An Architecture to Share 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 address the problem of sharing metadata between different archives spread across a geographic region. In particular we consider archives of the Italian Veneto Region. Initially we analyze the Veneto Region information system based on domain gateway system called 'SIRV-INTEROP project" and we propose a solution to provide advanced services against the regional archives. We face these issues in the context of a the SIAR – Regional Archival Information System – project.
The aim of this work is to integrate different archive realities in order to provide an unique public access to archival information. Moreover we propose a non-intrusive, flexible and scalable solution that preserves archives identity and autonomy. This system is called SIRV-PMH because it joints together SIRV-INTEROP project and *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)*. It permits to address the problem of metadata sharing, and in the future it could be integrated inside the Italian National Net of services for Public Administration.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: H.3.7 Digital Libraries.

General Terms
Systems, Architecture.

Keywords
Digital Library Architecture, Digital Library Service, Digital Archives, Open Archive Initiative (OAI)

1 Introduction

The experience gained in the context of the DELOS cooperative activities on service architectures for *Digital Library Systems (DLSs)* has permitted the launching and participation to a new project of interest of the Italian Veneto Region for the management of metadata on public archives distributed over the region. The project has been named *Sistema Informativo Archivistico Regionale (SIAR)* that 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 that are maintained in several repositories spread across the Veneto Region territory.

In this study, we discuss how to address the problem of sharing archival metadata stored into different repositories geographically far one 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: the autonomy of all these institutions has to be preserved as well as their way of managing and organizing the 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 to external users the possibility of access and obtain archival metadata stored in the regional archives. This service should provide a coherent way of accessing the archival information and should preserve users from having to physically visit an archive.

Finally, the proposed system 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 the 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 that is based on domain gateways, it explains also how works the SIRV-INTEROP project developed by the Veneto Region. Section 3 addresses the design of the SIAR infrastructure and presents a conceptual architecture of the system which involves the Veneto Region and the archive keepers in the regional territory. Section 4 presents the SIRV-PMH architecture. Section 5 draws some conclusions.

2 The National Telematic Infrastructure for Public Administration

Veneto Region participates to the Italian National Net of services that permits to create an infrastructure for public administrations interconnection. Veneto Region participates to the National Net by means of its SIRV-INTEROP project that implements a Domain Gateway System based on applicatory cooperation principles [1].

The SIRV-INTEROP project implements a Domain Gateway System based on Applicatory Cooperation principles. Veneto Region through this system could participates to the Italian National Net of services that improves cooperation and integration between the various administrations and provides various services to external users.

The domain gateway system main goal is to integrate different administrations services. A big important issue for this system is to maintain the independence of each single information system that cooperates. In this way any system that wants to fulfill a service to the Net community could maintain inalterate its internal structure. We define as **domain** the set of resources and policies of a particular organization. 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 the informative objects (e.g. data and procedures) and the different domains policies.

The fundamental element of this system is represented by the modalities through which a server domain exports its services for the clients domains. **Domain gateway** is the technological element to realize this system; it has a proxy function for the resources access. Domain gateway represents the summa of all that necessaries things to access domain resources.

From an architectural point of view, domain gateways are seen as adaptors that permit the cooperation between the National Net and many different information systems.

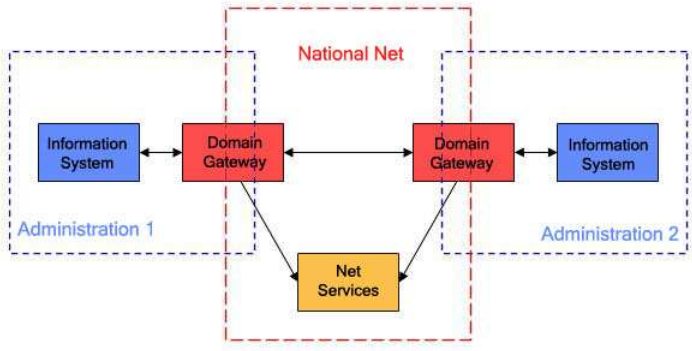

Fig. 1: Domain Gateway System

Domain gateways are divided into two main classes that are:

- **Applicative Gateway**: is the domain gateway that grants services; every domain that can distribute services carries out this function through this gateway. Applicative gateway interfaced information systems 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 the *net services* to realize the collaboration.

