A Distributed Digital Library System Architecture for Archive Metadata

Nicola Ferro and Gianmaria Silvello

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

I. Motivations and Goals

The role of *Digital Library Systems (DLSs)* in collecting, managing and preserving our cultural heritage is increasingly preponderant in several contexts. In particular, DLSs are not merely the digital counterpart of traditional libraries, rather they can be seen as the technology of choice for managing the information resources of different kinds of organizations, which range from libraries, and museums to archives.

We present a solution to address the problem of sharing metadata between different archives spread across a geographic region. In particular, we consider the Italian Veneto Region archives which promote a related project called *Sistema Informativo Archivistico Regionale (SIAR)*. The main goal of the SIAR project is to develop a DLS for sharing archive metadata spread across the territory.

Archive metadata are geographically distributed across the Veneto Region and they are preserved in several local archives; in this context, independent, private or public local archives keep metadata without sharing them and this prevents the offering of common advanced services on metadata. The SIAR objective is to develop a DLS ables to provide advanced services on regional archive metadata, such as a full-text search.

The SIAR design must take into consideration several issues: it has to preserve local archives autonomy and gather their metadata to perform advanced services. Moreover, the SIAR project has to be developed in the Italian Veneto Region context, considering various constraints imposed by the pre-existing regional information systems.

To design the SIAR DLS, we have to satisfy several requirements. On the one hand we have to guarantee the local bodies maintain archives management autonomy. On the other, we have to build up regional coordination so Veneto Region can have an integrated global vision of local archives participating in SIAR.

The added value of the SIAR is that DLS shares metadata exploiting advanced *Digital Library (DL)* technology that can be integrated and adapted with pre-existing systems using different technologies. The result is a scalable, adaptive, flexible and widely-adoptable architecture to share information in a distributed environment.

Section II explains SIAR design choices. Section III discusses two works related to the SIAR project. Section IV presents the SIAR DLS architecture. Section V describes SIAR prototype development details and section VI concludes and discusses the future developments of SIAR.

II. SIAR Project Coordinates

SIAR is constituted by a federation of autonomous archives characterized by a unique access point represented by the Veneto Region portal. It supplies advanced research and integration services providing a common starting point; these services could also be implemented at a later date.

A. Archival Descriptive Metadata

In archival context, data exchanged between local archives are often represented by metadata. Indeed, most of the archival resources are not available in digital form, but they are described and represented by metadata. Sometimes archival resources are metadata themselves. So considering the SIAR project focus, we have to share metadata between archives spread across the territory. In order to perform advanced services on metadata, the Veneto Region has to collect metadata from local archives.

The structure and the requirements of archives have to be taken into account in order to adopt and use metadata in the SIAR project. Indeed, in the archival context metadata are often used to represent archival descriptions. An archival description is defined in [14] as "the process analyzing, organizing, and recording details about the formal elements of a record or collection of records[1], to facilitate the work's identification, management, and understanding"; archival descriptions have to reflect the peculiarities of the archive, retain all the informative power of a record, and keep trace of the provenance and original order in which resources have been collected and filed by archival institutions [7]. Thus, the metadata formats used and exchanged in the SIAR, have to respect all the archival informative requirements. Metadata have to retain information about the context of a given record, have to reflect the archive organization. Archives are described in a multi-leveled fashion and the descriptions assume a hierarchical structure. Archival descriptions are define by an international standard called *International Standard for Archival Description (General) (ISAD(G))* [9]; ISAD(G) stands for general standard for archival description and it is defined by the *International*

[1] A record is defined in [13] as: "Any document made or received and set aside in the course of a practical activity"

Council on Archives (ICA)[2]. ISAD(G) defines that archival descriptions have to proceed from the general to the specific unit; archival descriptions assume a tree structure expressing the hierarchical structure of the archive.

Context and hierarchy are two fundamental requirements that must be realized by archival descriptive metadata; indeed, they permit to express the whole informative power of archival descriptions. Thanks to context and hierarchy, provenance is retained and relationships between archival descriptions are preserved as well as the relationships with the production and preservation environment. In a distributed environment another requisite must be taken into account; we define this requirement as variable granularity. Archival descriptive metadata has to facilitate the access to a particular archival item, with the desired degree of detail and without accessing the whole archive hierarchy.

