemento Língua, é recomendado o uso do RFC 1766 (IETF, 1995) e da ISSO 3166 (ISO, 2009), para as siglas da língua e respectivo país. De igual forma, é indicado como boa prática, a utilização de um sistema de identificação formal nos elementos Identificador, Fonte e Relação.

O Dublin Core qualificado (Tabela 4) é uma extensão do Dublin Core simplificado. Como tal, utiliza os 15 elementos referentes ao Dublin Core simplificado mais os elementos: *Audience* (Audiência), *Provenance* (Proveniência), *Rights holder* (Titular de direitos), *Instructional method* (Método de instrução), *Accrual method* (Método de ingestão), *Accrual periodicity* (Periodicidade de ingestão) e *Accrual policity* (Política de ingestão) (DCMI, 2008).

Elemento	Descrição
Audience	Conjunto de entidades para quem o recurso se destina. Pode ser determinada pelo criador, editor ou terceiros
Provenance	Alterações na propriedade do recurso, desde a sua criação, que são importantes para a sua autenticidade e integridade
Rights holder	Entidade proprietária ou gestora dos direitos sobre o recurso
Instructional method	Processo a que o recurso deve dar apoio (formas de apresentação de materiais pedagógicos, tipos de interacções pedagógicas)
Accrual method	Método pelo qual os itens são adicionados a uma colecção
Accrual periodicity	Frequência com que os itens são adicionados a uma colecção
Accrual policity	Política que rege a adição de itens a uma colecção

Tabela 4 – Elementos do Dublin Core qualificado

Para além destas extensões, o Dublin Core qualificado, utiliza uma série de qualificadores (Tabela 5) que especificam com maior precisão o significado do elemento que se pretende qualificar. Existem ainda os qualificadores de esquemas de codificação. Estes identificam os esquemas que interpretam os valores de cada elemento (DCMI, 2000).

Elemento	Qualificador	Esquema de codificação
Title	Alternative	-
Creator	-	-
Subject	-	LCSH, MeSH, DDC, LCC, UDC
Description	Table of Contents Abstract	-
Publisher	-	-
Contributor	-	-
Date	Created Valid Available Issued Modified	DCMI Period, W3C-DTF
Type	-	DCMI Type Vocabulary
Format	Extent Medium	IMT
Identifier	-	URI
Source	-	URI
Language	-	ISO 639-2, RFC 1766
Relation	Is version of Has version Is replaced by Replaces Is required by Requires Is part of Is referenced by References Is format of Has format	URI
Coverage	Spatial Temporal	DCMI Point, ISO 3166, DCMI Box, TGN DCMI Period, W3C-DTF
Rights	Access Rights License	URI

Audience	Mediator Education Level	-
Provenance	-	-
Rights holder	-	-
Instructional method	-	-
Accrual method	-	-
Accrual periodicity	-	-
Accrual policity	-	-

Tabela 5 – Qualificadores do Dublin Core qualificado

3.3 ESE - *Europeana Semantic Elements*

O conjunto de elementos semânticos da Europeana (Tabela 6) corresponde a um subconjunto de elementos do Dublin Core mais alguns elementos definidos pela Europeana.

A Europeana (Europeana, 2009b) é a biblioteca digital europeia, com o propósito de recolha, agregação, indexação e disponibilização de conteúdos bibliográficos, museológicos e arquivísticos. Foi um projecto que teve uma duração de 2 anos, tendo início em Julho de 2007. Disponibiliza um site protótipo dando aos utilizadores acesso directo a cerca de 2 milhões de objectos digitais, incluindo o material de filmes, fotografias, pinturas, sons, mapas, manuscritos, livros, jornais e documentos de arquivo. O protótipo foi lançado em Novembro de 2008 por *Viviane Reding*, Comissária Europeia para a Sociedade da Informação e Media.

Fonte	Elemento	Qualificador	Descrição
DC	*Title*	Alternative	Nome pelo qual o recurso é conhecido
DC	*Creator*	-	Entidade (individuo ou instituição) responsável pela existência do recurso
DC	*Subject*		Tópicos do conteúdo do recurso
DC	*Description*	tableOfContents	Descrição do conteúdo do recurso
DC	*Publisher*		Entidade responsável por tornar o recurso acessível
DC	*Contributor*		Entidade responsável por qualquer contribuição para o conteúdo do recurso
DC	*Date*	created; issued	Data associada a um evento do ciclo de vida do recurso
DC	*Type*		Natureza do conteúdo do recurso

Europeana			
DC	*Format*		Manifestação física ou digital do recurso. Deve incluir a identificação das aplicações ou equipamento necessário para utilizar o recurso bem como as dimensões (tamanho e duração) do recurso
DC	*Identifier*		Referência do recurso. Geralmente através de um sistema de identificação formal, como URL, DOI ou ISBN
DC	*Source*		Referência ao recurso de onde o presente recurso possa ter derivado
DC	*Language*		Idioma do conteúdo intelectual do recurso
DC Europeana	*Relation*	isVersionOf; hasVersion; isReplacedBy; replaces; isRequiredBy; requires; isPartOf; hasPart; isReferencedBy; references; isFormatOf; hasFormat; conformsTo isShownBy; isShownAt	Referência a um recurso relacionado
DC	*Coverage*	spatial; temporal	Extensão ou alcance do recurso
DC	*Rights*		Informação sobre os direitos do recurso (direitos de autor, de propriedade intelectual ou outros direitos)
DC terms	*Provenance*		Alterações na propriedade do recurso, desde a sua criação, que são importantes para a sua autenticidade e integridade
Europeana	*userTag*		Marca criada por um utilizador através da interface Europeana
Europeana	*Unstored*		Este elemento contém todas as informações relevantes que não podem ser mapeados para um outro elemento na ESE
Europeana	*Object*		URL (não URI) que faz referência ao objecto digital no site do fornecedor de conteúdo, para gerar uma imagem miniatura ou amostra. Este elemento geralmente mapeia para <europeana:isShownBy>
Europeana	*Language*		Idioma do recurso
Europeana	*Provider*		Nome da organização que detém o objecto digital
Europeana	*Type*		Natureza ou género do recurso. O tipo inclui termos que descrevem categorias, funções, géneros, ou níveis de agregação de conteúdos. É recomendado seleccionar um valor a partir de um vocabulário

			controlado (http://dublincore.org/documents/dcmi-type-vocabulary/)
Europeana	*Uri*		URI para o recurso, dentro do contexto Europeana
Europeana	*Year*		Ponto ou período de tempo associado a um acontecimento na vida do objecto original analógico ou digital
Europeana	*hasObject*		Indica a disponibilidade de miniaturas de objectos digitais para serem processadas pelo sistema da Europeana. Dois valores são permitidos: True ou False. O valor corresponde a <europeana:object>.
Europeana	*Country*		Nome do país em que o fornecedor de conteúdo se baseia ou "Europa" no caso de projectos à escala europeia.

Tabela 6 – Elementos semânticos da Europeana

3.4 Considerações finais

Neste capítulo foram estudados 3 formatos de metainformação que podem suportar a partilha de metadados entre as entidades que participam no sistema. Foram estudados formatos com naturezas e características diferentes, entre eles, o EAD, o DC e o ESE.

O EAD representa a natureza complexa e hierárquica da informação em arquivo por níveis de descrição. Desta forma, os registos de metainformação estão descritos numa estrutura hierárquica organizada por contextos. Este formato apresenta uma complexidade e variantes de utilização que podem tornar complexa a comunicação entre as entidades do sistema.

Pelo contrário, o Dublin Core é caracterizado pela sua simplicidade, interoperabilidade semântica e consenso internacional. No entanto este formato não é suficiente para suportar a descrição de informação de arquivo.