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

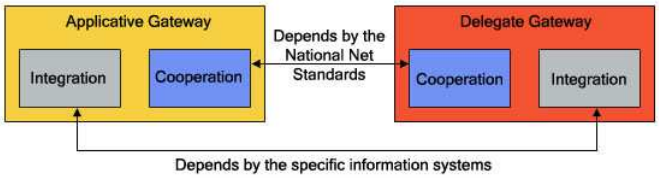

Fig. 2: Internal Structure of Domain Gateway System

- **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 by the National Net standards, instead integration components depend both by the National Net standards and by the characteristics of the information system that they have to integrate.

Communications and data exchanges between Applicative and Delegate gateways take place by means of *Simple Object Access Protocol (SOAP)* protocol that uses *eXtensible Markup Language (XML)* technologies to define an extensible massaging framework [2].

Veneto Region through SIRV-INTEROP project could share services and information between many public administrations. SIRV-INTEROP project guarantees interoperability, so different public administrations could use different informative systems.

3 The SIAR Infrastructure

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

We have the necessity to guarantee the autonomy of the different juridical subjects for what is concerned with the archival and the information systems choices; a possible solution could be constituted by a net of autonomous archives that share regional authority lists and by a protocol for metadata exchange. With this kind of solution local archives would exchange with SIAR only the metadata that describe their archival resources and SIAR would store the harvested metadata in a central repository. This central repository would constitute the basis of an unique access portal to the regional archives. This system would permit to an user to individuate a particular archival resource and SIAR would give the information to physically or virtually reach the archive that contains the resource.

3.1 Integration Requirements

SIAR is constituted by a federation of autonomous archives characterized by an unique access point. It will supply advanced research and integration services granting a common starting point; these services could be implemented also in subsequent times. However it is important to individuate some basic prerequisites which are fundamental to the integration process.

Authority lists are the first requirement. It is necessary for the coordination and the integration of local archives spread across the regional territory. The definition of authority lists represents a required tool that permits data exchange and integration between two or more subjects. An authority list permits to uniquely identify the particular entity that it describes. Moreover authority lists supply a shared description of an entity; in this way identification and description are common to all the user that use the same authority list.

The first step towards an integrated access to the resources distributed across the regional territory is the utilization of a set of authority lists created and defined by the archival conservation subjects (archival keepers) with the coordination of Veneto Region.

SIAR will supply a common access point to the different archival resources distributed in the Veneto Region territory. It will be a portal that permits an integrated view of the Veneto Region's archival resources and it will represent an unique public access point to them.

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

Local bodies collaboration: Different archive keepers could obtain a benefit from the common authority lists defined by Veneto Region. Moreover they should form metadata following the rules defined by the common protocol chosen by Veneto Region.

3.2 Conceptual Architecture

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

As we have just said before, the international initiative called *Open Archives Initiative (OAI)* is very important in an archival context. The main goal of OAI is develop and promote interoperability standards to facilitate the dissemination of content, in this case the archival contents.

The SIAR project is an occasion to enhance local archives to disseminate their contents. In this context OAI could be the right choice to dispose of a methodological and technological equipment useful for designing a system that manages and shares archival contents.

OAI-PMH [3] allows us to realize a technological integration of the information systems that participate in the SIAR project. OAI-PMH is based on the distinction between the participants roles; they can be a *Data Provider* or a *Service Provider*. In our case study Veneto Region is the Service Provider, because it is the subject that 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 Veneto Region in order to provide services.

