

# Creating a National Federation of Archives using OAI-PMH

Luís Miguel Ferros<sup>1</sup>, José Carlos Ramalho<sup>1</sup> and Miguel Ferreira<sup>2</sup>

<sup>1</sup>Department of Informatics – University of Minho  
Campus de Gualtar, 4710 Braga – Portugal  
{lferros, jcr}@di.uminho.pt

<sup>2</sup>Department of Information Systems – University of Minho  
Campus de Azurém, 4800 Guimarães – Portugal  
mferreira@dsi.uminho.pt

**Abstract.** This paper describes the planning stages of the creation of a national level federation of archives. The Open Archives Initiative Protocol for Metadata Harvesting will be used for collecting metadata being produced by local archives using any compliant records management software. The first stage of implementation will harvest metadata produced with the DigitArq [1-3] platform in a pilot group of Portuguese Regional Archives. DigitArq relies on Encoded Archival Description (EAD) metadata to describe its collections. However, the overwhelming flexibility and complexity of EAD make the harvesting operation more complex than usual. This paper addresses possible ways of exchanging EAD records using OAI-PMH as the basis for a central repository of metadata that will enable the creation of advanced information services at the national level.

**Keywords:** Metadata, harvesting, data provider, service provider, interoperability, repository, digital archives, OAI-PMH, EAD, XML, Dublin Core

## 1 Introduction

The use of standards for describing items of information introduces one major benefit in the globalised information society that we live on: it enables information systems to be interoperable. The arrival of the Open Archives Initiative [1] enabled repositories and other types of information systems to share and concentrate information, and more often metadata, allowing a great deal of new information services to be developed, such as centralised search engines as Google Scholar [4] or OAIster [5] and global statistics such as the ones provided by the Registry of Open Access Repositories (ROAR) [6].

The purpose of this paper is to discuss the major obstacles one will face while attempting to centralize archival metadata coming from different regional

repositories. More specifically, we will address the problematic of using OAI-PMH to transfer metadata encoded in EAD (Encoded Archival Description) [7]. Due to its flexibility, hierarchical nature and complex structure, EAD presents several challenges to anyone trying to exchange this type of information in an efficient way.

This paper is organized as follows: in the section 2 we provide a short description of the EAD standard. Section 3 describes OAI-PMH and its common requests; in section 4 we describe the architecture of the proposed system and problems we expect to face during its implementation; and finally, in section 5 we draw some conclusions and outline some points of future work.

## **2 EAD as the archival standard for descriptive metadata**

EAD [2] is a non-proprietary standard for encoding archival of finding aids. The purpose of EAD is to provide information about archival resources in standard syntax and normalized language. An instance of an EAD document is composed of three parts: a header, a front matter and the archival description of collections (a collection of documents created by a single person, family or organization).

The header section contains information about the EAD document itself [2]. The front matter embeds information convenient for publishing or rendering the finding aid. The archival description contains the bulk of an EAD document instance, which describes the content, context, and extent of a body of archival materials, including administrative and complementary information that facilitates the use and the discovery of the material.

Information in an EAD instance is organized in unfolding hierarchical levels that account for an overview of the whole collection to be followed by a more detailed view of its constituent parts, e.g. sections, classes, documents, etc [1] (Fig. 1). Each level of description contains information that roughly follows the ISAD(g) model [8]. Examples of descriptors that are commonly found at one of these description levels are: title, range of dates, biographic history, archival history, scope and content, existence and location of originals and copies, physical characteristics, etc.

