

Design of a federation service for digital libraries: the case of historical archives in the PORTA EUROPA Portal (PEP) Pilot Project

Marco Pirri, Maria Chiara Pettenati, Dino Giuli

m.pirri@telemat.det.unifi.it, pettenati@achille.det.unifi.it, giuli@det.unifi.it

Authors are with the Electronics and Telecommunications Department, University of Florence
Via Santa Marta 3 50139 Florence (IT)

Abstract

Access to distributed and heterogeneous Internet resources is coming up as one of the major problem for future development of the next generation of Digital Libraries. Available data sources vary in terms of data representation and access interfaces, therefore a system for federating heterogeneous resources accessible via the Web is considered to be a crucial aspect in digital libraries research and development. Libraries as well as institutions and enterprises are struggling to find solutions that can offer the final user an easy and automatic way to rapidly find relevant needed resources among heterogeneous ones.

Our project starts from the recent results of Dublin Core Metadata Initiative (DCMI) and in particular from the Dublin Core Recommendations (DCMIR).

This paper reports our analysis of three different digital historical archives maintained by the European University Institute (EUI) in Florence and its mapping using a common Meta Resource Card based on Dublin Core Elements (DCMES). This situation requires careful consideration of interoperability issues related to uniform naming, metadata formats, document models and access protocols for the different data sources.

We also present our Porta Europa Portal (PEP) federated architecture that will support an XML Dublin Core implementation and in our aim should be easily open to RDF future support. The PEP pilot project specialised portal should provide high quality information, selected according to the criteria of originality, accuracy, credibility together with the cultural and political pluralism derived from the EUI's profile. The information in Porta Europa will be relevant, reliable, searchable and retrievable.

Keywords: Federated service, digital libraries, Dublin Core, Metadata, interoperability

1. Introduction

The integration of existing digital libraries and electronic catalogues of publication is considered to be one of the major issues for the digital library community. The purpose of digital library integration is to devise a proper architecture, a metadata structure and a suitable protocol to:

- provide a uniform interface hiding the specific features and restrictions of the single sources
- supply integrated view on the data.

These issues are tied to two main aspects (Endig et al. 2000):

- the access to data sources (the digital library) depends on the query interface and capabilities of specific data source which have therefore to be carefully described
- a specific data format is used in each single digital library, therefore mapping into a common format is required.

State of the art in digital libraries has shown an evolution of data integration approach along two main directions (Hanani & Frank 2000): from the Stand-alone Digital Libraries to Federated Digital Libraries. In the first case the Digital Library is maintained by a single institution and the data collection is self contained while the material is localised and centralised. The second case is related to a federation of several independent Digital Libraries in the network, possibly organised around a common theme or topic. The Federated Digital Library regroups many autonomous Stand-alone Digital Libraries forming a networked library accessible through a unique user interface.

The digital library federation service approach is therefore adopted to cope with this issues of data integration where the need of regrouping different Stand-alone Digital Libraries arises such it is the case of this project. It is worth remarking that, even if the archives are managed by a single institution, such it is our case, the digital libraries are considered to be stand-alone because of their heterogeneity in metadatada, document formats and access interfaces as it will be more clearly explained in the sequel.

The interoperability issue is consequently decomposed in the sub-problems related to uniform naming, metadata* formats, document models and access protocols.

* Paper Accepted for presentation at Dublin Core Conference, Florence, October 13-17 2002

This paper reports on the preliminary study for the design of a federation services for the integration of three different digital libraries (here also referred as *data sources*) – three heterogeneous archives related to historical topics – whose access has to be made uniform through a single portal - the Porta Europa Portal.

2. The History Pilot Project - The Porta Europa Portal

The PEP (Porta Europa Portal) Pilot Project refers to the integration of three digital libraries related to European history topics: Voices on Europe, Virtual Library and Biblio library catalogue.

Each of these data source is characterized by:

- a collection of data objects (digitized audio, html pages, records...) available locally or through the network
- a collection of metadata structures
- a collection of services (access methods, management functions, logging/statistics, etc.)
- a domain focus (topic)
- a community of users

The need of integrating the three data sources comes from the topic (European history) and users community which are common to all three archives.