As we can see in Figure 3 Veneto Region has to get an harvester software, instead archive keepers have to get a repository software that answers to the harvester requests. Repository software has to prepare and to send metadata in a specific and agreed format.

As we have just seen before, 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 necessity to propose an automatic or manual procedure to import metadata from the different keepers systems inside repositories that will be harvested by Veneto Region harvester software.

Moreover an archival integration between the different archive keepers will be fundamental; this is possible by means of the Veneto Region guidelines that have to be shared between them. Veneto region has to define the standards for the archival descriptions and it has to produce and to keep the authority lists. Authority lists sharing assures a first degree of integration, that could be used not only for metadata aggregation but also to constitute a common public access point to Veneto Region archive information. In this context archive keepers need for a mechanism that permits them to obtain the authority lists. Also in this occasion OAI-PMH could be used; so archive keepers would have an harvester software instead Veneto Region would have a repository software to manage the authority files. In this system 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 3.

4 Integration of OAI-PMH into the National Telematic Infrastructure

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

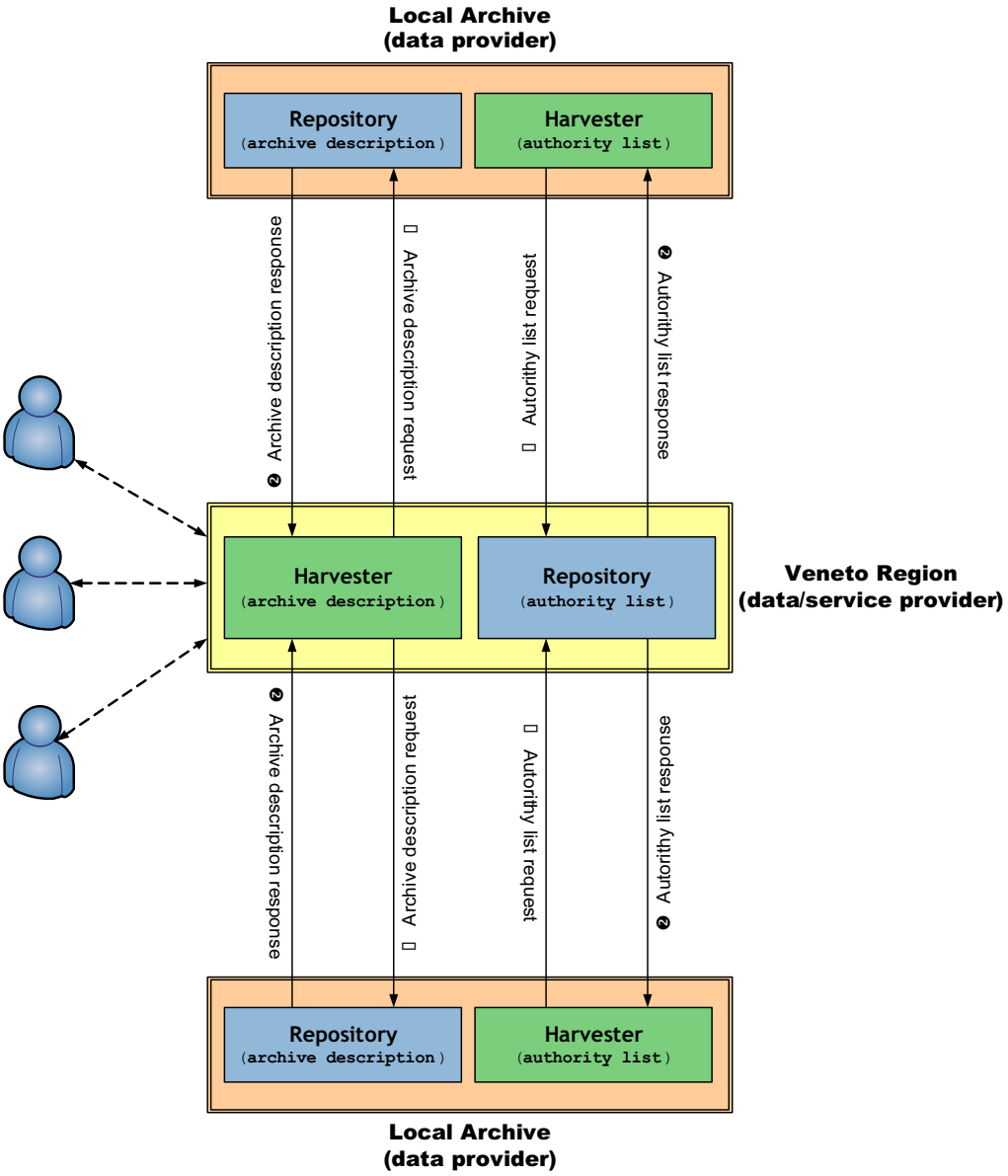

Fig. 3: Conceptual Architecture of SIAR

The main idea is that the Service Provider works as a delegate gateway that takes users' requests and requires services to the various applicative gateways. Users require archive metadata by means of a portal system and so the delegate gateway uses the six OAI-PMH verbs to harvests metadata records from the various repositories. We can say that delegate gateway harvests the repositories which look like applicative gateways.