O conjunto de elementos semânticos da Europeana corresponde a um subconjunto de elementos do Dublin Core mais alguns elementos definidos pela Europeana. Da mesma forma que o Dublin Core, este formato não é apropriado para captar de forma integral a informação necessária das descrições arquivísticas.

Estes diferentes formatos levantam algumas problemáticas quando são utilizados para a transferência de informação entre as entidades da rede. No capítulo seguinte são apresentadas possíveis resoluções e consensos.

Capítulo 4
Arquitectura do sistema

Neste capítulo descreve-se a arquitectura e funcionamento do sistema que será implementado no contexto da rede portuguesa de arquivos. A arquitectura proposta é resultado do estudo e análise dos modelos de arquitectura e dos formatos para transferência de metainformação entre as entidades participantes no sistema.

Conforme a análise comparativa efectuada no capítulo 2 - Modelos de arquitecturas, chegou-se à conclusão que o sistema de recolha de metainformação baseado no protocolo OAI-PMH é o que melhor cumpre os requisitos impostos.

Os formatos de metainformação que vão ser utilizados pelo protocolo OAI-PMH para a recolha de informação são: o EAD (*Encoded Archival Description*), o DC (*Dublin Core*) e o ESE (*Europeana Semantic Elements*). Estes diferentes formatos levantam algumas problemáticas quando são utilizados pelo protocolo OAI-PMH. Neste capítulo são apresentadas possíveis resoluções e consensos.

4.1 Recolha de metainformação no formato EAD

A Figura 17 mostra um esquema simplificado da arquitectura do sistema seguindo o protocolo OAI-PMH para recolha de metadados no formato EAD e posterior armazenamento centralizado para o fornecimento de serviços. O protocolo prevê, como mostra no esquema, dois tipos principais de intervenientes: os *data providers* e os *service providers*. Neste caso, os *data providers* correspondem a repositórios digitais de entidades de arquivo que disponibilizam a metainformação dos documentos custodiados. Para garantir a interoperabilidade, os *data providers* devem disponibilizar os seus metadados segundo formatos normalizados, neste caso a norma de descrição arquivística EAD. O *service provider* oferece serviços a partir dos dados colectados e armazenados num repositório central. A recolha dos dados é feita pelo módulo de agregação pelo envio de pedidos OAI-PMH aos diversos *data providers*, que de acordo com a requisição solicitada, enviam como resposta os metadados solicitados em formato XML EAD. Os metadados colectados pelo módulo anterior são posteriormente integrados e armazenados no repositório central (base de dados MySQL).

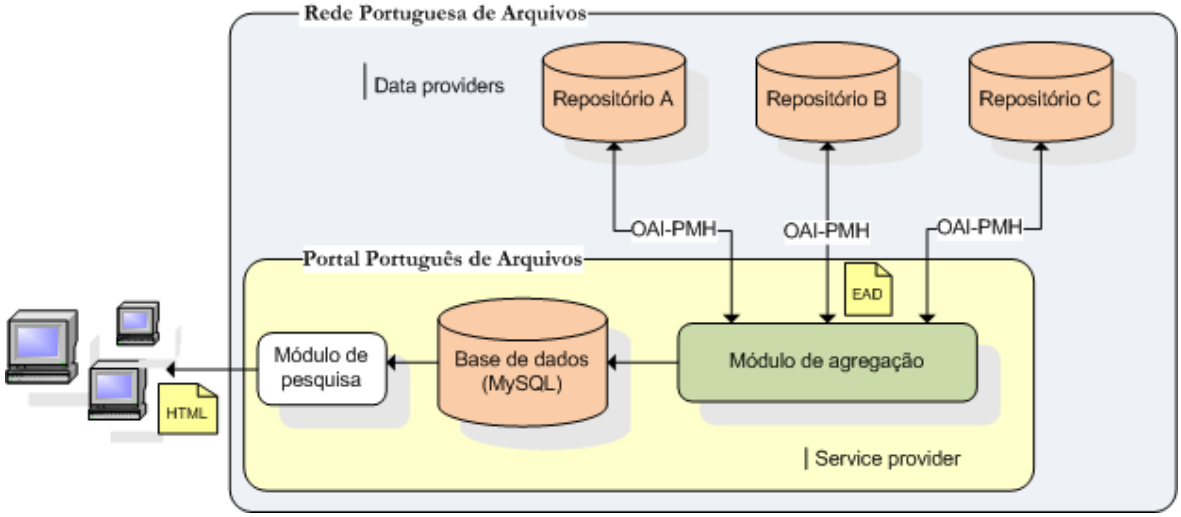

Figura 17 – Arquitectura do sistema com recolha de metainformação no formato EAD

O EAD é um formato que representa a natureza complexa e hierárquica da informação em arquivo por níveis de descrição. Desta forma, os registos de metainformação estão descritos

numa estrutura hierárquica organizada por contextos, do mais geral para o mais particular. Um determinado registo de descrição só tem sentido quando enquadrado no respectivo contexto hierárquico, onde a sua descrição é complementada pelos registos ascendentes da hierarquia até ao nível da raiz – nível de descrição fundo.

Visto que um registo do EAD só está totalmente descrito quando se tem em conta toda hierarquia a que pertence, surge de imediato uma questão. Como efectuar a recolha de registos EAD pelo protocolo OAI-PMH?

Este problema pode ser resolvido de 3 maneiras diferentes: recolher o fundo completo ao qual pertence esse registo, recolher o ramo da árvore desde o fundo até ao registo alterado, ou recolher apenas o registo alterado (Ferros, Ramalho, & Ferreira, 2008). Descreve-se a seguir o funcionamento de cada uma das soluções possíveis.

a) Recolher o fundo completo ao qual pertence esse registo

Esta solução é simples de implementar, uma vez que a resposta ao pedido apenas consiste em enviar o fundo respectivo após seleccionado e extraído a partir do *data provider*. O *service provider* apenas tem que fazer a integração do fundo recebido no repositório central, que se resume a substituir o antigo por este. Como se depreende facilmente, é uma solução muito pouco eficiente, pois uma simples alteração ou inserção de um novo registo num determinado fundo iria desencadear uma posterior recolha do fundo completo. Desta forma, alterações de registos em diversos fundos, os quais podem ter um número elevado de registos (na ordem das centenas de milhar), iria desencadear uma transferência de dados muito elevada.

b) 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 desses registos nos *service providers* mostra-se mais complexa. A extracção passa por seleccionar os registos ascendentes do registo pedido na árvore do fundo. A resposta ao pedido corresponde, então, a uma parte do fundo com esses registos. Por conseguinte, a integração será efectuada pela substituição no respectivo fundo dos registos antigos pelos registos recebidos respectivos.

c) 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 – fundo. Contudo, o *schema* do EAD não verifica este invariante ou condição para 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 da referência completa de um registo. Desta forma, a existência destes campos na referência completa de um registo e a unicidade das referências dentro de um dado repositório, garantem a unicidade das referências no universo dos repositórios que disponibilizam EAD.

Assim, a tarefa de *harvesting* 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 válidos pelo *schema* EAD.

4.2 Recolha de metainformação no formato DC

Para aumentar a interoperabilidade entre repositórios, os repositórios devem disponibilizar outros formatos de metainformação para além do EAD. Para tal, os repositórios devem disponibilizar o formato DC (DCMI, 2009c), visto ser um formato de metainformação simples, interoperável e largamente utilizado no contexto do OAI-PMH.

A Figura 18 mostra um esquema simplificado da arquitectura do sistema seguindo o protocolo OAI-PMH para recolha de metadados no formato DC e posterior armazenamento centralizado para o fornecimento de serviços.

Como é mostrado no esquema, os *data providers* correspondem a repositórios digitais de entidades detentoras de arquivo que disponibilizam a metainformação dos documentos custodiados. Tendo em conta que o formato de metainformação utilizado pela maioria destes repositórios é o EAD, torna-se necessário a definição e implementação de um transformador ou *crosswalk* de EAD para DC.