- **Voices on Europe;** (<http://wwwarc.iue.it/webpub/Welcome.html>) Voices on Europe is an archive containing the electronic audio version and electronic transcriptions about a hundred of interviews given by outstanding politician and historians.
- **WWW-VL (Virtual Library) on European History Integration;** (<http://vlib.iue.it/history/index.html>) The Virtual Library (VL) is the Web oldest catalogue, conceived by Tim Berners-Lee. Unlike commercial catalogues, it is run by a loose confederation of volunteers, who compile pages of relevant links for specific areas in which they are expert. The EUI Library Web site contains the complete list of VLs belonging to the **WWW VL History Project** in the University of Lawrence/Kansas (USA) and mirrored at the European University Institute's Library (EUI).
- **Biblio (the EUI historical archives);** (<http://www.iue.it/LIB/Catalogue/>) This is the library catalogue containing more than 250.000 bibliographic records. Access to resources is supported by INNOPAC, well known Library Automation System (INNOPAC).

Characteristics of the archive	VOICES ON EUROPE	VIRTUAL LIBRARY	BIBLIO Library Catalogue
Data objects	Digitized audio-video tapes Interviews written transcription (pdf)	HTML pages	Records
Collection of metadata structures	The archive is organised in Access Database	The archive is structured in Web pages	The archive is maintained in a proprietary database in USMARC format
Collection of services	The access to the interviews is currently performed via a Web interface through SQL queries. Resource management is allowed directly on the database. No logging or statistic functions are allowed.	The access is performed through the Web, maintenance and updating of the information is managed through the Web by a project administrator. No logging or statistic functions are allowed	Information management functions are performed through INNOPAC Library automation system.
Domain focus (topic)	European history		
Community of users	Everybody for information search On a case basis, restricted access for full documents consultation Administrators for information management		

Table 1: Main properties of the three data sources

As it is remarkable by the properties illustrated in Table 1, the heterogeneity of the three data sources are due to their difference in the types of data objects, in the collection of metadata structures and in the collection of services provided by each access interface. It is therefore clearly outstanding the need to provide a federation system to integrate access and management of the archives.

3. The PEP Project development phases

The cultural and operational context of the European University Institute and the presence of a top class library in the social sciences with an emphasis on European issues brought to the idea of building a **specific Portal Project** integrated inside the EUI Web Site and offering opportunities to link the currently dispersed European oriented information sources and to contribute also to a better visibility of the Institute. The proposal is to create a specialized portal - **Porta Europa** - which should answer to this need and position the Institute itself on the Web as a leader in the "European debate" and as a natural gateway, a logical point of access to high quality information on European issues.

To test the feasibility and the impact of the PEP project the EUI committed itself to the development of a PEP prototype concerning historic topics. To this extent, among the various available digital historical archives three of them were chosen for the implementation of the pilot, as described in the previous paragraph.

The PEP Pilot Project is being developed according to the following steps:

1. Analysis of the three data resources

In this part we analysed the current situation of the resources and we identified the main issues involved in each case. Each resource is characterised by different issues which have been elicited and therefore faced (see Table 1). This phase ended with a detailed description of the metadata formats, document models and access protocols for each of the data sources. The analysis revealed the strong points and the weakness of each digital library setting the basis for the definition of a common document description model. More specifically we defined a Meta Resource Card (MRC) with a detailed mapping of the relevant fields derived by each resource. Table 2 illustrates a synthesis of the MRC where each archive single fields are more detailed in the related internal reports to be shortly published by the EUI library (Pirri and Noiret 2002) (Pirri and Terzuoli 2002) (Pirri and Baglioni 2002).

Dublin Core Element	Voices on Europe	Virtual Library	Biblio
Title	Interviewee's surname/name	Title	Title
Creator	Name of Interviewer	Author	Author
Subject	Level 1,2,3 (eurovoc)	Type 3	Subject
Description	Full text Interview	Abstract	Note
Publisher	Eui	Type 1	Imprint
Contributor	Not used	Not used	Not used
Date	Date of recording	Date of insertion	Date of publication
Type	Video/Audio/Testo	Text (Html)	Text
Format	Pdf	Html	Pdf
Identifier	Url	Url	Isbn
Source	Not used	Not used	Not used
Language	Language	English	Lang
Relation	Additional Material	Not used	Not used
Coverage	Not used	Not used	Not used
Rights	User Profile	Free	User Profile

Table 2 – Resources Mapping in Meta Resource Card

2. Definition of the federation architecture

After the first phase, the analysis and definition of the federation architecture has to be covered. According to what available in literature (Endig et al. 2000) we agreed on the conceptually layered architecture described in paragraph 4, where each layer has to provide/use specific operation to/from adjacent layers. The objective of the federation services architecture is to provide uniform interface to the individual resources and to supply an integrated view on the data. Therefore the architecture must be conceived in order to accept queries on the global view (uniform data model), decompose them and translate them to allow processing from the single data sources.