OAI Data Providers are seen as applicative gateways that 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 SOAP protocol [2].

From this point of view, archives are open to the OAI-PMH and participate at the National Net of services.

In Figure 4 we can see how the National Net, which is based on the exchange of XML messages, could be used to harvest metadata with OAI-PMH. Communications between service/data providers and domain gateways occur by HTTP post/get method instead data communication between domain gateways occur by means of SOAP protocol. This consideration shows that the two systems do not change their internal functioning, indeed they always use their default transport protocols.

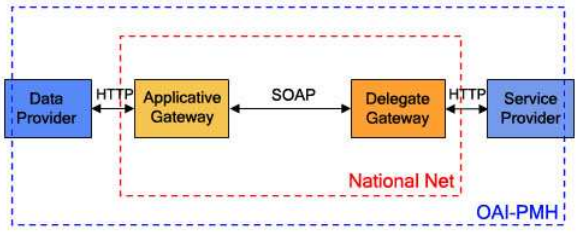

Fig. 4: OAI-PMH over National Net: General View

We have to do a few addictions to the Veneto Region system: we have to add a delegate gateway for the service provider and an applicative gateway for each repository that participates at 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 that participate to the National Net can offer other services besides those provided by OAI-PMH. In addition, the service provider maintains its OAI-PMH functions and delegate gateway works as an *"adaptor"* between OAI-PMH requests and the National Net. Delegate gateway harvests metadata; it crowns OAI-PMH requests inside XML SOAP messages.

Applicative gateway is connected to 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 to integrates OAI-PMH inside of the National Net is to carry OAI-PMH requests by SOAP protocol and do the same with the responses.

In this case the principal role of domain gateways is to incapsulate the requests or the responses into SOAP envelops. On the other side domain gateways have to extract OAI-PMH requests and responses from their SOAP envelop.

Domain gateway as we have just seen are composed by two main components that are *cooperation* and *integration*. These two parts are helpful to address our problem; in Figure 5 we can have a visual idea of how they act.

