CULTURAL HERITAGE on line
Empowering users: an active role for user communities
INTERNATIONAL CONFERENCE
Florence 15th - 16th December 2009

Maristella AGOSTI, Nicola FERRO, and Gianmaria SILVELLO

Enabling cross-language access to archival metadata

Abstract

In this paper we analyze the ratio between Digital Library (DL), archives and multilingualism. We focus our attention on the interoperability issues that need to be faced when you attempt to make different cultural institutions cooperate, to allow a selective and pinpoint online access to their resources, and to enable cross-language retrieval of their materials.

Introduction

Digital Library (DL) systems have been becoming the fundamental tool for managing, exchanging and searching cultural digital resources and as a research field has seen continuous growth over the last ten years. The central role of DL in fostering access to our cultural heritage is also enhanced by the European Commission which financially supports many projects related to DL, such as the TELplus project[1], which aims to offer a free service to access the resources of the 48 national libraries of Europe in 20 languages, or the Digital Repository Infrastructure Vision for European Research (DRIVER) project[2], the goal of which is to develop a pan-European Digital Repository Infrastructure by integrating existing individual repositories from European countries and developing a core number of services, including search, data collection, profiling and recommendation. Furthermore, the "European Commission Working Group on Digital Library Interoperability has the objective of providing recommendations for both a short term and a long term strategy towards the setting up of the European Digital Library as a common multilingual access point to Europe's distributed digital cultural heritage including all types of cultural heritage institutions" [4]. In particular, the recipient of these recommendations is Europeana[3] , which aims at addressing the interoperability issues among European museums, archives, audio-visual archives and libraries for the creation of the "European Digital Library". From this picture we can see that DL are not merely the digital counterpart of traditional libraries, but they are the fundamental tool for pursuing interoperability between different cultural organizations such as libraries, archives and museums. Collecting and managing the resources of these organizations is fundamental for providing wide, distributed and open access to our cultural heritage.

Currently, libraries are the foremost components of DL, this is due to the availability of technologies well-suited for them and that have been adopted by DL since their conception such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) that is the standard de-facto for metadata exchange in distributed environments and the Dublin Core[4] (DC) metadata format which is a tiny and lightweight metadata format that is getting the preponderant mean to exchange information. Archives and museums should adopt these technologies to exploit the services offered by the DL systems; two European projects pursue this goal: the APEnet[5] (Archives Portal of Europe on the Internet), which aims to build an Internet Gateway for Documents and Archives in Europe, and the Athena (Access to cultural heritage networks across Europe) project[6], which aims to reinforce, support and encourage the participation of museums and other institutions coming from those sectors of cultural heritage not fully involved yet in Europeana. Unfortunately, the process of adopting these technologies and exploiting the DL system advanced services is not as straightforward as it is for the libraries; this is due to the nature and the organization of the archives and of the museums as cultural institutions. In this paper we shall concentrate on archives because the problematic issues of museums can be related to those of archives; indeed, often museum resources are described and organized as archival resources. The archival structure is deeply hierarchical and the relationships between the documents must be retained to express their full informational power. These characteristics lead to the development of metadata standards such as the Encoded Archival Description (EAD) which are not particularly well-suited to be used within the DL systems. These standards may be a barrier towards the interoperability between the cultural institutions and towards the automatic processing of the data. These difficulties have moved archives away from full participation in DL, in particular they have limited the access to several services offered by DL systems. For both archives and

1 http://www.theeuropeanlibrary.org/telplus/

2 http://www.driver-repository.eu/

3 http://www.europeana.eu/

4 http://www.dublincore.org/

5 http://www.apenet.eu/

6 No Website yet available.

libraries, multilingual access to the resources is a key point especially in the European context; indeed, multilingualism also promoted the CACAO European project[7] which aims to offer an innovative approach for accessing, understanding and navigating multilingual textual content in digital libraries. Furthermore, the CACAO infrastructure will be adopted by "The European Library" to promote aggregation of different contents at the European level. In this paper we analyze the problematic issues which could prevent the use of the multilingual services within the archival digital resources. Moreover, we shall propose a methodology that permits us to exploit the techniques adopted by the libraries with the archival metadata, enabling a multilingual access to these valuable resources.
The paper is organized as follows: Section 2 introduces the three main techniques to address metadata-related challenges in a multilinguistic environment. In Section 3 we briefly describe the archival organization and we explain why EAD metadata format does not work well in distributed and multilingual environments. In Section 4 we present our methodology which maps the EAD files into a combination of sets and DC metadata enabling the use of the cross-language techniques. Finally, in section 5 we draw some conclusions.

