An Architecture for Sharing Metadata Among Geographically Distributed Archives

Maristella Agosti, Nicola Ferro, and Gianmaria Silvello

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

Abstract. We present a solution to the problem of sharing metadata between different archives spread across a geographic region. In particular we consider the Italian Veneto Region archives. Initially we analyze the Veneto Region information system based on a domain gateway system called "SIRV-INTEROP project" and we propose a solution to provide advanced services against the regional archives. We deal with these issues in the context of the SIAR – Regional Archival Information System – project.

The aim of this work is to integrate different archive realities in order to provide unique public access to archival information. Moreover we propose a non-intrusive, flexible and scalable solution that preserves archives identity and autonomy.

1 Introduction

The experience gained in the context of the DELOS cooperative activities on service architectures for *Digital Library Systems (DLSs)* has enabled the launch and participation in a new project of interest to the Italian Veneto Region for the management of metadata on public archives distributed throughout the region. The project has been named *Sistema Informativo Archivistico Regionale (SIAR)* and is a project for the design and development of a prototype able to manage metadata of interest for the building of a "Regional Archival Information System". In fact the aim of SIAR is to offer access to archival information which is maintained in several repositories spread across the Veneto Region.

In this study, we discuss how to address the problem of sharing archival metadata stored in different repositories geographically distant from each other.

The Veneto Region archives belong to different kinds of institutions, such as Municipalities; they are managed by different *Information Management Systems (IMSs)*. In this context, we have to satisfy a strong requirement for cooperation and inter-operability, while at the same time preserving not only the autonomy of all these institutions, but also their way of managing and organizing their archives. As a consequence, the different IMSs have to be considered as legacy systems and cannot be modified or changed in order to be integrated together.

Moreover, a central service has to be provided to give external users the possibility of accessing and obtaining archival metadata stored in the regional

C. Thanos, F. Borri, and L. Candela (Eds.): Digital Libraries: R&D, LNCS 4877, pp. 56–65, 2007.

archives. This service should save a coherent way of accessing the archival information and should preserve users from having to physically visit an archive.

Finally, the system proposed has to be integrated into the national telematic infrastructure for the Public Administration, which is being developed in order to provide the inter-operation between different applications of the public administrations.

The paper is organized as follows: Section 2 reports on the Italian National telematic infrastructure for the public administrations which is based on domain gateways, it explains also how SIRV-INTEROP works; this is a project developed by the Veneto Region. Section 3 addresses the design of the SIAR infrastructure and presents a conceptual system architecture which involves the Veneto Region and the regional territory archive keepers. Section 4 presents SIRV-PMH architecture. Section 5 draws some conclusions.

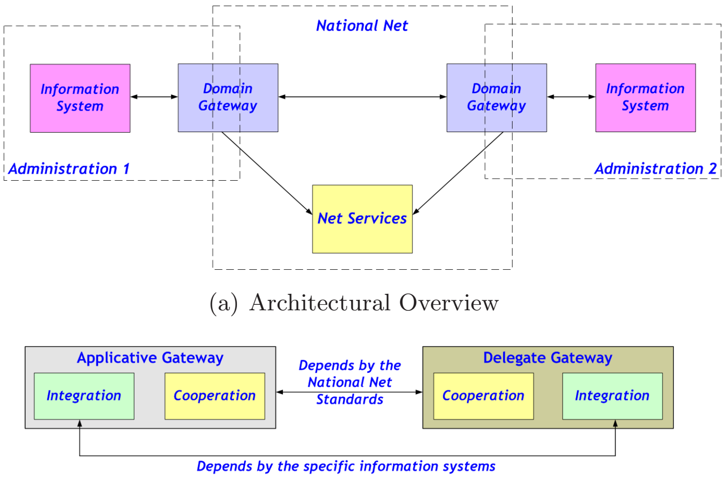
2 The National Telematic Infrastructure for the Public Administration

The Veneto Region participates in the creation of the Italian National Net of services which enables an infrastructure for the interconnection of public administrations. The Veneto Region participates to the National Net by means of its SIRV-INTEROP project. SIRV-INTEROP project implements a Domain Gateway System based on Applicatory Cooperation principles [3]. Through this system the Veneto Region is able to participate in the Italian National Net of services. The Italian National Net improves cooperation and integration between the various administrations and provides various services to external users.