O *service provider* oferece serviços a partir dos dados armazenados no repositório central. A recolha dos dados é feita pelo módulo de agregação pelo envio de pedidos OAI-PMH aos diversos *data providers*, que de acordo com a requisição solicitada, enviam como resposta os metadados solicitados em formato DC. Os metadados colectados pelo módulo anterior são posteriormente integrados e armazenados na base de dados central que alimenta o módulo de pesquisa e referência do Portal Português de Arquivos.

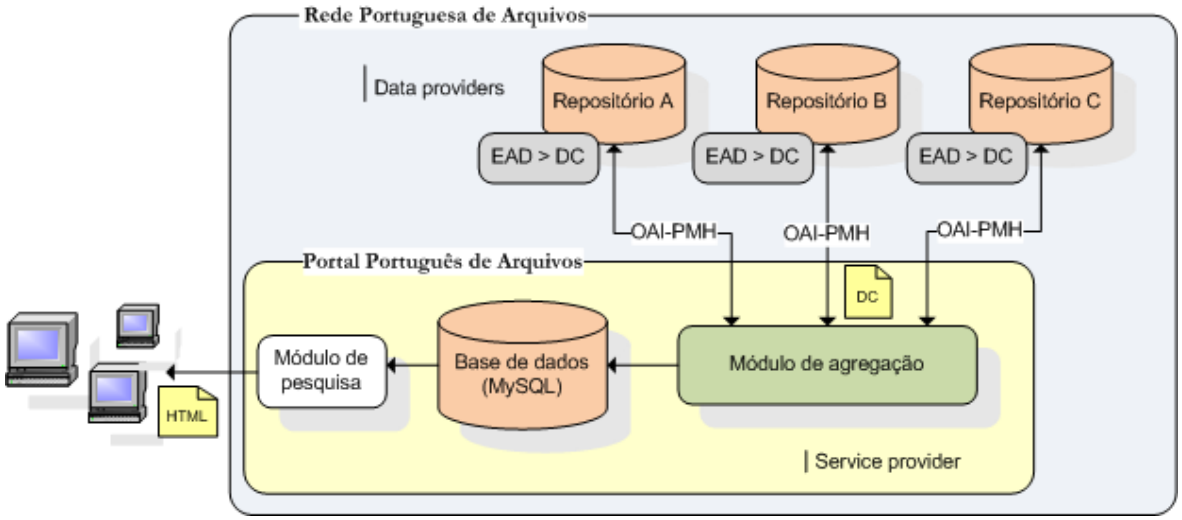

Figura 18 – Arquitectura do sistema com recolha de metainformação no formato DC

O formato de metainformação DC prima pela simplicidade, interoperabilidade semântica e consenso internacional. Em termos de comunicação e interoperabilidade, este formato apresenta as características óptimas para ser utilizado na transferência de informação entre as entidades que participam na rede. 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 consentâneo com a realidade da descrição arquivística, pois não possibilita uma representação hierárquica da informação em arquivo por níveis de descrição.

Como foi descrito atrás, aquando a apresentação dos formatos de metainformação, viu-se que o conjunto dos elementos do formato DC é um subconjunto do formato ESE. Desta forma, o mapeamento descrito na secção seguinte (ver Tabela 7) entre o formato EAD e ESE também capta o mapeamento entre o formato de metainformação EAD e DC.

4.3 Recolha de metainformação no formato ESE

A Figura 19 mostra um esquema simplificado da arquitectura do sistema seguindo o protocolo OAI-PMH para recolha de metadados no formato ESE – *Europeana Semantic Elements* e posterior armazenamento centralizado para o fornecimento de serviços.

Como é mostrado no esquema, os *data providers* correspondem a repositórios digitais de arquivos que disponibilizam a metainformação dos documentos custodiados. Uma vez que este tipo de repositórios segue, normalmente, o formato EAD para o armazenamento de informação, os *data providers* deverão dispor de um transformador ou *crosswalk* para transformar EAD em ESE.

O *service provider* oferece serviços a partir dos dados armazenados no repositório central. A recolha dos dados é feita pelo módulo de agregação pelo envio de pedidos OAI-PMH aos diversos *data providers*, que de acordo com a requisição solicitada, enviam como resposta os metadados solicitados em formato ESE. Os metadados colectados pelo módulo anterior são posteriormente integrados e armazenados na base de dados central, a qual alimenta o módulo de pesquisa do PPA.

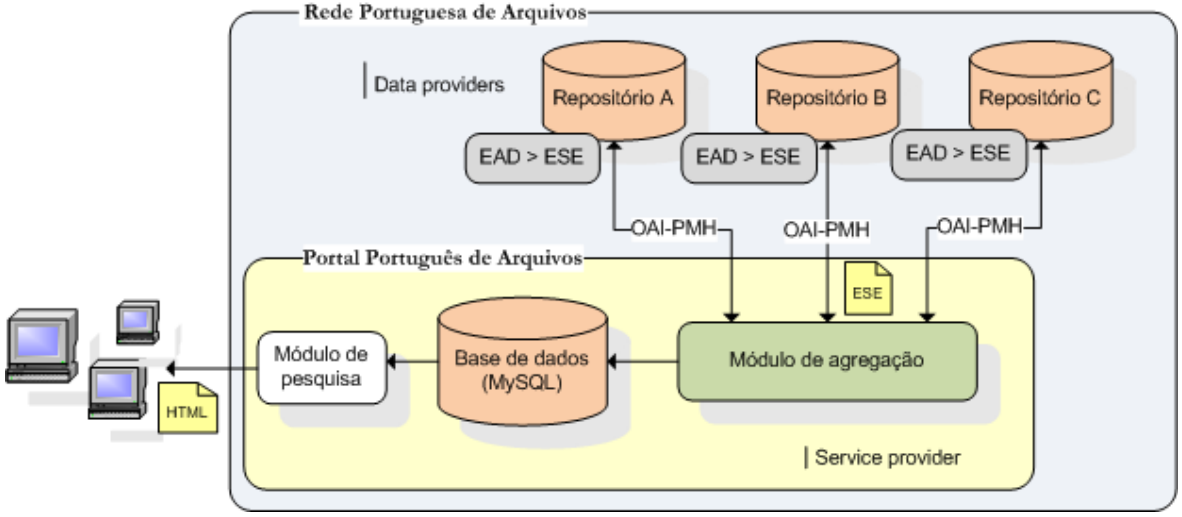

Figura 19 – Arquitectura do sistema com recolha de metainformação no formato ESE

Dado que o formato de metainformação utilizado pelas entidades detentoras de conteúdos de arquivo é, normalmente, o EAD, existe a necessidade de mapear a informação para o formato ESE. A Tabela 7 mostra o mapeamento entre o formato de metainformação EAD e ESE. Nesta tabela apresenta-se o elemento EAD e o seu correspondente no esquema ESE. Como já foi referido atrás, o conjunto de elementos do formato DC é um subconjunto do formato ESE, o que torna o mapeamento descrito a seguir também válido para o formato DC.