In the SIAR context we have to design an effectual system that permits to manage and exchange archival descriptive metadata retaining context and hierarchy and at the same time enabling variable granularity requirement.

The only standard defined for archival descriptive metadata is the *Encoded Archival Description (EAD)* metadata format[3]. EAD reflects the archival structure and holds relationships between entities in an archive [15]. EAD respects context and hierarchy requirements but it has been underlined [10], [16] that EAD metadata standard is not well-suited to be used in distributed environment; indeed, EAD files are usually heavy and hard-to-move. Moreover, EAD does not enable variable granularity requirement; EAD retains in a unique file all the descriptions of an archive and it is not possible to access a specific archival unit without accessing the whole archival hierarchy. In the context of SIAR the use of EAD is problematic because there is the necessity of flexibility and lightness that EAD cannot provide.

B. Authority Control

Authority control enables catalogers to disambiguate items with similar or identical headings and it is used by catalogers to collocate materials that logically belong together. Authority control is realized defining the authority files; authority file enables the unique identification of the particular entity it describes. Archive resources describe different kinds of realities, such as a person, a private organization or a public institution.

In a distributed environment like SIAR, we must guarantee the definition uniqueness of the entities described by archival resources; so it is very important to share common authority lists between the archives participating in SIAR. The Veneto Region has to define the authority lists that can be formed as metadata files and it also has to share them with local archives; they must use Veneto Region authority lists to define their entities in a well-formed and unique way.

C. SIAR Design Choices

The SIAR objectives are related to the current development guidelines of DLSs. In fact DLSs are service-oriented and can be composed by independent sub-systems that cooperate together to supply the required Digital Library functionalities. Moreover, DLSs aim to strengthen integration and interoperability between different systems.

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is our project choice to address the metadata exchange issue. OAI-PMH is based on the distinction between Data Provider and Service Provider which, respectively, offer metadata and harvest metadata to provide services [17]. Data Providers are the components that make metadata available to the Service Providers that harvest metadata. Each Data Provider manages its own metadata and it is independent and autonomous by the outside information systems. Service Provider role is to harvests metadata by the different Data Provider and to performs advanced services on these harvested metadata.

In our case study, the Veneto Region is the Service Provider which gives advanced services such as data and public access to metadata. Archive keepers act as Data Providers because they supply archive metadata. These metadata will be harvested, stored and indexed by the Veneto Region to provide services. On the other hand, in considering authority list metadata, the Veneto Region acts as a Data Provider and local archives as Service Providers that harvest authority lists metadata.

Dublin Core (DC)[4] metadata format is tiny, easy-to-move, shareable and remarkably suitable for a distributed environment. Thanks to these characteristics it is required as the lowest common denominator in OAI-PMH. Thus, DC metadata are very useful in information sharing; for this reason, it is used in several contexts ranging from Web to digital libraries.

We propose to use the DC metadata format in conjunction with the OAI-PMH protocol in SIAR, addressing archive requirements. Even if metadata are the foremost SIAR resources, we also have to contemplate that some archives start to keep archival resources in digital form. Therefore in SIAR we also have to look for a way to share digital resources between distributed archives.

III. Related Works

A. The European Library (TEL)

An interesting experience of different metadata-based systems integration in the Digital Library field is The European Library [5]. The European Library is a free service that offers access to the resources of the 47 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.

A national library can join The European Library in different ways depending on the information system it uses. The The European Library approach is to establish a minimum requirement a national library has to achieve to join the project. In this way every national library maintains its resource management autonomy and it has not to modify its information system [18].

[2] http://www.ica.org/
[3] http://www.loc.gov/ead/
[4] http://www.dublincore.org/
[5] http://www.theeuropeanlibrary.org/

The European Library project offers a concrete integration possibility based on OAI-PMH, used to collect the catalogue records of national libraries. In this context, The European Library acts as Service Provider and the national libraries as Data Providers; information and harvested metadata are stored in a central index that offers advanced search services. Moreover this central index could be the starting point for other services, such as multilingual information access [1].