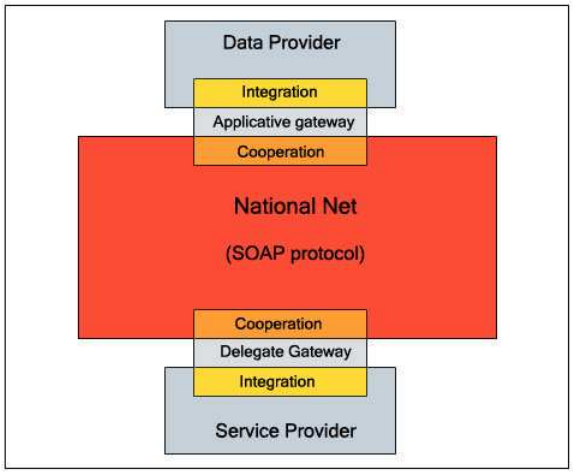

Fig. 5: The Role of Integration and Cooperation Components

Figure 5 clarifies the role of domain gateways and shows how cooperation and integration components work as interfaces between National Net and OAI Data/Service providers.

Now we have to consider how to implement solutions to integrate OAI-PMH with National Net without any substantial modifications of the two systems; the fundamental idea is to uses OAI-PMH as a layer over SOAP [4].

4.2 SIRV-PMH Design

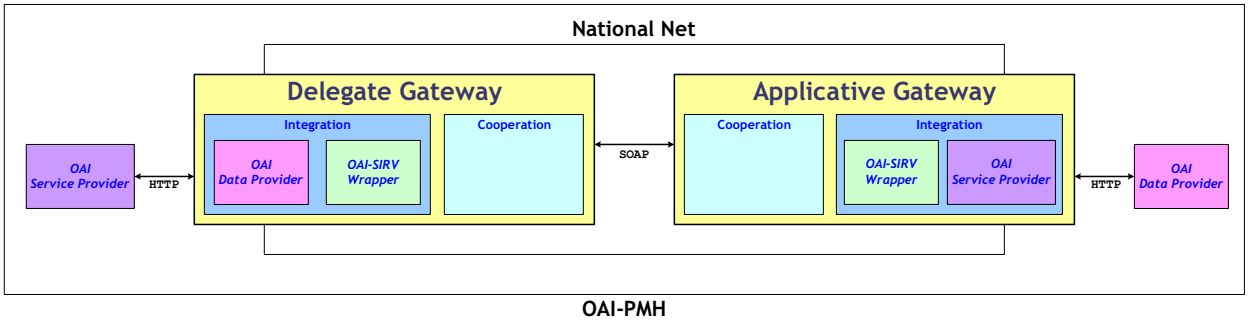

Fig. 6: SIRV-PMH Architecture Design

SIRV-PMH represents the union of the SIRV-INTEROP project and OAI-PMH. In Figure 6, we can see how SIRV-PMH is composed. Integration is a logical component which on one side acts as OAI Data Provider and on the other side acts as OAI Service Provider. In particular in the Delegate Gateway, the Integration component is composed by:

- **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 in Applicative Gateways the Integration component is composed by:

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

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

In this analysis of gateways operations we put into brackets the part (Cooperation or Integration components) that does the operation.

An Applicative Gateway has to:

1. receive a XML SOAP message (*Cooperation*);
2. remove SOAP tags from the message and extract OAI-PMH request (*OAI-SIRV Wrapper*).

The extracted request is sent by means of the O*AI Service Provider Integration component* to the OAI Data Provider that has to:

1. process OAI-PMH request;
2. query the repository to obtain required information and build-up an OAI-PMH response.

When the response is ready, it has passed again to the Applicative Gateway that:

1. receive the response (*OAI Service Provider*)
2. adds SOAP tags (*OAI-SIRV Wrapper*);
3. sends XML SOAP message through the National Net (*Cooperation*).

Delegate Gateway and Service Provider have a similar role and more or less work in the same manner. Service Provider has to:

1. formalize an OAI-PMH request;
2. pass it to the Delegate Gateway.