The main goal of the domain gateway system is to integrate the service of different administrations. A highly important issue for this system is to maintain the independence of each cooperative information system. In this way any system that wants to fulfill a service to the Net community could maintain its internal structure unchanged. We define as **domain** a particular organization set of resources and policies. The domain is also considered the organization *responsibility boundary.* The National Net is conceived as a domains federation. Communication takes place through uniform entities (domains) and the main goal of the cooperative architecture is to enable the integration of policies of the informative objects (e.g. data and procedures) and the different domains.

The fundamental element of this system is represented by the modalities through which a server domain exports its services for the clients domains. The technological element for realizing this system is a **Domain gateway**; it has a resources access proxy function. A domain gateway represents the summa of all things necessary for accessing the domain resources, as shown in figure 1(a).

From an architectural point of view, domain gateways are seen as adaptors which enable cooperation between the National Net and many different information systems, as shown in figure 1(b). Domain gateways are divided into two main classes:



(a) Architectural Overview

(b) Internal Structure

Fig. 1. Domain Gateway System

- **Applicative Gateway:** This is the domain gateway which grants services; every domain which can distribute services carries out this function through this gateway. Information systems are connected by applicative gateway through a particular module called *Wrapper*;
- **Delegate Gateway:** Is the domain gateway that requests services to application gateways. Delegate gateway is realized by the information systems that use *net services* to realize the collaboration.

From a logical point of view, every domain gateway is composed of two main components:

- **Cooperation:** Realizes data communications generical functions;
- **Integration:** Realizes the adaptation towards the information systems and guarantees that the applicatory content respects formats and coding policies.

Cooperation components depend on the National Net standards; in contrast integration components depend both on the National Net standards and the characteristics of the information system they have to integrate.

Communication and data exchanges between Applicative and Delegate gateways take place by means of *Simple Object Access Protocol (SOAP)*, protocol which use *eXtensible Markup Language (XML)* technologies to define an extensible messaging framework [4].

Through SIRV-INTEROP project the Veneto Region is able to share services and information between many public administrations. The SIRV-INTEROP project guarantees interoperability, in this way different public administrations can use different information systems.

3 The SIAR Infrastructure

The SIAR project backdrop is characterized by the presence of different institutional needs. To design the SIAR system we have to consider each of these needs. On the one hand we have to guarantee that the local bodies maintain the management autonomy of their archives. On the other hand, however, we have to build-up regional coordination so that we can have an integrated global vision of the local archives participating in SIAR; for this reason the Veneto Region has to define a set of useful rules for the coordination and integration of local archival bodies present in the Region.

We need to guarantee the different juridical subjects autonomy with regard to the archival and information system choices; a possible solution could be a net of autonomous archives which share regional authority lists together with a protocol for metadata exchange. With this kind of solution local archives would exchange with SIAR only the metadata describing their archival resources, SIAR would store the harvested metadata in a central repository. This central repository would constitute the basis of unique access portal to the regional archives. This system would enable a user to identify a particular archival resource and SIAR would provide the information enabling the user to physically or virtually reach the archive containing the resource.

3.1 Integration Requirements

SIAR is constituted by a federation of autonomous archives characterized by unique access point. It will supply advanced research and integration services granting a common starting point; these services could also be implemented at a later date. However, it is important to identify some basic pre-requisites which are fundamental to the integration process:

Authority lists are the first requirement. It is mandatory for local archive coordination and integration. The definition of authority lists is a necessary tool to enable data exchange and integration between two or more subjects. An authority list enable the unique identification of the particular entity it describes. Moreover, authority lists supply an entity shared description; in this way identification and description are common to all the users using the same authority list.

The first step towards integrated access to the resources distributed across the Region is the utilization of a set of authority lists created and defined by the archival conservation subjects (archive keepers) with the coordination of the Veneto Region. SIAR will supply a common access point to the different archival resources distributed through the Veneto Region. It will be a portal which enables an integrated view of the Veneto Region's archival resources and it will be a unique public access point.

Protocol for metadata exchange is another essential requirement. In general, a protocol for data exchange is a protocol that defines a rules set which fixes the data format, the channel and the means of communication. A part of this requirement is the data format choice, which is useful for the metadata exchange between archive keepers and the Veneto Region.

Collaboration of local bodies. Different archive keepers could obtain a benefit from the common authority lists defined by the Veneto Region. Moreover, they should form metadata following the rules defined by the common protocol chosen by the Veneto Region.