Furthermore, the TELplus project[6] is another building block towards European Digital Library and is aimed to strengthen, extend and improve The European Library service. In particular, it aims to improve and enhance the adoption of OAI-PMH as a means of integration.

B. DRIVER: Networking European Scientific Repositories

DRIVER[7] 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].

The DRIVER infrastructure is composed of a number of core resources that are exploited for making specific DL applications available to user communities. It implements a virtually centralized and homogeneous access point to heterogeneous and distributed content collected from existing repositories. DRIVER emphasizes the implementation of nominal, globally accepted standards in a real-life system, with a focus on metadata exchange, in particular by the use of OAI-PMH. One of the Digital Library application components provided by DRIVER is an OAI-Publisher Service that implements the OAI-PMH and it makes DRIVER an open archive Data Provider. In DRIVER multiple OAI repositories are aggregated to construct Search and OAI-Publisher Service; in this way DRIVER services operate upon the aggregated content of existing institutional OAI repositories [4].

IV. SIAR Development: The Nature of the Digital Library System

SIAR is a DLS developed as a three-layer architecture, composed of the metadata transport layer, the metadata management layer and the presentation layer.

In Figure 1 we can see the SIAR architecture layers. The transport layer called SIRV-PMH represents the SIAR transport infrastructure composed of the union of OAI-PMH and the regional information system called SIRV-INTEROP. At this stage there is the SIAR distributed core. Archive metadata have to be harvested by the Veneto Region without undermine the autonomy of local archives; thanks to the OAI-PMH protocol adoption, SIAR has a way of both transporting and sharing metadata in a distributed environment.

At the second level we find the management layer called SIAR-MM (SIAR - Metadata Management Layer). SIAR-MM layer is composed of an Application Logic part and a Data Logic part. By the use of Application Logic we

[6]http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/

[7]http://www.driver-repository.eu/

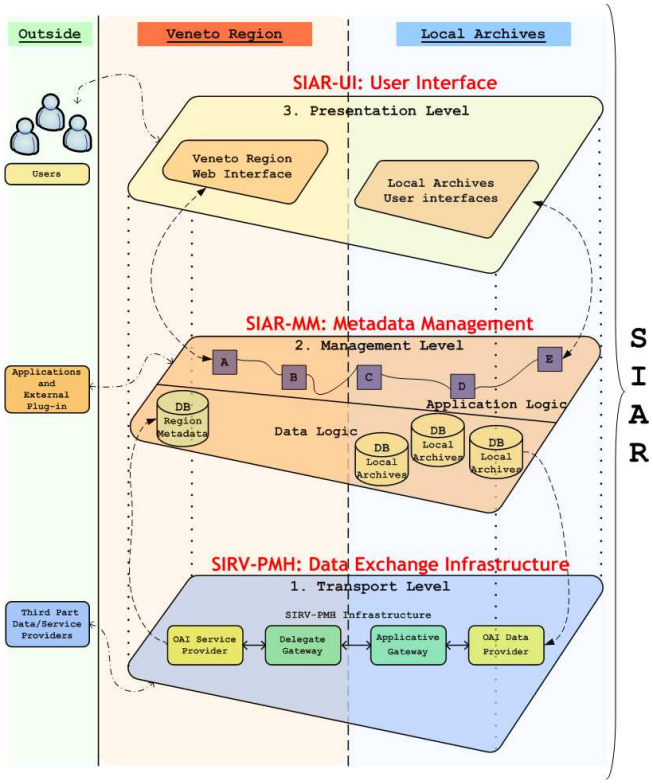

Fig. 1. SIAR Architecture Layers

can develop advanced services both on harvested metadata owned by the Veneto Region and on local archive metadata. The applications developed for SIAR can be used on Veneto Region metadata index and on local archives metadata too; indeed they are independent of the transport infrastructure. Thanks to this organization, adding a third-part service to the SIAR will be almost effortless. The Application Logic works on the metadata managed by the Data Logic composed of a central database owned by the Veneto Region and a set of distributed local databases owned by the local archives. SIAR-MM Data Logic preserves and manages the physical data of the SIAR project; so this sub-layer manages local archive data and Veneto Region archive data as well.