EAD can be used to describe all sorts of archival material, these being physical, like books, reports and photographs, or digital, such as databases, Web pages or spreadsheets.

```

<ead>
  <eadheader>
    <eadid />
    <filedesc>
      <titlestmt>
        <titleproper />
      </titlestmt>
    </filedesc>
  </eadheader>
  <archdesc level="otherlevel" otherlevel="F">
    <did>
      <abstract />
      <unitid countrycode="PT" repositorycode="ADPRT">EMP/BM</unitid>
      <physdesc>
        ...
      </physdesc>
    </did>
    <dsc>
      <c level="otherlevel" otherlevel="SC">
        <did>
          <unitid countrycode="PT" repositorycode="ADPRT">CR</unitid>
          <physdesc>
            ...
          </physdesc>
        </did>
      </c>
    </dsc>
  </archdesc>
</ead>

```

Fig. 1 - Extract of an EAD instance

### 3 OAI-PMH as the standard for metadata exchange

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [9] plays the important role of enabling repositories to become interoperable. The main goal of OAI-PMH is to allow geographically separated repositories to exchange metadata thus allowing the creation of repository federations.

The OAI-PMH defines a communication protocol that defines how the transference of metadata should be performed between two basic entities: data providers and service providers.

Data providers support the OAI-PMH as way to publish their metadata. The service providers send OAI-PMH requests to data providers and harvest their metadata that will serve as the basis for the development of more advanced services. The interaction between these two entities is depicted in the Fig. 1. As one can observe, a service provider that wants to harvest metadata sends a HTTP request to a data provider, which, according to the request, responds with a XML message.

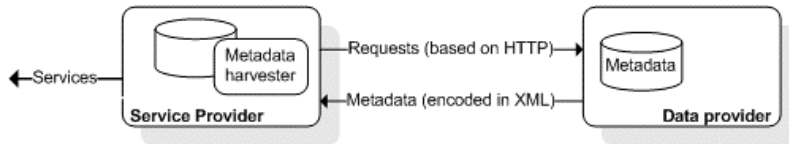


Fig. 2 - Interaction between OAI-PMH entities

For data providers to be able to publish their metadata through OAI-PMH, they must implement six types of requests (called the verbs in this context):

- **GetRecord** - This verb is used to retrieve an individual metadata record from a repository. Required arguments specify the identifier of the item from which the record is requested and the format of the metadata that should be included in the record [10].
- **Identify** - This verb is used to retrieve information about a repository. Some of the information returned is required as part of the OAI-PMH. Repositories may also employ the Identify verb to return additional descriptive information [10].
- **ListRecords** - This verb is used to harvest records from a repository. Optional arguments permit selective harvesting of records based on set membership and/or datestamp [10].
- **ListIdentifiers** - This verb is an abbreviated form of ListRecords, retrieving only headers rather than records. Optional arguments permit selective harvesting of headers based on set membership and/or datestamp [10].
- **ListMetadataFormats** - This verb is used to retrieve the metadata formats available from a repository. An optional argument restricts the request to the formats available for a specific item [10].
- **Listsets** - This verb is used to retrieve the set structure of a repository, useful for selective harvesting [10].

Fig. 3 shows an request of an request that list the metadata formats that can be disseminated from the repository <http://www.perseus.tufts.edu/cgi-bin/pdataprov> for the item with unique identifier `oai:perseus.tufts.edu:Perseus:text:1999.02.0119`.

The response to this request (Fig. 4) shows that 3 metadata formats are supported for the given identifier: `oai_dc`, `olac` and `perseus`. For each of the formats, the location of an XML Schema describing the format, as well as the XML Namespace URI is given.

```

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