3.2 Conceptual Architecture

We have to consider that SIAR is an institutional project and that there has been a recent digital administration legal framework. Also for these reasons, we think it is important to propose an architecture based on standards and on open source softwares. Standards enable the development of interoperable systems which are also based on a methodological study which guarantees a long lifetime to the project itself. The use of open source tools is desirable both because it is consistent with the most recent legal dispositions about digital administration and because it is supported by a community of developers and users which guarantee its development and analysis in a continuous way.

As we have just mentioned, the international initiative called *Open Archives Initiative (OAI)* is very important in an archival context. The main goal of OAI is to develop and promote interoperability standards to facilitate the dissemination of content, in this case the archival contents. The SIAR project is an occasion to promote local archives to disseminate their contents. In this context OAI could be the right choice for setting up the methodological and technological equipment useful for designing a system which manages and shares archival contents.

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [2] allows us to achieve a technological integration of the information systems participating in the SIAR project. OAI-PMH is based on the distinction between the roles of the participants; they can be a *Data Provider* or a *Service Provider*. In our case study the Veneto Region is the Service Provider, because it is the subject which gives advanced services such as data and public access to them. Archive keepers are seen as Data Providers because they supply archive metadata. These metadata will be harvested, stored and probably indexed by the Veneto Region in order to provide services.

As we can see in Figure 2 the Veneto Region has to get harvester software, instead archive keepers have to get repository software which answers the harvester requests. Repository software has to prepare and send metadata in a specific and agreed format.

As we have just seen, archive keepers autonomy is very important, in this way there could be the presence and the co-existence of many different archive management information systems. There will be the need to propose an automatic or manual procedure to import metadata from the different keepers systems inside repositories which will be harvested by the Veneto Region harvester software.

Moreover, archival integration between the different archive keepers will be fundamental; this is possible by means of the Veneto Region guidelines which have to be shared between them. The Veneto region has to define the standards

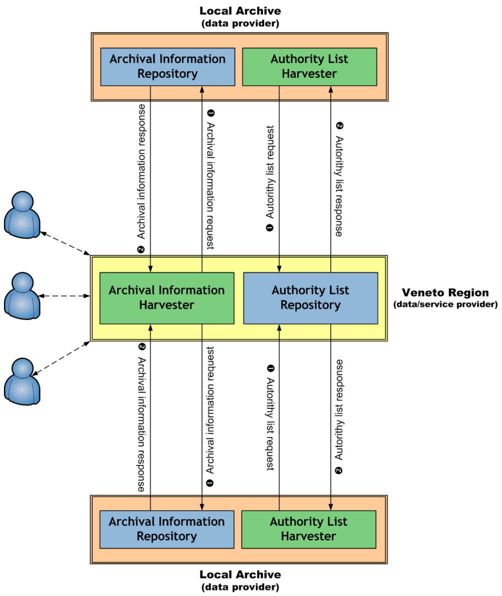

Local Archive
(data provider)
Archival Information Repository
Authority List Harvester
❷ Archival information response
❶ Archival information request
❶ Autorithy list request
❷ Autorithy list response
Archival Information Harvester
Authority List Repository
Veneto Region
(data/service provider)
❷ Archival information response
❶ Archival information request
❶ Autorithy list request
❷ Autorithy list response
Archival Information Repository
Authority List Harvester
Local Archive
(data provider)

Fig. 2. Conceptual Architecture of SIAR

for the archival descriptions and it has to produce and keep the authority lists. Authority list sharing assures a first degree of integration, this could be used not only for metadata aggregation but also for constituting a common public access point to the Veneto Region archive information. In this context archive keepers need a mechanism enabling them to obtain the authority lists. OAI-PMH could be used on this occasion too; archive keepers would have harvester software whereas the Veneto Region would have repository software for managing the authority files. In this system the Veneto Region would be a Service Provider when it offers advanced services on metadata and a Data Provider when it keeps and delivers authority files to the archive keepers. We can see these peculiarities in the conceptual architecture design in Figure 2.

4 Integration of OAI-PMH into the National Telematic Infrastructure

In this section we propose a solution capable of adapting OAI-PMH functionalities to the domain gateway system. We perform OAI Data and Service Provider as domain gateways, in this way the two systems can be adapted to each other without any particular modification.