Delegate Gateway has to:

1. receive OAI request (*OAI Data Provider*)
2. encapsulate OAI-PMH request inside a XML SOAP message (*OAI-SIRV Wrapper*);
3. send the message to the Applicative Gateway by means of the National Net (*Cooperation*);
4. receive the answer message by the Applicative Gateway (*Cooperation*);
5. remove SOAP tags from the message and extract OAI-PMH response (*OAI-SIRV Wrapper*);
6. send the OAI response to the OAI Service Provider (*OAI Data Provider*)

If we consider a typical OAI-PMH request, for example a ListIdentifier [5] which harvest 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. This one receives the request by means of its integration component that in Figure 6 is represented by a Data Provider. OAI-SIRV Wrapper adds SOAP tags to the request so that it could be sent to the correct Applicative Gateway through the National Net. When Applicative Gateway receives the request, it could extract the OAI-PMH request that could be sent to the specificated Data Provider by means of a Service Provider. Data Provider elaborates the request and builds-up the response that follows the inverse procedure of the request to reach the Service Provider. We can see SIRV-PMH functioning in Figure 7.

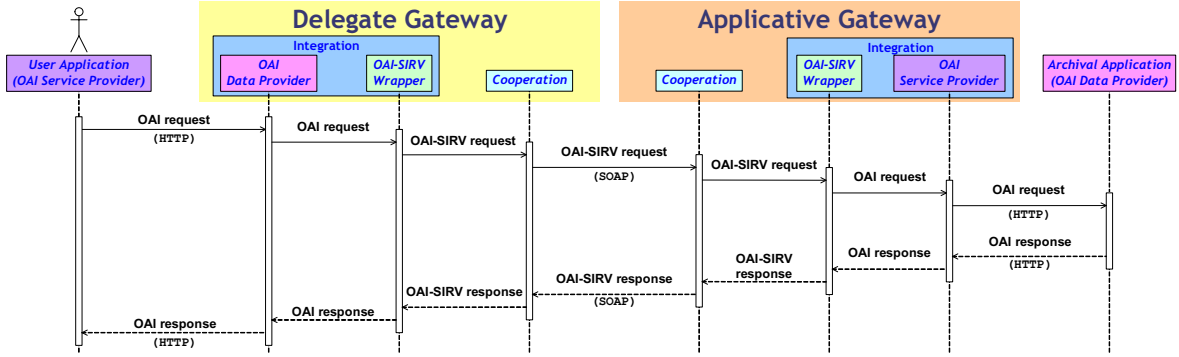

Fig. 7: SIRV-PMH Functioning

In this way the metadata contained in Data Providers could be harvested also by Service Providers that do not partecipate to the National Net of services.

5 Conclusions

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

Now we have to implement OAI-SIRV wrapper module and experimentally verify the efficiency of SIRV-PMH system. Also Data and Service Provider softwares 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 necessity 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), in particular in the context of the activities related to Task 1.1 and Task 1.5 of Work Package 1 on Digital Library Architecture.

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

References

1. Gazzetta Ufficiale N.78 del 3 Aprile 2002. ALLEGATO n. 2: Rete Nazionale: caratteristiche e principi di cooperazione applicativa.
2. M. Gudgin, M. Hadley, N. Mendelsohn, J. Moreau, and H. F. Nielson. SOAP Version 1.2 Part 1: Messaging Framework and Part 2: Adjuncts. Technical report, 2003.
3. OAI. The Open Archives Initiative Protocol for Metadata Harvesting – Version 2.0. `http://www.openarchives.org/OAI/openarchivesprotocol.html` [last visited 2006, October 2], October 2004.
4. S. Congia, M. Gaylord, B. Merchant, and H. Suleman. Applying SOAP to OAI-PMH. In *ECDL*, pages 411–420, 2004.
5. H. Van de Sompel, C. Lagoze, M. Nelson, and S. Warner. The Open Archives Initiative Protocol for Metadata Harvesting. Technical report, 2004.

[1] **OCLC OAICat**: http://www.oclc.org/research/software/oai/cat.htm [last visit January 21, 2007]
[2] **OCLC Harvester2**: http://www.oclc.org/research/software/oai/harvester2.htm [last visit January 21, 2007]