# The Design of a DLS for the Management of Very Large Collections of Archival Objects

Maristella Agosti, Nicola Ferro, and Gianmaria Silvello

Department of Information Engineering, University of Padua, Italy
{agosti, ferro, silvello}@dei.unipd.it

**Abstract.** This work presents the design of a *Digital Library System (DLS)* able to collect, manage and share archival metadata in a distributed environment. Archive characteristics are presented as well as the reasons that make the management of archival resources challenging. In particular, interoperability and heterogeneity are the two most relevant and peculiar challenges to the field. Furthermore, *Compound Digital Objects (CDOs)* are defined in the archival context and an extension of the proposed distributed DLS architecture able to manage this kind of digital objects is described.

## 1 Introduction

The role of *Digital Library Systems (DLSs)* in collecting, managing, sharing and preserving our cultural heritage is increasingly crucial in several contexts. DLSs have been becoming 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 a wide, distributed and open access to our cultural heritage.

In this wide and heterogeneous scenario, interoperability is the most relevant issue that a DLS has to face. In a distributed environment the first problem is interoperability between different information systems; a DLS must be able to collect resources shared by a wide number of different systems without compromising their autonomy and independence. In this context the interoperability issue is emphasized also by another necessity: the designing of a unique access point to several resources widely different in nature.

In this paper we consider a challenging kind of information resource: the archival documents. When archives are considered, interoperability between the archives themselves, between archival resources and between archival and other types of resources must be taken into account. In the work that we have been carrying out we have underlined that DLS technologies need to be revisited to be well-suited and successfully applied to the management of archival metadata and digital objects [1]. In this paper we briefly describe the nature of archival resources and we present a DLS architecture that enables them to be included in a *Digital Library (DL)*. Moreover, we suggest an extension to this DLS architecture able to manage not only archival metadata but also *Compound Digital Objects (CDOs)*.

The paper is organized as follows: in Section 2 we present background projects and initiatives that constitute the context of the work reported in the paper. Section 3 reports why archives indicate very large solutions. Section 4 presents the distributed DLS architecture we defined in order to share and develop advanced services on archival metadata in a distributed environment. In Section 5 we discuss the extension of our solutions to manage, share and retrieve CDOs and in Section 6 we make some final remarks.

## 2 Background

In order to provide wide access to large and broad collections of digital resources and to address interoperability issues, several initiatives have been instituted. The DELOS Network of Excellence on Digital Libraries[1] has proposed and developed a reference model for laying the foundations of digital libraries [2] which takes into account the perspectives and needs of different cultural heritage institutions and provides a coherent view on the main concepts which constitute the universe of digital libraries in order to facilitate the co-operation among different systems.

The "European Commission Working Group on Digital Library Interoperability", active from January to June 2007, had 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" [5]. In particular, the recipient of these recommendations is the Europeana thematic network[2]; which is a project launched in July 2007 with the aim of addressing the interoperability issues among European museums, archives, audio-visual archives and libraries towards the creation of the "European Digital Library".

Interoperability between different systems has been promoted by the *Open Archives Initiative (OAI)*[3] through *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)* [11], a flexible and lightweight protocol for metadata harvesting, which is becoming the *de-facto* standard in metadata exchange in distributed environments. This protocol permits metadata harvesting between different repositories in a straightforward fashion, in order to create aggregated metadata collections and to enable the creation of advanced services on them. At the same time *Dublin Core (DC)*, a tiny and lightweight metadata format, is getting the preponderant mean to exchange information in a wide distributed environment. Indeed, the characteristics of DC have enabled it to address several interoperability problems and it has been chosen as the minimum common denominator in the OAI-PMH environment. Libraries have been using for the couple OAI-PMH and DC since a relatively long time with good results.

[1] `http://www.delos.info/`
[2] `http://www.europeana.eu/`
[3] `http://www.openarchives.org/`

Two relevant European initiatives which both enjoy the benefits of OAI-PMH are *The European Library* portal[4] and *Digital Repository Infrastructure Vision for European Research (DRIVER)*[5]. *The European Library* is a free service that offers access to the resources of the 48 national libraries of Europe in 20 languages. The goal of The European Library is to create a single access point to all the European national libraries. The European Library project offers a concrete integration possibility based on OAI-PMH, used to collect the catalogue records of national libraries. Furthermore, the TELplus project will form another building block of the European Digital Library and is aimed at strengthening, extending and improving The European Library service. In particular, to contribute to interoperability among different organizations cooperating in The European Library, it aims to improve and enhance the adoption of OAI-PMH as a means of integration. DRIVER is a European project whose goal 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 [3]. DRIVER emphasizes the implementation of nominal, globally accepted standards in a real-life system, with a focus on metadata exchange, in particular using OAI-PMH. One of the Digital Library application components provided by DRIVER is an OAI-Publisher Service; in this way DRIVER services operate upon the aggregated content of existing institutional OAI repositories.