```

Fig. 3 – OAI-PMH request with the 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>

```

Fig. 4 – OAI-PMH Response

## 4 Architecture and system operation

In this section, we describe the architecture and operational characteristics of a system that uses the OAI-PMH to offer two important services: centralization of metadata and interoperability between repositories.

### Metadata centralization

Fig. 2 shows a simplified diagram of the system's architecture that uses the OAI-PMH to harvest metadata from several EAD repositories [1-3].

The protocol provides, as shown in the diagram, two main types of participants: the data providers and the service providers. In this particular example, the data providers are the digital repositories that hold the archival metadata. To ensure interoperability, the data providers must provide their metadata according to common descriptive metadata standard. In this case, this should be the EAD. The service provider offers builds additional added-value services from the metadata harvested and stored in its central repository (CR).

The harvesting task is performed by the Metadata harvesting module by sending OAI-PMH requests to data providers, which according to the type of request, will receive appropriate XML responses. Some of those will deliver the EAD metadata that is hosted in the data provider.

The metadata harvested by the previous module is transformed and adapted as necessary to fit the central repository (CR). This process is carried out by the “XML EAD to CR” component.

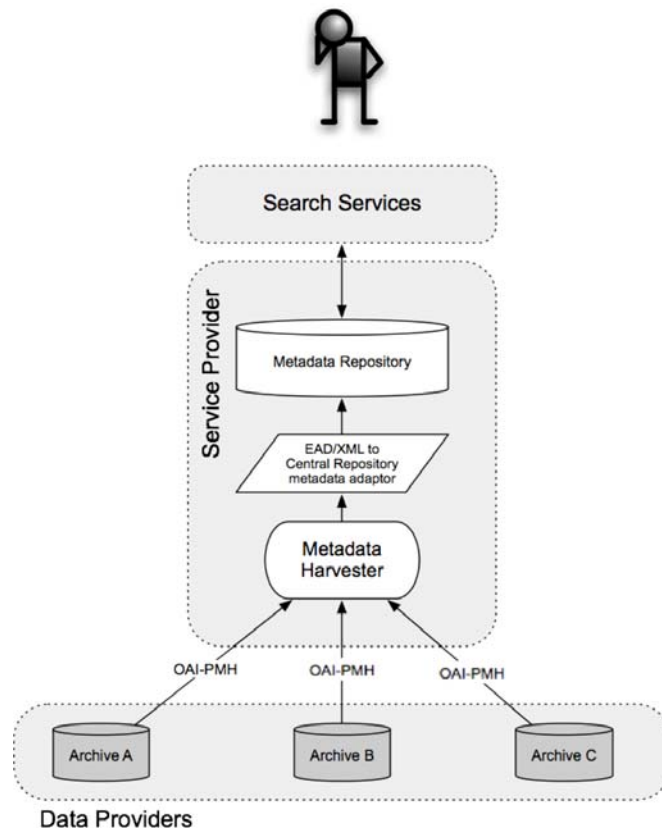


Fig. 5 - Overview of the system's architecture

EAD represents hierarchically the overall records that compose a collection. A single collection may range from a few dozen nodes to a staggering size of millions. This being said, a question may be raised: If the atomic unit in an EAD file is the collection (with all its complexity and size), how does one harvest this type metadata if only one of its nodes gets updated? Well, this problem can be addressed in three different ways:

### **1. Harvest the complete collection that holds the updated record.**

This solution is simple to implement, since the answer to the request consists in sending the corresponding updated collection. The service provider must only integrate the new version of the collection in its central repository by simply replacing the old one by the new collection. Although easy to implement, it is an inefficient solution, because a simple change or insertion of a new record in a previously harvested collection will trigger a subsequent harvest of the complete set of nodes that compose the collection. This strategy is very inefficient at the bandwidth usage level.

### **2. Harvest the whole branch that contains the updated record.**

This method triggers a transfer of data much lower than previously described. However, the operation of extracting the nodes from data providers and the integration of these records with the service providers is much more complex. The extraction involves selecting the nodes all the way up the branch of the collection's tree. Consequently, the integration will be achieved by the replacement of old records by the received records on the corresponding collection in the central repository.

### **3. Harvest only the updated record.**

This is certainly the most efficient approach, as only the new or changed records are harvested independently of their position in the collection structure. The problem in this approach is that an EAD file is not valid if the whole collection is not present. Using this strategy implies that the nodes are identified uniquely at the national level so as to guarantee that they are integrated in the right collection in the correct position. Analysing EAD one can verify the existence of a CountryCode (code of the country) and a RepositoryCode (code of the repository) elements, which compose the complete reference of a record. This way, the existence of these fields in the complete reference of a record and the unique references inside a given repository, guarantee the uniqueness of references in the repositories universe that publish EAD. So, the task of harvesting only a record (or a set of records), independently of the hierarchical organization can be possible, as long as we assume to be exchanging incomplete EADs that are not valid from the EAD Schema perspective.