Descrição	EAD	ESSE	Notas
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	Apenas se for fundo
Código do país	CountryCode	europeana.country	
História custodial	CustodHist	dc.provenance	
Dimensão e suporte	Dimensions	dc.format.extent	
Tipologia	GenreForm	dc.format.medium	Apenas para suportes físicos
Localidade	GeogName	-	
Idioma/Escrita	LangMaterial	dc.language	Em ISO-639-1
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	-	Ver DSpace
Informações do processo	ProcessInfo	-	
Materiais relacionados	RelatedMaterial	dc.relation.refences	
Entidade detentora	Repository	dc.publisher	

Código do repositório	RepositoryCode	-	Pode-se acrescentar ao dc.publisher
Âmbito e conteúdo	ScopeContent	dc.subject	
Material separado	SeparatedMaterial	dc.relation.hasPart	?
Datas	UnitDate	dc.date	
Referência	Unitid	dc.identifier	Colocar também o URL (como 1º elemento)
Título do documento	UnitTitle	dc.title	
Tipo título	UnitTitleType	-	Açúcar semântico para o dc.title
Condições de reprodução	UseRestrict	dc.rights	
Plano de classificação	FilePlan	-	
-	-	europena.type	Sempre image
-	-	europeana.uri	dc.identifier
-	-	europeana.year	year(dc.date)
-	-	europeana.provider	dc.publisher
-	-	europeana.hasObject	hasDigitalObject()
-	-	europeana.object	firstUrl(DigitalObject)
-	-	europeana.userTag	-
-	-	europeana.language	-
-	-	europeana.country	-
-	-	europeana.isShownBy	-
-	-	europeana.isShownAt	Url(DigitalObject)

Tabela 7 – Mapeamento de EAD para ESE

O processo de mapeamento descrito na tabela anterior revelou dificuldades ao nível da inexistência por parte da ESE (e 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 ESE 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, ainda torna o problema mais crítico.

O exercício de mapear EAD em ESE não faz sentido quando se pretende mapear todos os campos do EAD, uma vez que a ESE não contempla todos os elementos relevantes para a descrição arquivística. Em vez disso, os *data providers* devem disponibilizar directamente EAD Component (um registo do EAD), conforme descrito na secção 4.1, onde se descreve a

recolha de metainformação no formato EAD. Desta forma, o procedimento será mais consentâneo com a realidade da descrição arquivística.

Também se constata que não existe vantagem de mapear EAD em ESE para a disponibilização de informação por parte dos *data providers* para o PPA, mas apenas no momento da recolha de informação por parte da Europeana no PPA.

Após alguma ponderação concluiu-se da não necessidade de recolha de todos os elementos descritivos patentes do EAD, processo esse que seria consideravelmente complicado conforme o que acima foi referido. No entanto a eficácia e propósitos do PPA não serão comprometidos, visto que o portal não se pretende substituir aos repositórios, funcionando antes como uma plataforma de direccionamento dos utilizadores para os vários repositórios aderentes. Assim considerou-se que apenas seria necessário recolher os seguintes dados:

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

Este conjunto de elementos são suficientes para dar ao utilizador uma percepção completa dos registos que lhe interessam, podendo este navegar directamente a partir do registo até ao repositório da entidade detentora respectiva e aí visualizar toda a informação existente relativa a esse registo, assim como consultar o contexto descritivo onde o registo está situado.

Como todos estes elementos são directamente transponíveis para campos do DC (simplificado) (ver Tabela 8), o problema ficou aparentemente resolvido, não sendo necessário mapear todos os campos do EAD em ESE (ou DC). Considerado este facto, o exercício de mapeamento partindo de todos os elementos do EAD foi abandonado por ser desnecessário.

Descrição	EAD	DC	Notas	Obrigatoriedade
Referência	Unitid	dc.identifier	Colocar URL como 1º elemento.	Sim
Título do documento	UnitTitle	dc.title		Sim
Datas	UnitDate	dc.date		Sim
Nível de descrição	OtherLevel	dc.type		Sim
Entidade detentora	Repository	dc.publisher		Sim
Dimensão e suporte	Dimensions	dc.format.extent		Não
Âmbito e conteúdo	ScopeContent	dc.subject		Não
História custodial	CustodHist	dc.provenance		Não
-	-	dc.relation	URL para miniatura	Não

Tabela 8 – Mapeamento de EAD para DC simplificado

A recolha de metainformação no formato DC pode ser considerado um formato adequado uma vez que cumpre e resolve os requisitos impostos pelo PPA. Isto porque:

- Apesar de, geralmente, a descrição arquivística seguir esquemas mais complexos, como o modelo ISAD(G) ou o EAD, o Dublin Core apresenta-se como um formato suficiente para alimentar o PPA e os seus serviços de pesquisa;
- Trata-se de um esquema simples e que pode ser facilmente implementado por qualquer repositório de dados;
- É interoperável, na medida em que os dados são expostos num formato padrão utilizado pela maioria dos repositórios.

4.4 Arquitectura detalhada do PPA

O PPA na terminologia do OAI-PMH corresponde ao *service provider*, o qual oferece serviços de consulta e referência sobre os dados das entidades aderentes colectados. Para prestar este serviço é necessário reunir um conjunto de módulos que implementem os requisitos funcionais de acordo com a sua função e público-alvo.

A Figura 20 apresenta os vários módulos funcionais que constituem o PPA, nomeadamente, o módulo de validação, módulo de agregação, módulo de registo, módulo de pesquisa e módulo

de administração. A figura ilustra também a interacção com os *data providers*, que correspondem aos repositórios das entidades aderentes da rede.

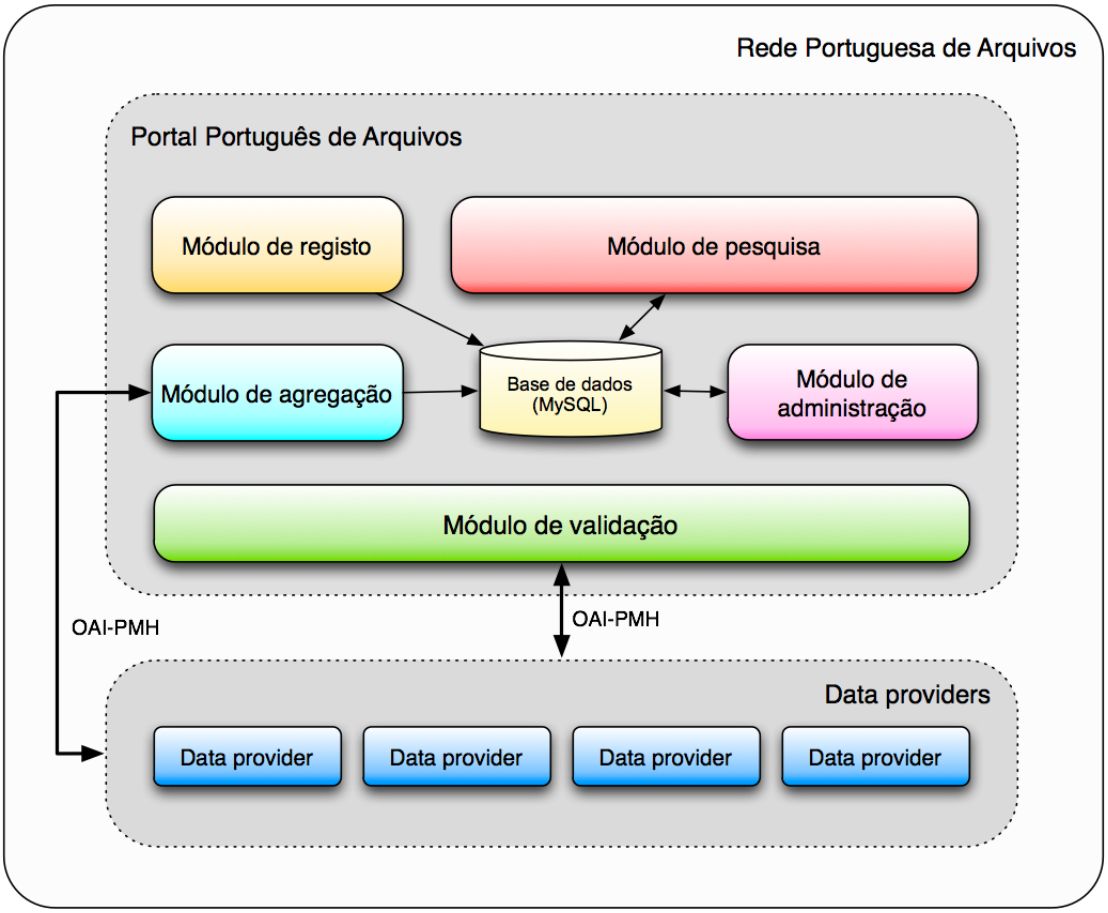

Figura 20 - Arquitectura detalhada do PPA

4.4.1 *Data provider*

O PPA será alimentado com informação proveniente de vários repositórios geograficamente distribuídos. Os repositórios que desejarem aderir à rede de arquivos deverão cumprir as directrizes definidas no âmbito do projecto. O cumprimento dessas directrizes por parte das entidades aderentes será verificado periodicamente recorrendo a uma ferramenta desenvolvida no âmbito do projecto, designada por Módulo de validação.