Cross-Language Access: Metadata-Related Challenges and Solutions

In the European Union (EU) there is a huge need to provide cross-language access to information; this is due to the diversity and multilingual EU environment where there are 23 official languages spoken in 27 member states. Cross-language access to information leads to problems of both semantic and syntactic interoperability [6]. Many solutions such as those adopted by the CACAO Project aim to address these problems mainly through the use of metadata, which provide access to a multilingual corpus of cultural resources.
A system which has to provide cross-language access to information must address two important metadata-related challenges which can be tackled by specifying the language of the metadata fields [6]: false friends and term ambiguity. To address these issues three main solutions are usually considered:

- Translation: A query formulated in the user language is automatically translated in the other supported languages and then submitted to the system. This solution is not free from the false friends issue.
- Enrichment of Metadata: The aim is to make the intended meaning of information resources explicit and machine-processable, to allow machines and humans to better identify and access the resources. The language would thus be provided in the metadata itself.
- Association to a Class: Terms are associated to a fairly broad class in a library classification system such as the Dewey Decimal Classification (DDC). This is a common solution for the term ambiguity problem and is similar to synsets used in WordNet[8].

The specification of the language of metadata field enables the full exploitation of metadata for cross-language purposes. If metadata do not come with or cannot be enriched with the language of the field, it is useful to rely on the association to a class technique. This useful technique relies on the use of the subject field of metadata; it is not always possible to determine the subject of a metadata or of a term. This is particularly true for archival metadata where determining the subject can be very difficult.

Archival Metadata and the EAD Format

An archive is a complex cultural organization which is not simply constituted by a series of objects that have been accumulated and filed with the passing of time. Archives have to keep the context in which their documents have been created and the network of relationships among them in order to preserve their informative content and provide understandable and useful information over time. The context and the relationships between the documents are preserved thanks to the strongly hierarchical organization of the documents inside the archive. Indeed, an archive is divided by fonds and then by sub-fonds and then by series and then by sub-series and so on; at every level we can find documents belonging to a particular division of the archive or documents describing the nature of the considered level of the archive (e.g. a fond, a sub-fonds, etc.).
The union of all these documents, the relationships and the context information permits the full informational power of the archival documents to be maintained. In the digital environment an archive and its components are described by the use of metadata; these need to be able to express and maintain such structure and relationships. The standard format of metadata for representing the complex hierarchical structure of the archive is EAD [7], which reflects the archival structure and holds relations between documents in the archive. In addition, EAD encourages archivists to use collective and multilevel description, and because of its flexible structure and broad applicability, it has been embraced by many repositories [7]. The use of EAD is widespread

7 http://www.cacaoproject.eu/home/

8 http://wordnet.princeton.edu/

in the United States of America and also in the EU; for instance the "Nationaal Archief"[9] in the Netherlands preserves a big collection of EAD metadata in Dutch or the "Archives Napoleon"[10] is based on EAD metadata in French. It is important to include archival metadata in DL because they retain unique and valuable information and at the same time it is very useful to enable multilingual services to access and retrieve them.
Unfortunately, the structure of EAD turns out to be a very large eXtensible Markup Language (XML) file with a deep hierarchical internal structure. On the other hand, EAD allows for several degrees of freedom in tagging practice, which may turn out to be problematic in the automatic processing of EAD files, since it is difficult to know in advance how an institution will use the hierarchical elements. The EAD permissive data model may undermine the very interoperability it is intended to foster. Indeed, it has been underlined that only EAD files meeting stringent best practice guidelines are shareable and searchable [10]. Moreover, there is also a second relevant problem related to the level of material that is being described. The EAD schema rarely requires a standardized description of the level of the materials being described and this possibility is often ignored, as pointed out by Pitti in [7]. Therefore, the access to individual items might be difficult without taking into consideration the whole hierarchy. This issue compromises the possibility of automatically enriching the metadata for multiliguality purposes. A single EAD metadata is used to describe an entire archive, thus in a single metadata we can find very different subjects. With this organization it is very difficult to disambiguate the terms or to identify the subject of metadata; with the EAD metadata the "association to a class" solution is essentially unworkable. Moreover, sharing and searching archival description might be made difficult by the typical size of EAD files which could be several megabytes with a very deep hierarchical structure. Indeed, each EAD file is a hierarchical description of a whole collection of items rather than the description of an individual item. On the other hand, users are often interested in the information described at the item level, which is typically buried very deeply in the hierarchy and might be difficult to reach.