At the third level we have the presentation layer called SIAR-UI (SIAR - User Interface) constituted by the user interfaces. SIAR 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 SIAR; the second is dedicated to specialized users who through this interface can add, remove or update archive metadata.

V. SIAR Metadata Transport Layer

SIRV-PMH is the first layer of the three-layers SIAR architecture and it provides the transport layer. SIRV-PMH is the infrastructure which permits metadata exchange between local archives spread across the territory and the central archive constituted by the Veneto Region. This infrastructure is the

outcome of the fusion between OAI-PMH and the Veneto Region information system SIRV-INTEROP.

A. SIRV-INTEROP: Veneto Region Information System

SIRV-INTEROP is the system that has to guarantee interoperability between different applications adopted by the Veneto Region. SIRV-INTEROP project implements a Domain Gateway System based on Applicatory Cooperation principles [6]. 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.

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 it provides various services to external users. The main goal of the domain gateway system is to integrate the services of different administrations. A highly important issue for this system is to maintain the independence of each cooperative system. In this way any system that wanted to fulfill a service to the National Net community could maintain its internal structure unchanged.

The main goal of the domain gateway system is to integrate the service of different administrations.

Domain gateways are divided into two main classes: *Applicative Gateway* and *Delegate Gateway*. Applicative Gateway provides services and every domain which can distribute services carries out this function through this gateway; Delegate Gateway requests services from Application Gateways.

Communication and data exchanges between Applicative and Delegate Gateways take place by means of *Simple Object Access Protocol (SOAP)*, which use *eXtensible Markup Language (XML)* technologies to define an extensible messaging framework [8].

From a logical point of view, every domain gateway is composed of two main components: *Cooperation* (performs data communications generic functions) and *Integration* (performs the adaptation towards the 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 systems they have to integrate.

Through the 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 technologies.

B. SIRV-PMH: Data Exchange Infrastructure

The purpose of SIRV-PMH is to integrate OAI-PMH and SIRV-INTEROP without any substantial modification of their functioning; to do this OAI-PMH has to be used as a layer over SIRV-INTEROP that in turn is built-on SOAP [5].

The biggest issue for integrating these systems is to transport OAI-PMH requests over the 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 to extract OAI-PMH requests and responses from their SOAP envelope [2].

We can see in Figure 2 how this system works.

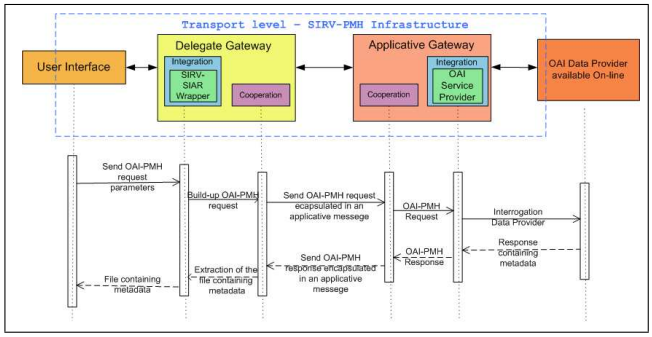

Fig. 2. SIRV-PMH Functioning

Moreover, we analyze 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 a *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 an OAI-PMH response. When the response is ready, it is 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.

C. SIRV-PMH Prototype Development

The whole SIAR project lays on the SIRV-PMH layer, so we developed a prototype of the transport infrastructure that will be integrated with the other layers of SIAR.

In Figure 3 we can see the SIAR Prototipe Interface. The user inserts the parameters needed by the Service Provider to constitute the OAI-PMH request by means of this graphical user interface. This information is: the metadata format, the *Uniform Resource Locator (URL)* of the Data Provider available on the Web that we want to interrogate and the XML file path where harvested metadata will be stored. Moreover the user inserts also SIRV-INTEROP parameters like the name of the Applicative Gateway to call and the applicative service request.

The OAI-PMH request parameters are sent to the Delegate Gateway which encapsulates the request parameters inside an

Fig. 3. SIAR Prototype Interface

applicative message that can travel inside SIRV-INTEROP. The Applicative Gateway receives the message, extracts the OAI-PMH parameters and calls a Service Provider that constitutes the OAI-PMH request and calls the required Data Provider.