3. Definition of the user roles for information access

Due to the variety of information accessible through the different digital libraries of this project, an important step consists in the definition of the users role and access rights.

For the scope of this project, we can identify the functions reported in Table 3 which are to be associated to the related users in order to allow the maximum flexibility in the management and access to the resources.

Users functions and roles have been used in the archives analysis phase as for the Dublin Core Rights field and will also be used in the next development of the project.

Function	User
General Administration (information management)	Administrator and Project Leaders
Information search	Public
Full information access	Internal users (IUE member, professors, students, etc.)
Restricted information access (restriction is due to property right on some resources contained in the archives)	External users, groups of users
Personalised services	Registered users

Table 3: Users and related roles for information access

4. The PEP federated architecture

The architecture of our federation service (Endig 2000) is structured in three layers: the *data source layer* where all information is stored with autonomy of representation and access interfaces, the *adapter layer* where special adapters (harvesters) have to be implemented to provide uniform access and transform the data source specific model into the global model of the federated system, and the *federation layer* which is responsible for global data integration using an on purpose database.

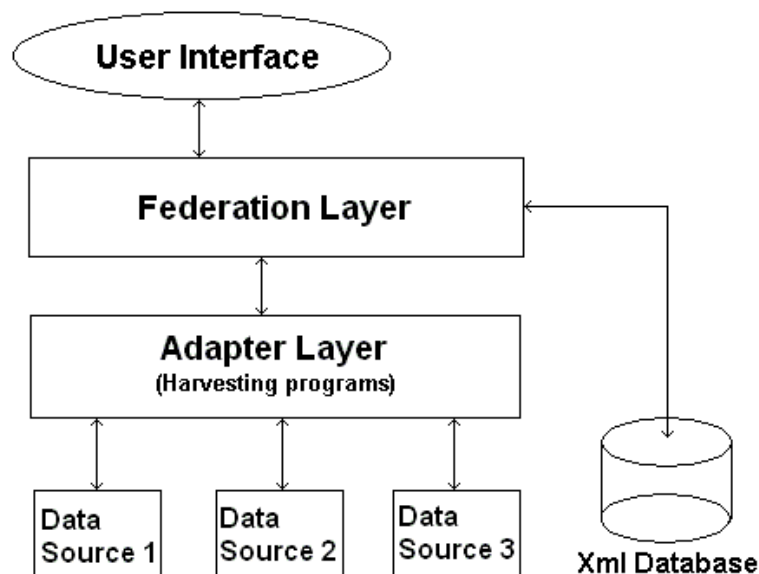


Figure 1

Data Source Layer: these are the archives (digital libraries) whose integration we deal with: Voices on Europe, Virtual Library and Biblio library catalogue.

Adapter Layer: this layer provides uniform access to the information, hiding the differences in the data models and query interfaces. Here the metadata are mapped from the source specific model into the global model of the federated system - the Meta Resource Card derived according to the Dublin Core Elements.

Relevant work has been done in literature as for the role of the Adapter layer. At this stage of the project we are considering the possibility to use the approach defined in the Open Archive Initiative (OAI) (Lagoze & Van den Sompel 2001) (Lynch 2001) where the Data Sources function as a Data Provider adopting OAI technical framework to expose metadata about their content. On the other side Service Providers (for instance the Federation Service) harvest metadata from data providers using the OAI protocol, to provide value-added services. According to this approach the Adapter layer would implement all the Harvester and the OAI protocol.

It is worth highlighting that the OAI approach addresses the interoperability issues requiring that all data providers (Data Source) provide the metadata in a common format, namely the Dublin Core Metadata Element Set (Weibel 1998). This approach has been adopted in successful initiative concerning digital libraries federation (Liu et al. 2001).

Federation Layer: in this layer the services for definition and query of the integrated data vision are provided. Metadata describing information of the three different resources are stored in a unique XML database.