O repositório aderente deverá ter a capacidade de interoperar com o PPA respeitando o esquema de metainformação definido nas directrizes. Para esse efeito, um repositório aderente deverá dispor de *crosswalks* (i.e. mapeamentos) que assegurem a equivalência entre esquemas de metainformação que utilizam e os esquemas exigidos pelo PPA.

O incumprimento das directrizes definidas no âmbito do projecto implicará a exclusão do repositório da rede portuguesa de arquivos.

Foi desenvolvido, no âmbito desta dissertação, um módulo para que o *software* de gestão de arquivo definitivo DigitArq (Ferreira & Ramalho, 2004a, 2004b; Ramalho, Ferreira, Ferros, Lima, & Sousa, 2006) tenha a capacidade de *data provider*. A inclusão deste módulo faz com que este *software* seja compatível com o Portal Português de Arquivos, assim como com outros agregadores de carácter internacional como a Europeana ou o Portal de Arquivos Europeu (ainda em desenvolvimento) (APEnet, 2009), respeitando na sua globalidade as directrizes em vigor e tornando as instituições que o utilizam prontas para adesão à Rede Portuguesa de Arquivos (DGARQ, 2008a, 2008b).

O módulo desenvolvido suporta dois esquemas de metainformação: oai_dc (Dublin Core simplificado) e ESE. Um protótipo funcional do mesmo pode ser consultado em http://digitarq.keep.pt/oai/.

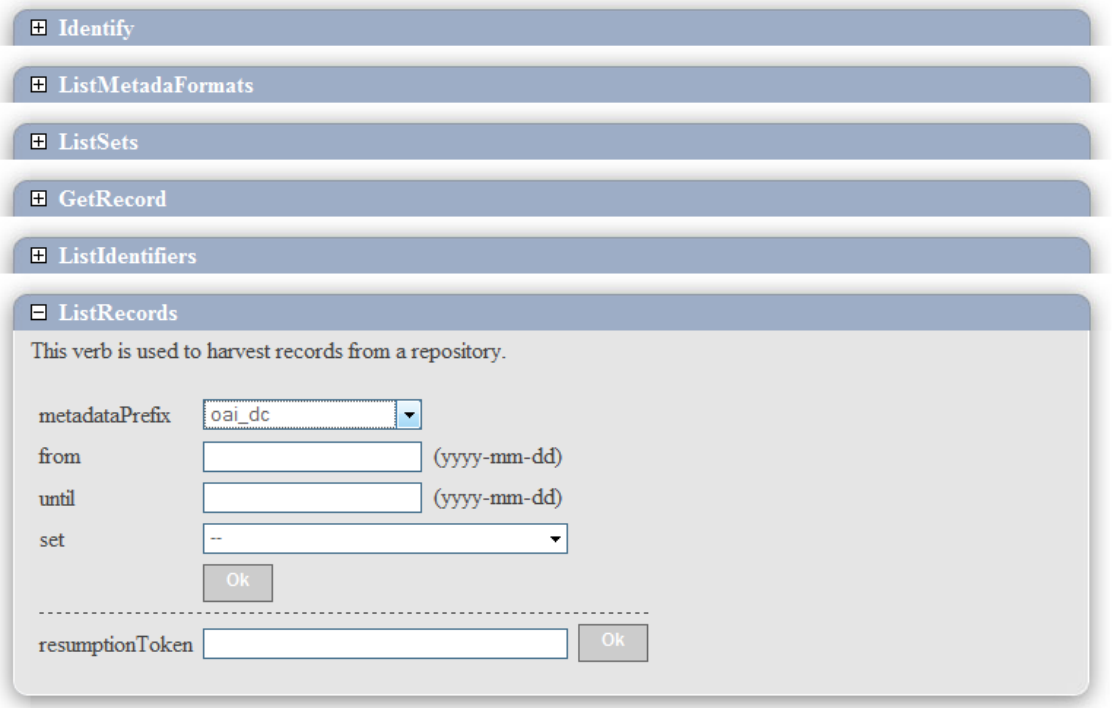

Figura 21 - Módulo OAI-PMH do DigitArq

4.4.2 Módulo de registo

Este módulo permite a qualquer entidade que pretenda aderir à rede manifestar essa intenção preenchendo um formulário de adesão. Nesse formulário deverá preencher um conjunto de elementos identificativos e técnicos que serão definidos ao longo do projecto.

4.4.3 Módulo de validação

De modo a suportar/apoiar a adesão das instituições ao PPA, será desenvolvido um sistema de validação de repositórios (i.e. *data providers*). Este sistema assenta no registo de informação sobre o repositório a validar, nomeadamente, o nome do repositório, a instituição que o administra e quais as suas interfaces Web e OAI.

Depois de efectuado o registo, um processo assíncrono é responsável por recolher todos os metadados do repositório e validá-lo seguindo um conjunto de directrizes definidas no âmbito do projecto.

Após a validação, será produzido um relatório contendo, para além de uma listagem de todas as anomalias encontradas, um conjunto de estatísticas que poderão ser úteis ao gestor do repositório. Após este processo, o relatório é enviado por e-mail para quem solicitou a validação.

4.4.4 Módulo de agregação

O módulo de agregação é responsável por recolher periodicamente a metainformação produzida pelas entidades aderentes. A agregação (i.e. *harvest*) será realizada de acordo com o protocolo OAI-PMH. A periodicidade da recolha será configurada de acordo com as necessidades de actualização do PPA.

Faz também parte deste processo o processamento da metainformação recolhida e a sua adaptação de modo a alimentar adequadamente o Módulo de pesquisa do PPA.

4.4.5 Módulo de pesquisa

O PPA incorpora um módulo de pesquisa que irá permitir a localização e recuperação de metainformação de acordo com os elementos apresentados na secção 1.1 deste documento.

Este módulo permite ao utilizador recuperar conteúdos, i.e. as representações digitais dos documentos descritos, desde que estes se encontrem em linha e estejam associados à metainformação recolhida. Por exemplo, caso haja imagens associadas a um registo descritivo de um documento, essas imagens serão recuperáveis através do portal de pesquisa, não de forma directa, mas permitindo uma ligação ao repositório que detém os dados e às suas interfaces de visualização de conteúdos.

O módulo de pesquisa irá permitir ao utilizador realizar pesquisas inter-repositórios ou apenas em alguns repositórios. Por exemplo, se um utilizador quiser apenas pesquisar no Arquivo Distrital do Porto deve poder fazê-lo retornando apenas metainformação desse repositório. Se quiser pesquisar no Arquivo Distrital de Aveiro e na Câmara Municipal do Corvo, deve poder fazê-lo. Se quiser pesquisar em todos os repositórios simultaneamente, também poderá fazê-lo.

O módulo de pesquisa permitirá ao utilizador recuperar apenas a metainformação que contenha conteúdos associados. A Figura 22 apresenta o mapa de navegação no Portal de pesquisa.