The Data Provider answer follows the inverse request iter when it is returned to the Delegate Gateway. In this prototype the Delegate Gateway saves the harvested metadata in an XML file that is sent to the user.

In Figure 4 we can see the output response of a test performed by the SIAR prototype.

In the SIAR architecture the Data Provider available on-line will be mapped in a local archive Data Provider. Every local archive will have a Data Provider linked to a specific Applicative Gateway. The Veneto Region will query the local archives through a Service Provider linked to a Delegate Gateway. By means of a specific application (SIAR-MM) a user could research directly through the metadata harvested by Veneto Region.

These metadata will also be made available to third party Service Providers that would like to harvest SIAR archive metadata.

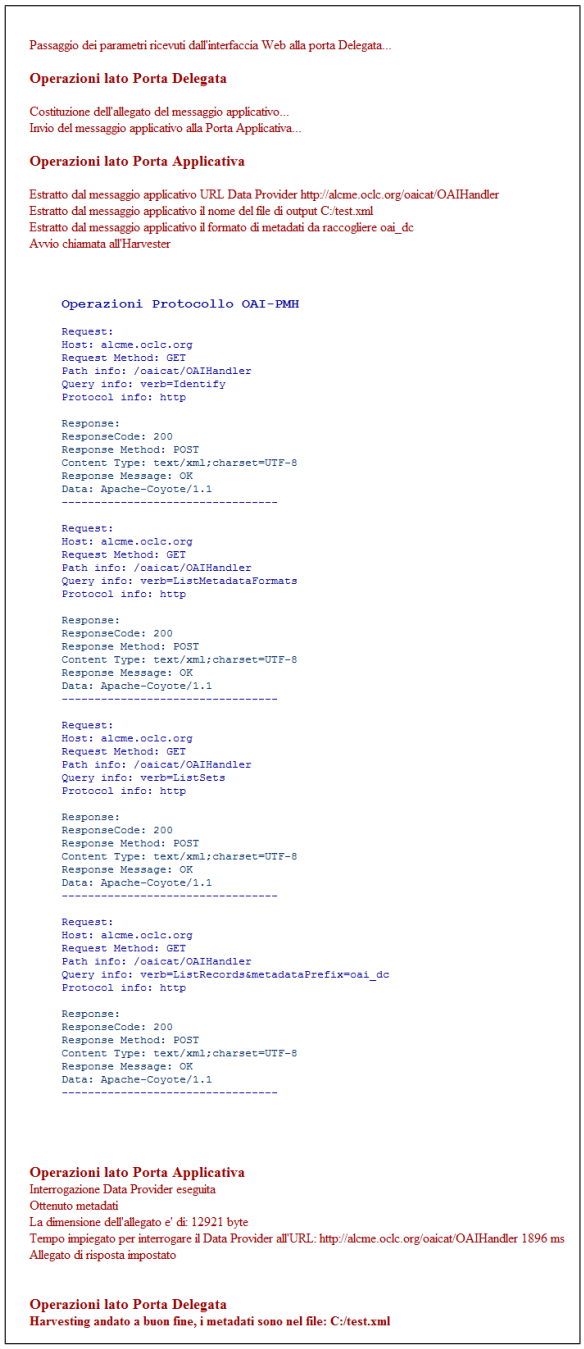

Fig. 4. Prototype Output Response

VI. Conclusions and Future Work

We presented the SIAR project, a DLS that permits the sharing metadata between several archive in a flexible and scalable way. The proposed architecture preserves local archives autonomy and at the same time provides advanced services on archive metadata.

Future work will concern the continuation of the development of the SIAR-MM layer and the SIAR-UI layer.

SIAR-MM application logic will provide a search service over the metadata that searches and retrieves metadata starting with a user query written in natural language.

A future development will be the direct access to digital resources, if available. This feature involves the sharing of complex digital objects [11] for which the *Open Archives Initiative - Object Reuse and Exchange (OAI-ORE)* protocol [12] would be an interesting possibility to take into consideration.