Nevertheless, in the archive context neither general interoperability efforts nor the adoption of specific solutions such as OAI-PMH are common and widespread; this contributes to the exclusion of the archival documents from forming a valuable part of the cultural heritage managed by a DLS.

## 3 Why Archives Indicate Very Large Solutions

An archive is the trace of the activities of physical people or juridical organizations in the course of their business which is preserved because of their continued value over 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. In this way archives are able to preserve the provenance of their documents. The preservation of digital resources provenance is an important issue currently being investigated by the scientific research community [8]; it must be considered as a key feature of a DLS, because it is through provenance information that authenticity can be demonstrated, and the history of archival documents can be preserved.

Archival documents are unique and valuable resources that should be prominently part of a DL content. Most archival documents are not available in digital form, but they are described and represented by metadata; sometimes archival resources are metadata themselves. In the archival context metadata are named

[4] `http://www.theeuropeanlibrary.org/`
[5] `http://www.driver-repository.eu/`

*archival descriptive metadata* and they represent archival descriptions. Archival descriptions have to reflect the peculiarities of the archive, retain all the informative power of a document, and keep trace of the provenance and original order in which resources have been collected and filed by archival institutions.

Whilst pursuing interoperability among archives, we have to deal with the lack of metadata standards. Indeed, the only standard defined for archival descriptive metadata is the *Encoded Archival Description (EAD)*. EAD has a flexible data model that reflects the archival structure and holds relations between entities in an archive. On the other hand, the EAD permissive data model may undermine the very interoperability it is intended to foster [9]. Moreover, it has been underlined that the EAD metadata standard is not well-suited to being used in a distributed environment [6]. Different solutions which address the interoperability issue have been studied to permit archival descriptive metadata exchange in a distributed environment. The proposed solutions suggest the couple DC and OAI-PMH as the means to enable the sharing of archival descriptive metadata and to map EAD files in shareable metadata format. The solution proposed in [10] suggests mapping an EAD file into many tiny and easy-to-move DC metadata. In this approach every DC metadata record contains a link to the original EAD file which causes a strong dependency of DC metadata with the original EAD file which narrows the exchange possibilities of the metadata [9]. Another solution proposed in [4] defines a methodology that joins and exploits the characteristics of OAI-PMH and DC. This methodology enables archive hierarchy to be expressed and meaningful relations between archival entities to be preserved by levering the role of OAI sets. The main idea is to map the archive hierarchy into a combination of OAI sets and DC metadata records. This methodology permits archival descriptive metadata to be exchanged in a distributed environment by facing interoperability problems and maintaining the whole archival informative power of the metadata.

Furthermore, to pursue interoperability *authority control* also need to be considered. Authority control enables archivists to disambiguate items with similar or identical headings and to collocate materials that logically belong together. Authority control is implemented by defining the authority files; an authority file enables the unique identification of an entity which is described. Archive resources describe different kinds of realities, such as a person, a private organization or a public institution. In a distributed environment we have to guarantee the definition uniqueness of the entities described by archival resources. Authority files, are a means of interoperability also between archives and libraries or other organizations. A distributed DLS needs to share common authority files to enable interoperability between the participating organizations.

A DLS aimed at collecting and managing archival resources has to face the complex nature of archives. In particular, interoperability and heterogeneity need to be addressed. Indeed, a DLS has to consider a large number of different archives distributed in a territory; each archive exposes a large number of metadata that have to be collected and managed preserving their whole informative power and thus a large amount of additional information.

If we consider not only archival metadata but also archival digital objects the dimension of the DLS notably increases. Thus, a very large DL is required for two main reasons: for managing a wide number of heterogeneous archives and for governing high space demanding digital objects.

## 4 The Conceived Distributed DLS Architecture

The constitution of a DLS whose goal is to put archival resources together must take into account the structure and the size of the participating archives. Archives preserve resources that are unique and valuable pieces, also small and medium archives need to partecipate in the system, because they provide original contributions. Usually, independent private and/or public archives keep archival metadata without sharing them and this prevents the offering of common advanced services on metadata.

A DLS architecture to be used in the archival context must take into consideration two aspects: the maintenance archives management autonomy and, at the same time a coordination view serves to give an integrated vision of the archives participating in the system. The added value of this DLS architecture is that it shares metadata exploiting DLS advances that can be integrated and adapted with preexisting systems using different technologies. In [1] we started the definition of the architecture and the result is a scalable, flexible and widely-adoptable architecture for sharing information in a distributed environment.