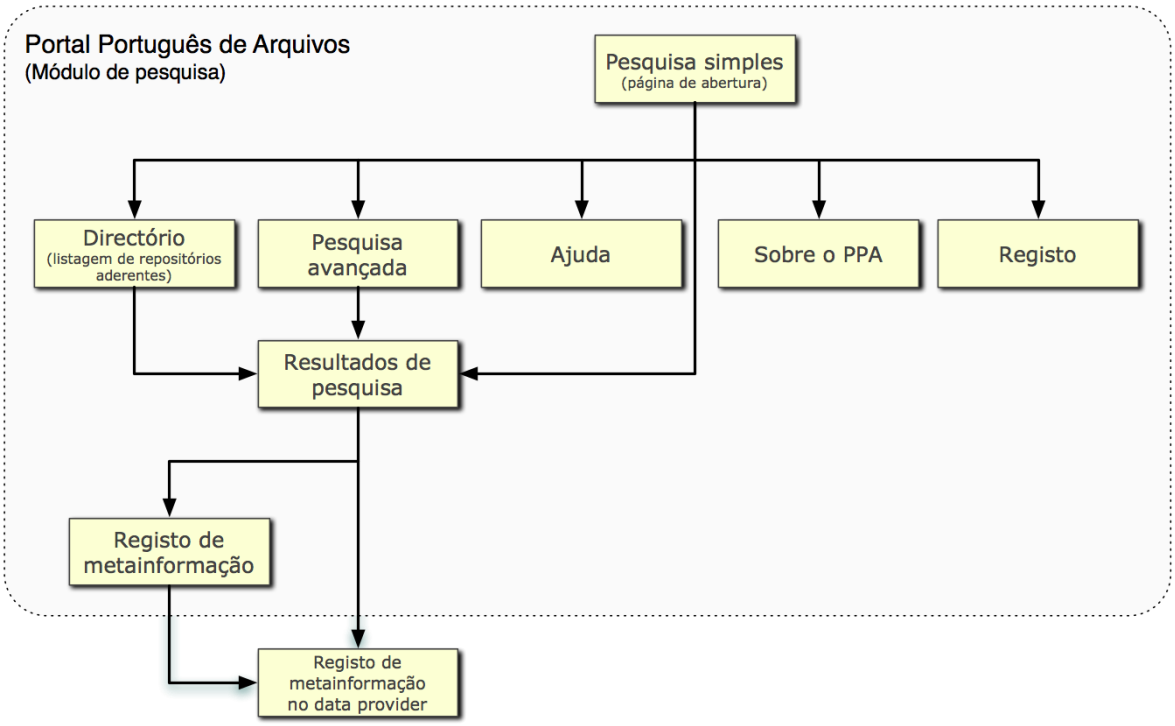

Figura 22 – Mapa de navegação no Portal Português de Arquivos

4.4.6 Módulo de administração

O PPA será acompanhado de um módulo de administração que permitirá ao gestor do portal realizar as seguintes operações:

1. Consultar estatísticas de agregação;
2. Consultar relatórios de problemas detectados durante a agregação;
3. Listar e adicionar novos repositórios para agregação e disponibilização nas interfaces de pesquisa;
4. Iniciar a agregação de novos repositórios.

4.5 Integração com a Europeana

A Europeana é a biblioteca digital europeia, com o propósito de recolha, agregação, indexação e disponibilização de conteúdos bibliográficos, museológicos e arquivísticos. Foi um projecto que teve uma duração de 2 anos, tendo início em Julho de 2007. Disponibiliza um site protótipo, dando aos utilizadores acesso directo a cerca de 2 milhões de objectos digitais, incluindo o material de filmes, fotografias, pinturas, sons, mapas, manuscritos, livros, jornais e documentos de arquivo. O protótipo foi lançado em Novembro de 2008 por *Viviane Reding*, Comissária Europeia para a Sociedade da Informação e Media.

Um dos objectivos da RPA é a sua integração com a Europeana. Para isso, o repositório central da RPA que serve de suporte ao PPA na qualidade de *service provider*, terá também a função de *data provider* na disponibilização de informação para a Europeana.

Conforme se mostra na Figura 23, o *service provider* da Europeana ao enviar um pedido OAI-PMH ao *data provider* do PPA, este responde ao pedido conforme o protocolo.

É claro que a participação da RPA na Europena implica o cumprimento dos requisitos de adesão impostos pela Europeana. Um dos requisitos técnicos de grande importância diz respeito aos protocolos e formatos suportados na disponibilização de informação. Foi neste sentido que um dos formatos de metainformação disponibilizados pelos *data providers* da RPA é a ESE. O formato ESE também será disponibilizado pelo *data provider* do PPA.

A integração do PPA com a Europeana oferece de forma imediata a participação das entidades aderentes da RPA (entidades detentoras de conteúdo de arquivo) na Europeana, sem terem de ser aderentes directos da Europeana, pois o *service provider* do PPA ao ter a função de *data provider* para a Europeana está implicitamente a disponibilizar os dados de todas as entidades aderentes da RPA à Europeana.

Por razões de direitos de informação, as entidades aderentes da RPA deverão ter a possibilidade de controlar qual a informação que desejam disponibilizar à Europeana, mesmo que essa informação seja disponibilizada ao PPA.