A Methodology to Enable Both Cross-Language Access and Exchange of EAD Metadata

In [2] a solution was proposed to enable the sharing of EAD metadata in a distributed environment and enabling the variable granularity access to the data; this solution maintains also the integrity and the structure of the described archive exploiting OAI-PMH inner structure and the DC metadata; indeed, it is based on a methodology which enables an EAD file to be represented as a combination of OAI-sets and several DC metadata. To properly understand this methodology it is worthwhile briefly describing the functionality of OAI-PMH called selective harvesting and how its internal organization based on OAI-sets can be used to express a hierarchical structure as an organization of nested sets [3].
Selective harvesting is based on the concept of OAI-set, which enables logical data partitioning by defining groups of records. Selective harvesting is the procedure which enables the harvesting only of metadata owned by a specified OAI-set. In OAI-PMH a set is defined by three components: setSpec which is mandatory and a unique identifier for the set within the repository, setName which is a mandatory short human-readable string naming the set, and setDesc which may hold community-specific XML-encoded data about the set. OAI-set organization may be hierarchical, where hierarchy is expressed in the setSpec field by the use of a colon [:] separated list indicating the path from the root of the set hierarchy to the respective node. For example, if we define an OAI-set whose setSpec is "A", its subset "B" would have "A:B" as setSpec. When a repository defines a set organization it must include set membership information in the headers of the records returned to the harvester requests. We exploit this structure to represent a hierarchical structure such as a tree data structure as an organization of nested sets as shown in Figure 1. Here we can see that each node of the tree can be mapped into a set, where child nodes become proper subsets of the set created from the parent node. Every set is subset of at least one set; the set corresponding to the tree root is the only set without any supersets and every set in the hierarchy is subset of the root set. The external nodes are sets with no subsets. The tree structure is maintained thanks to the nested organization and the relationships between the sets are expressed by the set inclusion order [3]. This methodology allows us to decompose the EAD tree structure into an organization of OAI-sets where the elements belonging to a set are metadata records. The structure of the EAD is maintained by the OAI-sets and the data are mapped into many DC records. As far as the mapping of the actual content of EAD items into DC records is concerned, we adopt the mapping proposed by Prom and Habing [9]. Our solution differs from [9] from a syntactic point-of-view: we propose to maintain the hierarchical structure of EAD throughout an organization of OAI sets containing the DC records mapping the content of EAD items. In [9] the hierarchical structure is maintained by means of several pointers connecting the DC records to the original EAD file.

9 http://www.nationaalarchief.nl/

10 http://www.archivesnationales.culture.gouv.fr/chan/chan/archives_napoleon-averti. htm

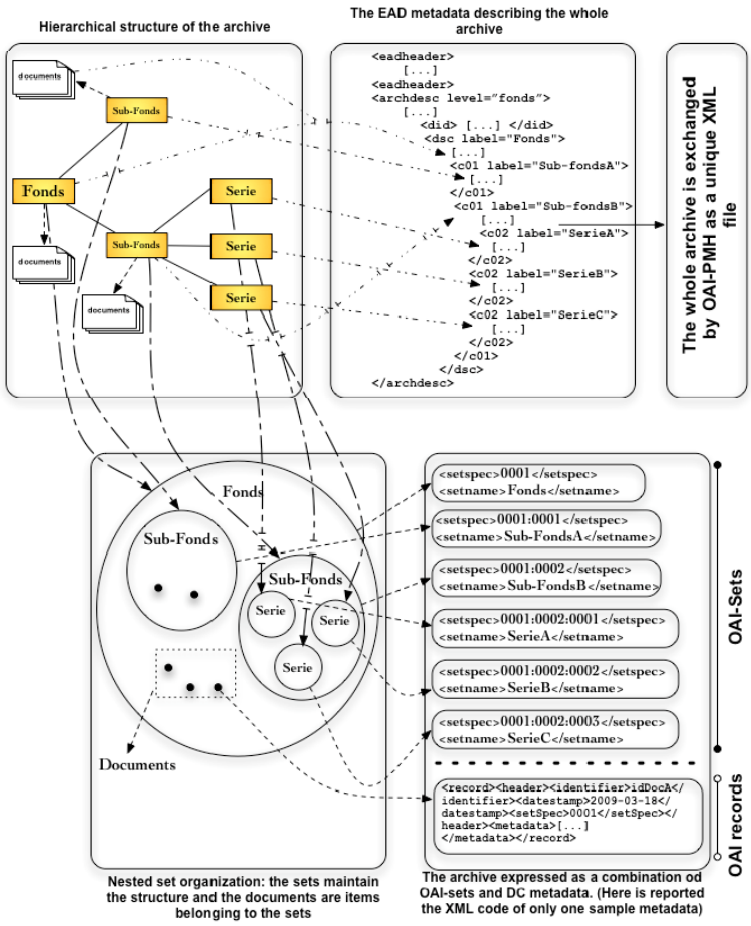

Figure 1 An EAD file mapped into a collection of OAI-sets and DC metadata records.