Such an architecture exploits the characteristics of the protocol OAI-PMH based on the distinction between Data and Service Provider and the DC metadata format. The DLS architecture we designed is symmetric in sharing and managing both archival descriptive metadata and authority files treated as metadata too. Indeed, archives act as Data Provider by exposing their descriptive metadata and also as a Service Provider by harvesting the authority files exposed by the *Digital Library (DL)*. The DL acts in the same way as a Data Provider furnishing authority files and as Service Provider harvesting archival descriptive metadata. Moreover, the DL acts as the central authority that constitutes authority files.

The DLS is developed as a three-layer architecture, composed of the metadata transport layer, the metadata management layer and the presentation layer. The transport layer represents the DLS transport infrastructure based on OAI-PMH. The archives participating in the system provide archival metadata, whereas the DL harvests the metadata. As stated before, archival metadata have to retain context and hierarchy information; we addressed this issue thanks to a methodology that combines OAI-PMH sets and DC [4]. In order to retain these useful and fundamental information the Service Provider has to harvest not only the metadata but also the whole set organization of the Data Provider. Selective harvesting is an OAI-PMH native procedure and it permits effective metadata harvesting that preserves archival information. In this way the archive organization expressed through sets and metadata is recreated in the Service Provider, thus enabling the implementation of advanced services on fully expressive archival

metadata. The architecture is open to third party OAI-PMH components that for example can harvest the DL Service Provider.

At the second layer we find the management level called DLS-MM, which is composed of an Application Logic part and a Data Logic part. By the use of Application Logic we can develop advanced services both on harvested metadata owned by the DL and on the metadata of the archives. The applications developed for the DLS can be used on DL metadata index and on archive metadata too; indeed they are independent of the transport infrastructure. Thanks to this organization, adding a third-party service to the system will be almost effortless. The Application Logic works on the metadata managed by the Data Logic composed of a database and a set of distributed databases owned by the archives. DLS-MM Data Logic preserves and manages the physical data of the system; so this sub-layer manages archive data and DL archive data as well.

At the third level we have the presentation layer called DLS-UI constituted by the user interfaces. The system presents two main interfaces: the first is a general-purpose interface dedicated to a generic user-type such as archivists, historical researchers, public administrations or private organizations that will use the advanced services available in the DLS; the second is dedicated to specialized users who through this interface can add, remove or update archival metadata.

## 5 Next Steps: Managing Metadata and Compound Digital Objects in a Distributed Environment

The proposed distributed architecture deals with the interoperability issue and it is particularly suited for metadata exchange; it deals with the heterogeneity issue enabling the proper management of archival metadata while retaining context and hierarchy information. To provide advanced services and to improve the delivery and discovery of the information managed by a DLS, both metadata and digital objects need to be managed and shared. From an informative point-of-view, the digital objects to be considered are the *compound digital objects* (CDO). We define a CDO as an aggregation of distinct information units which are combined together in order to shape a logical unique object. CDOs are digital objects that include information about context, provenance and relationships between the resources. Considering system interoperability, the use of CDOs is challenging. In fact metadata would address interoperability issues by means of their structured and standardized nature, whereas the structure of CDOs has not yet been standardized and this is already a moot point.

A digital archival resource represents a CDO; furthermore, we can say that the whole archive represents a unique CDO. The DLS we presented is able to manage metadata while retaining the whole package of information that forms the CDOs. An extension of our model based on archival metadata permits the management of archival digital resources.

In Fig. 1 we can see the extension of the distributed DLS architecture; as an example we explode a node of the network of archives participating the system. Every node participates the system sharing its metadata and harvesting the