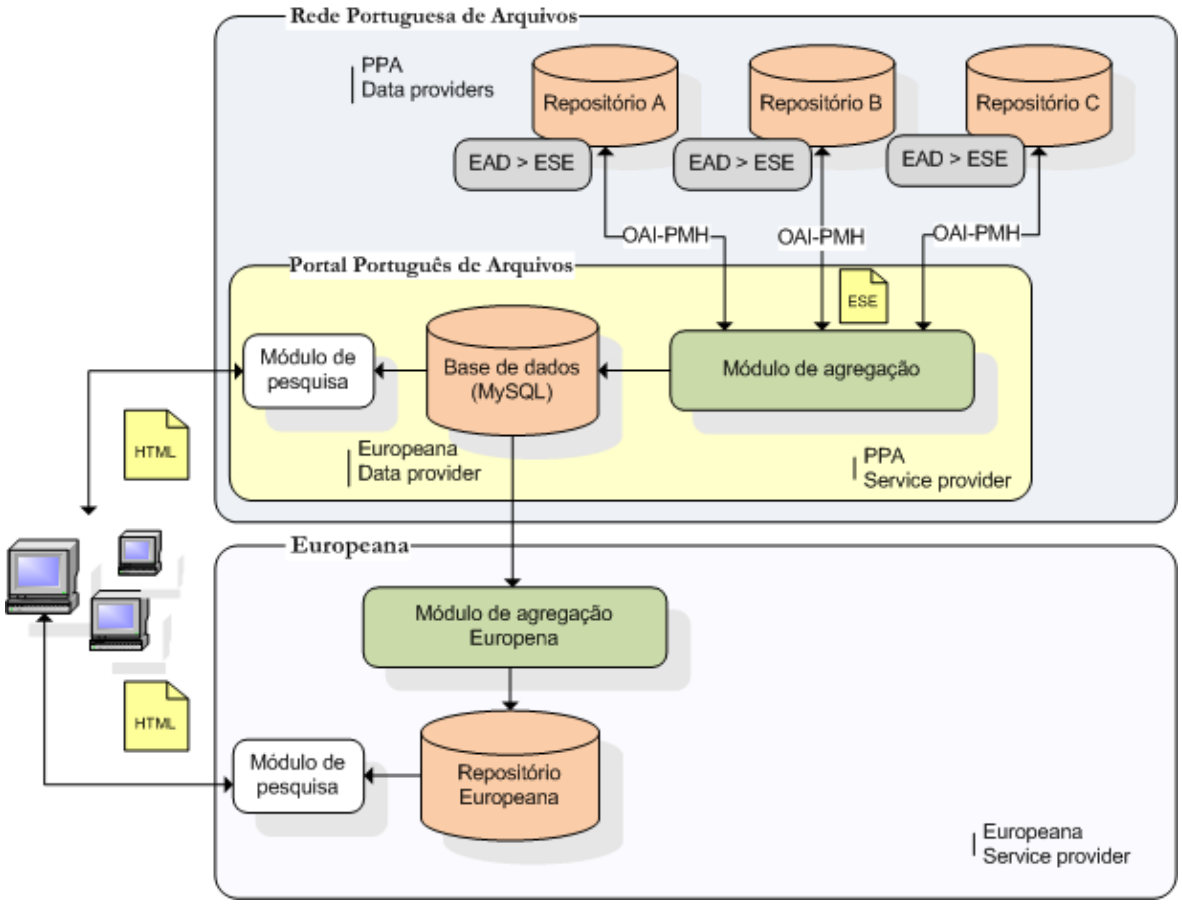

Figura 23 – Integração do PPA com a Europeana

4.6 Considerações finais

Neste capítulo apresentaram-se algumas resoluções e consensos para tornear algumas problemáticas quando são utilizados alguns formatos de metainformação pelo protocolo OAI-PMH.

Viu-se que o processo de recolha de metadados no formato EAD poderá ser difícil de implementar, apesar de não ser necessária a implementação de *crosswalks* para mapear os dados armazenados em EAD, pois este formato é, normalmente, já o formato de metainformação utilizado pelas entidades detentoras de arquivo. Contudo a sua natureza complexa e hierárquica para representar a informação de arquivo por níveis de descrição poderá tornar o

processo complexo, uma vez que um registo descritivo só está totalmente descrito quando se tem em conta toda hierarquia a que pertence.

Concluiu-se que a tarefa de *harvesting* de apenas um registo (ou um conjunto de registos) do EAD, independentemente da organização hierárquica pode ser praticável. Isto porque, apesar de um ficheiro EAD não ser considerado válido quando não existe o nível de descrição raiz – fundo, o *schema* do EAD não verifica este invariante ou condição para a validação de uma instância EAD. Além disso, pela análise do EAD, verifica-se a existência dos campos *CountryCode* (código do país) e *RepositoryCode* (código do repositório) os quais fazem parte da referência completa de um registo. Desta forma, a existência destes campos na referência completa de um registo e a unicidade das referências dentro de um dado repositório, garantem a unicidade das referências no universo dos repositórios que disponibilizam EAD.

Apesar de estudo aprofundado na recolha de metainformação no formato EAD, concluiu-se da não necessidade de recolha de todos os elementos descritivos patentes do EAD, processo esse que seria consideravelmente difícil. No entanto a eficácia e propósitos do PPA não serão comprometidos, visto que o portal não se pretende substituir aos repositórios, funcionando antes como uma plataforma de direccionamento dos utilizadores para os vários repositórios aderentes. Assim considerou-se que apenas seria necessário recolher alguns campos de metainformação que podem ser captados no formato de metainformação DC simplificado.

Também foi apresentada a arquitectura detalhada do PPA onde foram descritas as principais funcionalidades dos módulos constituintes de acordo com os requisitos funcionais do PPA, função e público-alvo. À luz do protocolo OAI-PMH o PPA corresponde ao *service provider*, o qual oferece serviços de consulta e referência sobre os dados das entidades aderentes colectados. A prestação deste serviço é feita pela colaboração dos vários módulos apresentados, os quais implementem os requisitos funcionais necessários.

Finalmente, mostrou-se o modelo de integração com a Europeana, onde a participação do PPA na Europeana oferece de forma imediata a participação das entidades aderentes da RPA (entidades detentoras de conteúdo de arquivo) na Europeana, sem terem de ser aderentes directos da Europeana, pois o *service provider* da RPA ao ter a função de *data provider* para a

Europeana está implicitamente a disponibilizar os dados de todas as entidades aderentes da RPA à Europeana.

Capítulo 5
Conclusões e trabalho futuro

Neste capítulo apresentam-se as principais conclusões que resultaram do trabalho realizado em torno desta dissertação de mestrado.

Este capítulo está dividido em 2 secções. Na primeira secção enumeram-se as principais conclusões, discussões e considerações finais que foram retiradas deste trabalho. Na segunda secção apresenta-se um conjunto de pontos que se podem realizar como trabalho futuro.

5.1 Conclusões e discussão

Nesta dissertação, começou-se por apresentar vários modelos de arquitecturas com características que dotam o modelo com capacidade de ser instanciado com o cenário em estudo. Contudo, viu-se que as características de alguns podem condicionar a viabilidade e funcionamento do sistema.

Após a apresentação e descrição das principais características de cada um dos modelos e posterior análise comparativa entre eles, verificou-se que o modelo que melhor se adequa aos requisitos da rede portuguesa de arquivos é o sistema de recolha de metadados baseado no protocolo OAI-PMH. Nessa análise constou-se que a aplicação das métricas estudadas a este modelo resulta numa valorização superior comparativamente aos restantes modelos. Além disso, esta arquitectura tem a vantagem de se reger por protocolos e normas, exigindo desta forma que os repositórios de dados que queiram aderir à rede, e posteriormente agregados pelo Portal

Português de Arquivos usem um conjunto de directrizes e normas comuns, no sentido de garantir a interoperabilidade e qualidade dos resultados das pesquisas. Outra característica que coloca em vantagem esta arquitectura em relação às outras é o facto de a sua implementação ser independente de plataformas, o que facilita as entidades que queiram aderir à rede. Esta característica é de extrema importância, assim como a facilidade de adesão à rede, para não colocar entraves à adesão e participação de entidades detentoras à rede.

Após estudada a arquitectura que melhor se adequa ao problema em análise, isto é, o modelo do sistema que deve ser seguido pela rede portuguesa de arquivos, foram estudados 3 formatos de metainformação que podem suportar a partilha de metadados entre as entidades que participam no sistema. Foram estudados formatos com naturezas e características diferentes, entre eles, o EAD, o DC e o ESE.

O EAD é, normalmente, o formato utilizado para representar a natureza complexa e hierárquica da informação em arquivo por níveis de descrição, onde os registos de metainformação estão descritos numa estrutura hierárquica organizada por contextos. Este formato apesar de ser o mais adequado para representar a informação de arquivo apresenta uma complexidade e variantes de utilização que podem tornar complexa a comunicação entre as entidades do sistema.

Por outro lado, o Dublin Core é mais adequado para a transferência de metadados entre duas entidades por ser caracterizado pela sua simplicidade, interoperabilidade semântica e consenso internacional, mas não é suficiente para captar integralmente a descrição de informação de arquivo.

Da mesma forma que o Dublin Core, o conjunto de elementos semânticos da Europeana, não é um formato apropriado para captar de forma integral a informação necessária das descrições arquivísticas.

No capítulo referente à arquitectura do sistema apresentaram-se algumas soluções e consensos para mitigar alguns problemas de transferência de informação entre as entidades que participam na rede, isto é, entre os *data providers* e os *service providers*.

Viu-se que o processo de recolha de metadados no formato EAD pode ser difícil de implementar, apesar de não ser necessária a implementação de *crosswalks* para mapear os dados

para o formato EAD, pois este formato é, normalmente, já o formato de metainformação utilizado pelas entidades detentoras de arquivo. Contudo a sua natureza complexa e hierárquica para representar a informação de arquivo por níveis de descrição torna o processo complexo, uma vez que um registo descritivo só está totalmente descrito quando se tem em conta toda hierarquia a que pertence, isto é, desde o nível de descrição fundo (raiz) até ao registo em causa.