In Figure 1 we can see two approaches to representing the archival organization and documents. The first approach is the EAD-like one in which the whole archive is mapped inside a single XML file. All information about fonds, sub-fonds or series as well as the documents belonging to a specific archival division are mapped into several XML elements in the same XML file. With this approach we cannot exchange precise metadata through OAI-PMH, rather we have to exchange the whole archive. At the same time it is not possible to determine a specific subject or to access a specific piece of information without considering or accessing the whole hierarchy.

By means of our approach, which graphical representation is shown in the lower part of Figure 1 we can transform archival metadata into a collection of DC metadata and OAI-sets. This solution is particularly well suited for use in the context of the several European projects and in particular for the CACAO project which relies on OAI-PMH to harvest the metadata and on DC records as minimum metadata requirement. In this way the solutions proposed to enable cross-language access to digital contents can be applied also with the archival metadata opening these valuable resources to a significant service offered by the DL technology. Indeed, the decomposition of an archive from a single EAD file into several DC metadata makes it easier to determine the subject of each single metadata and thus to apply the “association to a class” solution; in the same way the metadata enrichment can be adopted because the DC metadata are well-suited to automatic processing. As we can see, thanks to this methodology, the cross-language solutions developed for the library context can be easily adopted in the archival context without any additional efforts.

Acknowledgments

The work reported in this paper has been partially supported by a grant from the Italian Veneto Region. The study is also partially supported by the TELplus Targeted Project for Digital Libraries, as part of the eContentplus Program of the European Commission (Contract ECP-2006-DILI- 510003).

References

[1] A. Bosca and L. Dini. CACAO Project at the TEL@CLEF 2008 Task.
[2] N. Ferro and G. Silvello. A Methodology for Sharing Archival Descriptive Metadata in a Distributed Environment. In Proc. 12th European Conf. on Research and Advanced Technology for DL (ECDL 2008), pages 268-279. Lecture Notes in Computer Science (LNCS) 5173, Springer, 2008.
[3] N. Ferro and G. Silvello. The NESTOR Framework: How to Handle Hierarchical Data Structures, In Proc. 13th European Conf. on Research and Advanced Technology for DLs (ECDL 2009), pages 215-226. Lecture Notes in Computer Science 5714, Springer, 2009.
[4] S. Gradmann. Interoperability of Digital Libraries: Report on the work of the EC working group on DL interoperability. In Seminar on Disclosure and Preservation: Fostering European Culture in The Digital Landscape. National Library of Portugal, September 2007.
[5] K. Kiesling. Metadata, Metadata, Everywhere - But Where Is the Hook? OCLC Systems & Services, 17(2), pages 84-88, 2001.
[6] B. Levergood, S. Farrenkopf, and E. Frasnelli. The Specification of the Language of the Field and Interoperability: Cross-Language Access to Catalogues and Online Libraries (CACAO). In Proc. Of the Int'l Conf. on Dublin Core and Metadata Applications 2008, pages 191-196.
[7] D. V. Pitti. Encoded Archival Description. An Introduction and Overview. D-Lib Magazine, 5(11), 1999.
[8] C. J. Prom. Does EAD Play Well with Other Metadata Standards? Searching and Retrieving EAD Using the OAI Protocols. Journal of Archival Organization, 1(3), pages 51-72, 2002.
[9] C. J. Prom and T. G. Habing. Using the Open Archives Initiative Protocols with EAD. In Proc. 2nd ACM/IEEE Joint Conference on Digital Libraries, (JCDL 2002), pages 171-180. ACM, 2002.
[10] C. J. Prom, C. A. Rishel, S. W. Schwartz, and K. J. Fox. A Unified Platform for Archival Description and Access. In Proc. 7th ACM/IEEE Joint Conf. on DL, (JCDL 2007), pages 157-166. ACM, 2007.