Acknowledgements

The work reported has been 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) and 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] M. Agosti, M. Braschler, N. Ferro, C. Peters, and S. Siebinga. Roadmap for MultiLingual Information Access in the European Library. In N. Fuhr, L. Kovacs, and C. Meghini, editors, *Proc. 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)*, pages 136–147. Lecture Notes in Computer Science (LNCS) 4675, Springer, Heidelberg, Germany, 2007.

[2] M. Agosti, N. Ferro, and G. Silvello. An Architecture for Sharing Metadata among Geographically Distributed Archives. In C. Thanos, F. Borri, and L. Candela, editors, *DELOS Conference*, volume 4877 of *Lecture Notes in Computer Science*, pages 56–65. Springer, Heidelberg, Germany, 2007.

[3] L. Candela, D. Castelli, P. Manghi, and P. Pagano. Enabling Services in Knowledge Infrastructures: The DRIVER Experience. In M. Agosti, F. Esposito, and C. Thanos, editors, *Post-proceedings of the Third Italian Research Conference on Digital Library Systems (IRCDL 2007)*, pages 71–77. ISTI-CNR at Gruppo ALI, Pisa, Italy, November 2007.

[4] L. Candela, D. Castelli, P. Manghi, and P. Pagano. OAI-Aggregator Services for Enhanced Cross-Repository Services. In M. Agosti, F. Esposito, and C. Thanos, editors, *Post-proceedings of the Third Italian Research Conference on Digital Library Systems (IRCDL 2007)*, pages 61–66. ISTI-CNR at Gruppo ALI, Pisa, Italy, November 2007.

[5] S. Congia, M. Gaylord, B. Merchant, and H. Suleman. Applying SOAP to OAI-PMH. In R. Heery and L. Lyon, editors, *Proc. 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2004)*, volume 3232 of *Lecture Notes in Computer Science*, pages 411–420. Springer, Heidelberg, Germany, 2004.

[6] Gazzetta Ufficiale N.78 del 3 Aprile 2002. ALLEGATO n. 2: Rete Nazionale: caratteristiche e principi di cooperazione applicativa.

[7] A. J. Gilliland-Swetland. *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment*. Council on Library and Information Resources, 2000.

[8] 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, W3C, pp. 40, 2003.

[9] International Council on Archives. ISAD(G): General International Standard Archival Description, 2nd edition. Ottawa: International Council on Archives, 1999.

[10] K. Kiesling. Metadata, Metadata, Everywhere - But Where Is the Hook? *OCLC Systems & Services*, 17(2):84–88, 2001.

[11] C. Lagoze, S. Payette, E. Shin, and C. Wilper. An architecture for complex objects and their relationships. *International Journal on Digital Library*, 6(2):124–238, 2005.

[12] C. A. Lynch, S. Parastatidis, N. Jacobs, H. Van de Sompel, and C. Lagoze. The OAI-ORE effort: progress, challenges, synergies. In E. M. Rasmussen, R. R. Larson, E. Toms, and S. Sugimoto, editors, *Proc. 7th ACM/IEEE Joint Conference on Digital Libraries, (JCDL 2007)*, page 80. ACM Press, New York, USA, 2007.

[13] H. MacNeil, C. Wei, L. Duranti, A. Gilliland-Swetland, M. Guercio, Y. Hackett, B. Hamidzadeh, L. Iacovino, B. Lee, S. McKemmish, J. Roeder, S. Ross, W. Wan, and Z. Zhon Xiu. *Authenticity Task Force Report*. InterPARES Project: Vancouver, Canada, 2001.

[14] R. Pearce-Moses. *Glossary of Archival And Records Terminology.* Society of American Archivists, 2005.

[15] D. V. Pitti. Encoded Archival Description. An Introduction and Overview. *D-Lib Magazine*, 5(11), 1999.

[16] C. J. Prom. Does EAD Play Well with Other Metadata Standards? Searching and Retrieving EAD Using the OAI Protocols. *Journal of Archival Organization*, 1(3):51–72, 2002.

[17] 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.

[18] T. van Veen and B. Oldroyd. Search and Retrieval in The European Library. A New Approach. *D-Lib Magazine*, 10(2), February 2004.