The main idea is that the Service Provider works as a delegate gateway which takes user requests and requires services to the various applicative gateways. Users require archive metadata by means of a portal system and so the delegate gateway harvests metadata records from the various repositories using the six

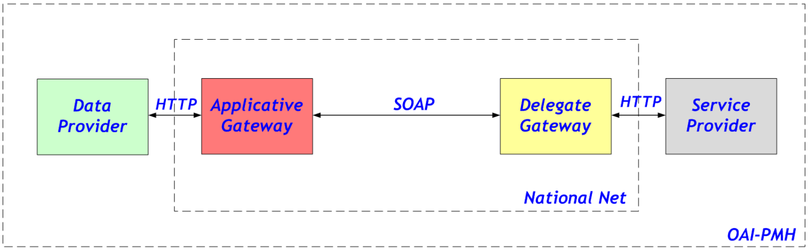


Fig. 3. OAI-PMH over National Net: General View

OAI-PMH verbs. We can say that the delegate gateway harvests the repositories which look like applicative gateways. OAI Data Providers are seen as applicative gateways which supply services; in this context they answer to the six verbs of the metadata harvesting protocol. Metadata requests and responses occur between applicative and delegate gateways through the SOAP protocol [4]. From this point of view, archives are open to the OAI-PMH and participate in the National Net of services.

In Figure 3 we can see how the National Net, which is based on the exchange of XML messages, could be used to harvest metadata by means of OAI-PMH. Service/data providers and domain gateway communication takes place by HTTP post/get method. In contrast data communication between domain gateways takes place by means of a SOAP protocol. This consideration shows that the two systems do not change their internal functioning, indeed they always use their default transport protocols.

We have to make a few additions to the Veneto Region system: we have to add a delegate gateway for the service provider and an applicative gateway for each repository participating in the system. We utilize an applicative gateway for each repository because different repositories constitute different domains and different domains implicate different domain gateways. In this way, repositories participating in the National Net can offer other services besides those provided by OAI-PMH. In addition, the service provider maintains its OAI-PMH functions and the delegate gateway works as an *"adaptor"* between OAI-PMH requests and the National Net. The delegate gateway harvests metadata; it crowns OAI-PMH requests inside XML SOAP messages. The applicative gateway is connected to the data provider and offers services to the National Net by means of wrappers around the information systems which are the base of a specific service; in this context applicative gateways interface data providers.

4.1 A Conceptual Architecture

The biggest issue for integrating OAI-PMH inside the National Net is to carry OAI-PMH requests by SOAP protocol and to do the same with the responses.

In this case the principal role of domain gateways is to encapsulate the requests or the responses into SOAP envelopes. On the other hand, domain gateways have

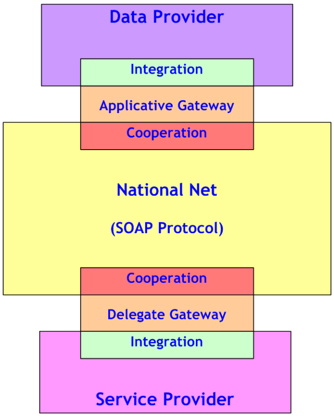


Fig. 4. The Role of Integration and Cooperation Components

to extract OAI-PMH requests and responses from their SOAP envelope. Domain gateways are composed of two main components: *cooperation* and *integration*. These two parts are helpful for addressing our problem; Figure 4 is a visual representation of how they act and shows how cooperation and integration components work as interfaces between the National Net and OAI Data/Service providers.

Now we have to consider how to implement solutions to integrate OAI-PMH with the National Net without making any substantial modifications to the two systems; the fundamental idea is to use OAI-PMH as a layer over SOAP [1].

4.2 SIRV-PMH Design

SIRV-PMH represents the joining of the SIRV-INTEROP project and OAI-PMH. Integration is a logical component which on the one hand acts as OAI Data Provider and on the other hand acts as an OAI Service Provider. In particular, the Integration component in the Delegate Gateway is composed of:

- **OAI Data Provider:** Receives OAI requests and answers with the required metadata;
- **OAI-SIRV Wrapper:** Encapsulates the OAI request inside a SOAP message consistent with the National Net.

Instead the Integration component in Applicative Gateways is composed of:

- **OAI-SIRV Wrapper:** Extracts the OAI request from the SOAP envelopment;
- **OAI Service Provider:** Sends OAI requests to the OAI Data Provider and receives the required metadata.