Concluiu-se que a tarefa de *harvesting* de apenas um registo (ou um conjunto de registos) do EAD, independentemente da organização hierárquica pode ser exequível. Isto porque, apesar de um ficheiro EAD não ser considerado válido quando não existe o nível de descrição raiz – fundo, o *schema* do EAD não verifica este invariante ou condição para a validação de uma instância EAD. Além disso, pela análise do EAD, verifica-se a existência dos campos *CountryCode* (código do país) e *RepositoryCode* (código do repositório) os quais fazem parte da referência completa de um registo. Desta forma, a existência destes campos na referência completa de um registo e a unicidade das referências dentro de um dado repositório, garantem a unicidade das referências no universo de todos os repositórios.

Apesar do estudo aprofundado na recolha de metainformação no formato EAD através do protocolo OAI-PMH, concluiu-se da não necessidade de recolha de todos os elementos descritivos patentes do EAD, processo esse que seria consideravelmente difícil. No entanto a eficácia e propósitos do PPA não serão comprometidos, visto que o portal não se pretende substituir aos repositórios, funcionando antes como uma plataforma de direccionamento dos utilizadores para os vários repositórios aderentes. Assim considerou-se que apenas seria necessário recolher alguns campos de metainformação que podem ser captados pelo formato de metainformação DC (simplificado).

Também foi apresentada a arquitectura detalhada do Portal Português de Arquivos, onde foram descritas as principais funcionalidades dos módulos constituintes de acordo com os requisitos funcionais do portal, função e público-alvo. À luz do protocolo OAI-PMH, o PPA corresponde ao *service provider*, o qual oferece serviços de consulta e referência sobre os dados colectados das entidades aderentes. A prestação destes serviços é feita pela colaboração dos vários módulos apresentados, os quais implementem os requisitos funcionais necessários.

Finalmente, mostrou-se o modelo de integração com a Europeana, onde a participação do PPA na Europeana oferece de forma imediata a participação das entidades aderentes da RPA (entidades detentoras de conteúdo de arquivo) na Europeana, sem terem de ser aderentes directos da Europeana, pois o *service provider* do PPA ao ter a função de *data provider* para a Europeana está implicitamente a disponibilizar os dados de todas as entidades aderentes da RPA à Europeana.

O formato que se entendeu mais adequado para a disponibilização de informação por parte dos *data providers* (repositórios aderentes) da RPA foi o DC, pelos motivos já apresentados. Contudo o *data provider* do PPA, isto é, o serviço que disponibiliza os dados colectados na base de dados central da RPA, deve disponibilizar também o formato ESE aquando a integração com a Europena. Desta forma é fornecida uma completa compatibilidade com este repositório.

5.2 Trabalho futuro

Aquando a apresentação da arquitectura detalhada do Portugal Português de Arquivos descreveram-se os vários módulos funcionais que o constituem, nomeadamente, o módulo de validação, módulo de agregação, módulo de registo, módulo de pesquisa, módulo de administração e *data providers* para a Rede Portuguesa de Arquivos.

Todos os módulos pertencentes ao *service provider,* isto é, o conjunto de módulos que constituem o PPA ainda não foram desenvolvidos. Estes módulos serão desenvolvidos no âmbito do projecto da responsabilidade da Direcção-Geral de Arquivos.

Após estes desenvolvimentos, para a integração do PPA com a Europeana será necessário torná-lo um *data provider*. O PPA ao ter a função de *data provider* para a Europeana está implicitamente a disponibilizar os dados de todas as entidades aderentes da RPA à Europeana, sem terem de ser seus aderentes directos.

A aplicação da arquitectura OAI-PMH definida nesta dissertação pode ser aplicada a outros cenários, para a agregação de conteúdos detidos por entidades de uma rede que funcionam de modo integrado e articulado na prossecução de objectivos comuns. Poderão ser redes governamentais, tal como a Rede Portuguesa de Arquivos, redes empresarias, redes institucionais, redes de repositórios de objectos de ensino (*e-lerning*), etc.

Referências

ANSI (2009). American National Standards Institute web page Retrieved 2009-10-15, from http://www.ansi.org/

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

Congress, L. o. (2008). CQL: Contextual Query Language (SRU Version 1.2 Specifications)

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

DCMI (1999). List of Resource Types Retrieved 2009-08-21, from http://dublincore.org/documents/resource-typelist/

DCMI (2000). Dublin Core Qualifiers Retrieved 2009-08-21, from http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/

DCMI (2008). DCMI Metadata Terms Retrieved 2009-08-21, from http://dublincore.org/documents/dcmi-terms/

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/

Deep Web Technologies (2009). Scitopia: Deep Federated Search Retrieved 2009-08-12, from http://www.scitopia.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 (2009a). Portal Português de Arquivos Retrieved 2009-10-22, from http://www.dgarq.gov.pt/rede-portuguesa-de-arquivos/portal-de-arquivos/

DGARQ (2009b). Rede Portuguesa de Arquivos Retrieved 2009-10-12, from http://www.dgarq.gov.pt/rede-portuguesa-de-arquivos/

Europeana (2009a). Europeana Semantic Elements specifications Retrieved 2009-10-05, from http://version1.europeana.eu/c/document_library/get_file?uuid=c56f82a4-8191-42fa-9379-4d5ff8c4ff75&groupId=10602

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

FCCN (2009a). b-on - Biblioteca do Conhecimento Online Retrieved 2009-10-10, from http://www.b-on.pt/

FCCN (2009b). Portal de pesquisa da b-on Retrieved 2009-10-10, from pesquisa.b-on.pt

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.

Fox, M. J. (2001). Locating EAD in the Descriptive Firmament. Retrieved from http://www.informaworld.com/smpp/ftinterface?content=a909291418&rt=0&format=pdf

IANA (2007). MIME Media Types Retrieved 2009-09-02, from http://www.iana.org/assignments/media-types/

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

IETF (1995). RFC 1766 - Tags for the Identification of Languages, from http://www.ietf.org/rfc/rfc1766.txt

ISO (2009). ISO 3166 - English country names and code elements Retrieved 2009-09-02, from http://www.iso.org/iso/english_country_names_and_code_elements

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

MERLOT (2009). Multimedia Educational Resource for Learning and Online Teaching Retrieved 2009-08-12, from http://www.merlot.org/

NISO (2009a). National Information Standards Organization web page Retrieved 2009-10-20, from www.niso.org/

NISO (2009b). National Information Standards Organization Z39.50 Information Retrieval Protocol Retrieved 2009-09-20, from http://www.niso.org/standards/resources/Z39.50_Resources

OAI-PMH. The Open Archives Initiative Protocol for Metadata Harvesting, from http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

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

Open Archives Initiative. The Open Archives Initiative Protocol for Metadata Harvesting Version 2.0 Retrieved 2007-11-22, from http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

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/

U.S. Department of Energy (2009a). Science Accelerator Retrieved 2009-08-12, from http://www.scienceaccelerator.gov/

U.S. Department of Energy (2009b). USA.gov for Science - Government Science Portal Retrieved 2009-08-12, from http://www.science.gov/

U.S. Department of Energy (2009c). WorldWideScience.org The Global Science Gateway Retrieved 2009-08-12, from http://worldwidescience.org/

University of Bristol (2008). MetaLib: your resource gateway Retrieved 2009-08-12, from metalib.bris.ac.uk

W3C (1997). ISO 8601 - Date and Time Formats, from http://www.w3.org/TR/NOTE-datetime

W3C (1999). XML Path Language (XPath) version 1.0. Retrieved 2009-02-14, from http://www.w3.org/TR/xpath

W3C (2000). Simple Object Access Protocol (SOAP) 1.1 Retrieved 2009-10-05, from http://www.w3.org/TR/2000/NOTE-SOAP-20000508/