To this extent a common metadata format (Meta Resource Card - MRC) must be devised for the three resources. To effectively address the interoperability issue, the Meta Resource Card should follow the unqualified Dublin Core Standard to define the common fields. This choice is compliant with the Open Archive Initiative intentions.

We are also investigating the possibility to find Federation layer solutions capable to become easily compliant with RDF approach.

On top of the Federation Layer we added the User Interface which will provide information access through the Web to all the users. The use of active pages will allow service personalization, according to the user's role and the actual function exploited as reported in Table 3.

5. Conclusion

This paper reports on the design of a federation service for three heterogeneous digital libraries. The scope of the federation service is to provide a common metadata format for gathering information from the available data sources and to provide a unique querying interface to access them.

At this stage of the project we analysed the state of the art in order to choose the most suitable realisation approach accounting for sound theoretic issues such as Dublin Core Metadata and Open Archive Initiative which are now being investigated in the digital libraries community. Our purpose is also to devise a simple yet easily realisable solution to validate the pilot requirements.

The three data sources analysis is now completed, highlighting the major differences of the three archives.

We therefore choose a federated model with a consequent layered architecture aiming at implementing the OAI protocol and the Dublin Core Metadata description.

We defined a Meta Resource Card, according to the Dublin Core Standard, to unify the description of the federated data to the PEP user.

We are now continuing the realisation of the pilot project whose the first results are expected by autumn 2002.

References

DCMI, Dublin Core Metadata Initiative, OCLC, Dublin Ohio.
<http://dublincore.org/>

DCMIR, Dublin Core Metadata Initiative Recommendationst
<http://dublincore.org/documents/>

DCMES, 1999. Dublin Core Metadata Element Set, Version 1.1: Reference Description
<http://dublincore.org/documents/dces/>

Endig, M, Hoding, M, Saake, G., Sattler, K.U. and Schallehn, E, 2000. Federation services for heterogeneous digital libraries accessing cooperative and non-cooperative sources. *In: International Conference on Digital Libraries: Research and Practice, 2000 Kyoto.* 120 -127.

Hanani, U. and Frank, A.J, 2000. The parallel evolution of search engines and digital libraries: their convergence to the Mega-Portal. *In: International Conference on Digital Libraries: Research and Practice, 2000 Kyoto, 211 -218.*

Lagoze, C. and Van de Sompel, H., 2001. The Open Archives Initiative: Building a low-barrier interoperability framework. *In: the ACM/IEEE Joint Conference on Digital Libraries, Roanoke VA June 24-28 2001, 54-62.*

INNOPAC Official Web site <http://www.iii.com/>

INNOPAC Users Mailing List <http://innopacusers.org/>

Liu, X., Maly, K. Zubair M., and Nelson. M.L. 2001. Arc - An OAI Service Provider for Cross Archiving Searching. *In: the ACM/IEEE Joint Conference on Digital Libraries, Roanoke VA June 24-28 2001, 65-66.*

Liu, X., Maly, K. Zubair M., and Nelson. M.L. 2001. Arc - An OAI Service Provider for Digital Library Federation. *D-Lib Magazine 7(4).*

Lynch, C. 2001. Metadata Harvesting and the Open Archives Initiative. *ARL Monthly Report 217, August 2001, <http://www.arl.org/newsltr/217/mhp.html>.*

OAI The Open Archives Initiative Protocol for Metadata Harvesting. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14.
<http://www.openarchives.org/OAI/openarchivesprotocol.html>

Pirri M. and Noiret S., 2002. WWW VL EUI History Project Report and Analysis for future development in the PORTA EUROPA Portal (PEP) Pilot Project.

Pirri M. and Terzuoli G., 2002. Voices on Europe for PORTA EUROPA Portal (PEP) Pilot Project.

Pirri M. and Baglioni P., 2002. Library Automation System Summary on INNOPAC Manual Description for PORTA EUROPA Portal (PEP) Pilot Project.

Rdf, Resource Description Framework (RDF)
<http://www.w3.org/RDF/>

Weibel S., 1998. The Dublin Core: A simple content description format for electronic resources. *NFAIS Netwletter, 1998. 40(7), 117-119.*

Xml, Extensible Markup Language (XML)
<http://www.w3.org/XML/>