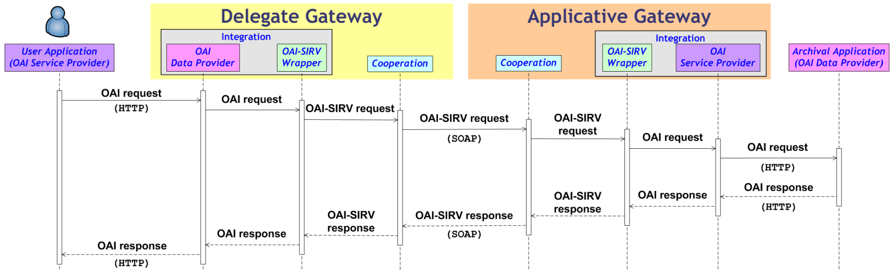


Fig. 5. SIRV-PMH Functioning

We can see how this system works by analyzing how OAI Service and Data Provider exchange metadata between each other.

The Applicative Gateway has to receive an XML SOAP message using the *cooperation component*; by the use of the *OAI-SIRV wrapper* it has to remove SOAP tags from the message and extract the OAI-PMH request. The extracted request is sent by means of the *OAI Service Provider Integration component* to the OAI Data Provider which has to process the OAI-PMH request and query the repository to obtain the required information. Afterwards, it has to build up an OAI-PMH response. When the response is ready, it has passed again to the Applicative Gateway which has to receive the response, add SOAP tags and send the XML SOAP message through the National Net.

The delegate Gateway and Service Provider have a similar role and more or less work in the same manner. In fact the Service Provider has to formalize an OAI-PMH request and pass it to the Delegate Gateway. The Delegate Gateway has to receive the OAI request, encapsulate the OAI-PMH request inside an XML SOAP message and send it to the Applicative Gateway by means of the National Net. Moreover, the Delegate Gateway has to receive the answer message by the Applicative Gateway, remove SOAP tags, extract the OAI-PMH response and send it to the OAI Service Provider.

If we consider a typical OAI-PMH request, for example a ListIdentifier [2] which harvests records headers from a repository, we can see how this request goes from the OAI Service Provider to the OAI Data Provider: an OAI Service Provider builds up a ListIdentifier request and sends it to the Delegate Gateway. The Delegate Gateway receives the request by means of its integration component. The OAI-SIRV Wrapper adds SOAP tags to the request so that it can be sent to the correct Applicative Gateway through the National Net. When the Applicative Gateway receives the request, it can extract the OAI-PMH request which can then be sent to the specificated Data Provider by means of a Service Provider. The Data Provider processes the request and builds up the response which follows the inverse procedure to reach the Service Provider. We can see SIRV-PMH functioning in Figure 5.

In this way the metadata contained in Data Providers can also be harvested by the Service Providers which do not participate in the National Net of services.

5 Conclusions

In this work we have presented an information system which addresses the problem of sharing archive metadata between different repositories geographically distant from each other. With this system, we can both preserve the autonomy of archive systems and provide a unique central public access point to the information they contain. We proposed a solution which integrates the Veneto Region system with an advanced, flexible and widely adopted protocol which is OAI-PMH. Moreover, we have seen how OAI-PMH can be adopted to work within SOAP and with different information systems. This system is called SIRV-PMH and would enable broad access to archive information which otherwise could only be reached by physically visiting the archives. SIRV-PMH does not modify the internal functioning of OAI-PMH and the SIRV-INTEROP project; it integrates these systems to provide advanced services on archives.

Now we have to implement the OAI-SIRV wrapper module and experimentally verify the efficiency of the SIRV-PMH system. The Data and Service Provider softwares also need to be taken into account, and we are evaluating *Online Computer Library Center (OCLC)* OAICat[1] and OCLC Harvester2[2] open source software tools. We have to verify if these software tools are truly effective for our purposes and if there is the need to adapt, add or change some of their functionalities.

Acknowledgements

The necessary background work of the study reported in this paper has been partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

The study is partially supported by a grant from the Italian Veneto Region.

References

1. Congia, S., Gaylord, M., Merchant, B., Suleman, H.: Applying SOAP to OAI-PMH. In: ECDL, pp. 411–420 (2004)
2. Van de Sompel, H., Lagoze, C., Nelson, M., Warner, S.: The Open Archives Initiative Protocol for Metadata Harvesting. Technical report (2004)
3. Gazzetta Ufficiale N. 78 del 3 Aprile 2002 ALLEGATO n. 2: Rete Nazionale: caratteristiche e principi di cooperazione applicativa
4. Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J., Nielson, H.F.: SOAP Version 1.2 Part 1: Messaging Framework and Part 2: Adjuncts. Technical report (2003)

[1] `http://www.oclc.org/research/software/oai/cat.htm`

[2] `http://www.oclc.org/research/software/oai/harvester2.htm`