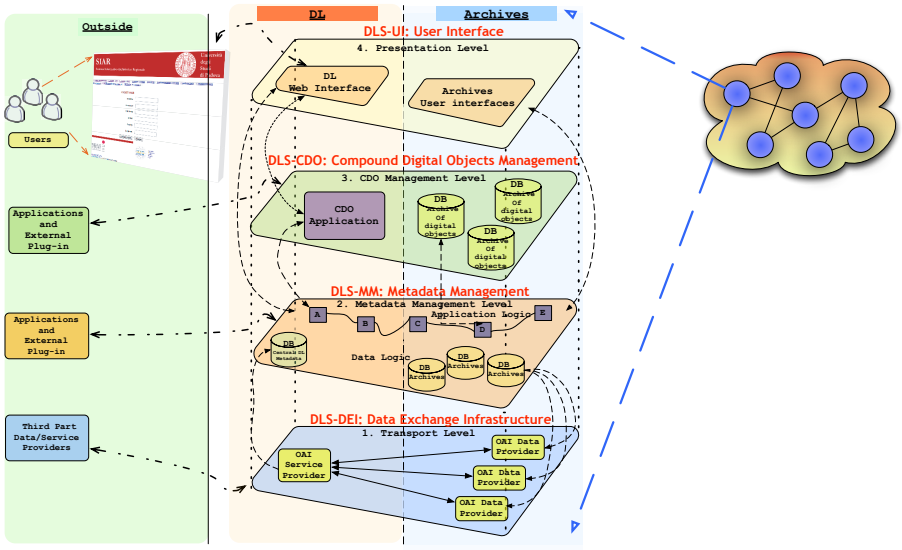


**Fig. 1.** Zoom on a node of the general architecture of the distributed system

authority files exploiting the designed distributed architecture. The final users have an overall view of the coordinated distributed system thanks to a unique interface. A new layer called the CDO Management level is built upon the DLS-MM layer and exploits it to manage, share and expose CDOs in the DL. In CDO Management level, data logic appertains to the archives side, whereas application logic is developed on the side of the DL. Both the logics rest upon and interact with the application logic of the layer below. The archives side contains the databases with the digital archival objects (digital documents or other archival digital goods) in a distributed way that maintains archive independence and the distribution of the preserving effort. Application logic of the DLS-MM layer links the metadata with the digital objects building CDOs that will be managed by the CDO Application. In this model, CDOs are managed, shared and retrieved through metadata which are the foremost entities that enable interoperability and that keep the system scalable, flexible and lightweight. This approach avoids digital object exchange that usually requires a major effort and thanks to the solution based on metadata enables access to CDOs.F urther analyses must consider and face the results of the international initiative named *Open Archives Initiative - Object Reuse and Exchange (OAI-ORE)* [7] which is studying the CDOs and designing an effective way to expose them in the Web.

## 6 Final Remarks

We presented a DLS architecture ables to share, collect and manage archival metadata in a distributed environment; the lightness and scalability of this ar-

chitectural solution have been exposed. We proposed an extension of this DLS architecture enabling the management of the CDOs.

Future works will involve new considerations about CDOs evaluating also the outcomes of international initiatives that are working in this field.

## Acknowledgements

The study is 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). The work of Gianmaria Silvello was partially supported by a grant from the Italian Veneto Region.

## References

1. M. Agosti, N. Ferro, and G. Silvello. An Architecture for Sharing Metadata among Geographically Distributed Archives. In *DELOS Conference*, LNCS 4877 , pages 56–65. Springer, Heidelberg, Germany, 2007.
2. L. Candela, D. Castelli, N. Ferro, G. Koutrika, C. Meghini, P. Pagano, S. Ross, D. Soergel, M. Agosti, M. Dobreva, V. Katifori, and H. Schuldt. *The DELOS Digital Library Reference Model. Foundations for Digital Libraries*. ISTI-CNR at Gruppo ALI, Pisa, Italy, 2007.
3. L. Candela, D. Castelli, P. Manghi, and P. Pagano. Enabling Services in Knowledge Infrastructures: The DRIVER Experience. In *Post-proc. of the 3rd Italian Research Conf. on Digital Library Systems (IRCDL 2007)*, pages 71–77. ISTI-CNR at Gruppo ALI, Pisa, Italy, 2007.
4. N. Ferro and G. Silvello. A Methodology for Sharing Archival Descriptive Metadata in a Distributed Environment. In *Proc. 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008)*. In print, 2008.
5. 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, Directorate-General of the Portuguese Archives, Lisbon, Portugal, 2007.
6. K. Kiesling. Metadata, Metadata, Everywhere - But Where Is the Hook? *OCLC Systems & Services*, 17(2):84–88, 2001.
7. C. A. Lynch, S. Parastatidis, N. Jacobs, H. Van de Sompel, and C. Lagoze. The OAI-ORE effort: progress, challenges, synergies. In *Proc. 7th ACM/IEEE Joint Conference on Digital Libraries, (JCDL 2007)*, page 80. ACM Press, USA, 2007.
8. L. Moreau, P. Groth, S. Miles, J. Vazquez-Salceda, J. Ibbotson, S. Jiang, S. Munroe, O. Rana, A. Schreiber, V. Tan, and L. Varga. The Provenance of Electronic Data. *Commununications of the ACM*, 51(4):52–58, 2008.
9. C. J. Prom. Does EAD Play Well with Other Metadata Standards? Searching and Retrieving EAD Using the OAI Protocols. *J. of Archival Org.*, 1(3):51–72, 2002.
10. 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 Press, USA, 2002.
11. H. Van de Sompel, C. Lagoze, M. Nelson, and S. Warner. The Open Archives Initiative Protocol for Metadata Harvesting (2nd ed.). Technical report, Open Archive Initiative, p. 24, 2003.