## **Interoperability between repositories**

To increase interoperability among repositories, these should disseminate their metadata in formats other than EAD. For example, it is common practice for the library community to disseminate its records in Dublin Core (DC) [11] independently of the metadata schemas that are being used in their information systems. Having a dissemination port for Dublin Core enables the EAD repositories to be compatible to an assortment of pre-existing service providers, e.g. OAIster, ROAR and Google Scholar.

Because the metadata schemas used in the archival repositories depicted in this paper are based in EAD [2], one must implement a crosswalk from EAD to DC. This issue has already been addressed by Prom and Habing in [12].

## 5 Conclusions and Future Work

In this paper was described the simple architecture of an information system capable of implementing the concept of a National Federation of Archives that is based on EAD that uses the OAI-PMH for exchanging metadata.

The EAD structure, due to its hierarchical nature and overwhelming flexibility makes the use of OAI-PMH a non-straightforward process. The paper identifies the causes that make this type metadata be so hard to harvest and synchronize. To address this issue, we have identified and described three possible solutions, that differ both in complexity and efficiency. Instead of exchanging full blown EAD collections we propose that solely the updated nodes are harvested and integrated in their corresponding collections. However, one should note that the description of a record outside of the context of a collection may not be sufficient to fully understand it, as the ascending nodes of description are required to make it complete and structured. However, for the sole purpose of transferring data and incrementally updating a central repository this solution seems highly appropriate.

As future work we will address in more detail the EAD to Dublin Core crosswalk strategy so that the national repositories may also disseminate their metadata and integrate with existing service providers.

It may also be interesting, on the service provider side, to develop modules that transform metadata in formats other than EAD so that one can provide services for other types of repositories.

## REFERENCES

- [1] M. Ferreira and J. C. Ramalho, "DigitArq - Creating and Managing a Digital Archive," presented at ICCC/IFIP International Conference on Electronic Publishing, Brasília, Brazil, 2004.
- [2] M. Ferreira and J. C. Ramalho, "DigitArq: Creating a Historical Digital Archive," presented at 5ª Conferência da Associação Portuguesa de Sistemas de Informação, Lisboa, 2004.
- [3] J. C. Ramalho, M. Ferreira, L. Ferros, M. J. P. Lima and A. Sousa, "DigitArq 2 - Nova plataforma aplicacional para gestão de Arquivos Definitivos," presented at 2nd International Conference on Enterprise Archives (2ª Conferência Internacional de Arquivos Empresariais), Seixal, Portugal, 2006.
- [4] Google, "Google Scholar." [Online]. Available: <http://scholar.google.com>.
- [5] University of Michigan, "OAIster." [Online]. Available: <http://www.oaister.org/>.
- [6] University of Southampton, "Registry of Open Access Repositories (ROAR)." [Online]. Available: <http://roar.eprints.org/>.

- [7] Library of Congress, "EAD - Encoded Archival Description," in Library of Congress, 1998. [Online]. Available: <http://www.loc.gov/ead/>. [Accessed 2004]
- [8] International Council on Archives, "ISAD(G): General International Standard Archival Description, Second edition," International Council on Archives 0-9696035-6-8, 1999.
- [9] Open Archives Initiative, "The Open Archives Initiative Protocol for Metadata Harvesting." [Online]. Available: <http://www.openarchives.org/pmh/>.
- [10] Open Archives Initiative, "The Open Archives Initiative Protocol for Metadata Harvesting Version 2.0." [Online]. Available: <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>.
- [11] Dublin Core Metadata Initiative, "Dublin Core Metadata Initiative." [Online]. Available: <http://dublincore.org/>.
- [12] C. J. Prom and T. G. Habing, "Using the open archives initiative protocols with EAD " presented at 2nd ACM/IEEE-CS joint conference on Digital libraries, Portland, Oregon, USA 2002.