BUILDING LARGE HETEROGENEOUS INTERCONNECTED DIGITAL LIBRARY INFRASTRUCTURES: THE INTEROPERABILITY CHALLENGE

C. Thanos, D. Castelli, L. Candela *

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche, 56124 Pisa, Italy - (thanos, castelli, candela)@isti.cnr.it

KEY WORDS: Interoperability, Digital Libraries, Standard, Mediators, Interoperability issues, interoperability model

ABSTRACT:

Interoperability is a multi-layered and context-specific concept, crucial when building heterogeneous interconnected Digital Library infrastructures. It encompasses different levels along a multidimensional spectrum. Two key problems hinder the achievement of interoperability: (i) the heterogeneity of the exchanged information which covers all types of syntactic, structural, and semantic diversities among systems used to modelling information and (ii) the inconsistency between the use of the information as intended by its originator and the intended exploitation of it by the recipient. Building heterogeneous interconnected Digital Library infrastructures requires addressing all dimensions of interoperability including content, user, functionality, architecture, quality, and policy. In this paper a comprehensive study of the interoperability issues involved when building interoperable Digital Libraries is performed.

1. INTRODUCTION

As technology becomes more far-reaching and interconnected, interoperability has become critical. It ranges along a wide multidimensional spectrum: at one end of the spectrum we have data and metadata interoperability while at the other end we distinguish organizational, legal and policy interoperability.
Interoperability is a multi-layered and context-specific concept, crucial when building heterogeneous interconnected Digital Library infrastructures. It encompasses different levels along a multidimensional spectrum. Two key problems hinder the achievement of interoperability: *(i)* the heterogeneity of the exchanged information which covers all types of syntactic, structural, and semantic diversities among systems used to modeling information and *(ii)* the inconsistency between the use of the information as intended by its originator and the intended exploitation of it by the recipient.
Building heterogeneous interconnected Digital Library infrastructures requires addressing all dimensions of interoperability including content, user, functionality, architecture, quality, and policy.
This paper is organized as follows. Section 2 describes a number of application scenarios where several interoperability issues are outlined. Section 3 gives a definition of interoperability and identifies the sources of problems which hinder the interoperability among federated Digital Libraries. Section 4 describes some approaches to achieving interoperability. Section 5 identifies and describes several types of interoperability. The last section contains some concluding remarks regarding best practices when implementing interoperability solutions.

2. INTEROPERABILITY SCENARIOS IN THE CONTEXT OF INTECONNECTED INTEROPERABLE DIGITAL LIBRARIS

In this section we illustrate some typical scenarios which occur in the context of interconnected interoperable digital libraries and emphasize interoperability problems.

2.1 Metadata Harvesting

In the open, networked environment that encompasses multiple user communities using a multitude of standards for description of digital resources, the need for interoperability among metadata is paramount. To enable federated searches and to facilitate metadata management, much effort has been devoted to achieving or improving interoperability among metadata schemes/records. The metadata harvesting scenario is the most commonly implemented.
In this scenario, information providers (institutional repositories/digital libraries) make metadata about their collections available for harvesting. Service providers (harvesters) use this metadata to create value added services.
If the service provider wants to collect metadata from different information providers and integrate them in order to provide a uniform access to them different interoperability problems could arise. For example, if the different information providers adopt different metadata models their integration could become problematic. Even if a common metadata model is adopted some other interoperability problems can still arise concerning, for example, the syntax rules for how the elements and their content should be encoded, content rules for how content must be formulated (for example, how to identify the main title), representation rules for content (for example, capitalization rules), and allowable content values (for example, terms must be used from a specified controlled vocabulary).
Additional interoperability problems can arise related to mismatching of transfer protocols.

* Corresponding author.

To facilitate interoperability, information providers are required to supply metadata which complies to a common schema (for example, the unqualified Dublin Core Metadata Element Set). The harvesters must issue requests according to a protocol (for example, OAI-PMH (Lagoze & Van de Sompel, 2001)).
The metadata harvesting approach is adopted by Europeana, the European Digital Library.
From the service provider side, Europeana must interoperate with the memory institutions in order to obtain the metadata on which it offers its services. In order to achieve this goal, Europeana must be able to interact with the providers' DLs at a functional level, to obtain the metadata required. This is obtained by adopting a standard solution, namely the OAI-PMH protocol for harvesting. The protocol defines the services to be implemented on both sides, so that interoperability between Europeana and the content producers can be achieved. Once the data is acquired from Europeana, it has to be mapped from the original format to the Europeana Data Model (EDM). This mapping requires knowledge of the semantics of the source and target data models.
From the information provider side, Europeana will make its contents available through a number of APIs, each one addressing the needs of a particular category. These APIs will be used by consumers to obtain services from Europeana, and will be the result of negotiations between the involved parties.

2.2 Digital Library – User Interaction

Each Digital Library has its own policies; for example, policies for document acquisition, document management, access and use, loans, charges for access, etc.
On the other hand, the different classes of patrons (end users, librarians, administrators, etc.) interact with the Digital Library according to a profile authenticated at the registration time. In order for the patrons to effectively perform their tasks these must be consistent with the policies of the Digital Library.
The Libraries' policies and the working requirements of their patrons must be compatible.
Adequate representation formalisms for representing DL policies as well as patron's requirements must be adopted and appropriate consistency checking techniques must be implemented.
In addition, in the context of interconnected Digital Libraries, there is the need for the user credentials and profiles to be propagated across several Digital Libraries in order to allow a user authenticated in one Digital Library to operate in another Library which trusts the user Library. There is the need for the interconnected and mutually trusted Digital Libraries to be able to exchange user credentials and profiles in a meaningful manner.

2.3 User – User Interaction

Another very important form of interaction between users regards the so-called "user collaboration" based on Digital Libraries' resources (content and functions/services). In order to support collaboration between users across different mutually trusted Digital Libraries these must enable their users to interact among them directly or indirectly. To achieve this the users must be able to exchange meaningful content and to invoke and/ or combine compatible functions/services across different Digital Libraries. This means the ability to perform consistency checking between the invoked and/or combined functions/services.

2.4 Trusted Digital Libraries

In the context of interconnected Digital Libraries it is of paramount importance the harmonization of their policies concerning the very different functionalities/services supported by them in order to guarantee a seamless interoperation between their patrons. Therefore, formalisms to express Digital Library policies and techniques to check their compatibility and enforce them are very important in this scenario.

3. THE INTEROPERABILITY PROBLEM

From the IEEE definition which characterises it as "*the ability of two or more systems or components to exchange information and to use the information that has been exchanged*" it follows that in order to achieve interoperability between two entities (producer, consumer) two conditions must be satisfied: (*i*) the two entities must be able to exchange "meaningful" information objects; (*ii*) the consumer entity must be able to use the exchanged information in order to perform a set of tasks that depend on the utilization of this information.
Therefore, there are two sources of problems which hinder the interoperability between the producer and consumer entities: (*i*) Heterogeneity of the exchanged information objects; (*ii*) Inconsistency between the use of the information object as intended by the producer entity and the intended exploitation of this object by the consumer entity.

3.1 Heterogeneity of exchanged information objects

There are three types of heterogeneity to be overcome in order to achieve a meaningful exchange of information objects (Candela, et al., 2008).
First, heterogeneity between the native data / query language (of the consumer entity) and the target data / query language (of the producer entity). When this heterogeneity is resolved we say that syntactic interoperability between the two entities has been achieved.
Second, heterogeneity between the models adopted by the producer and the consumer entities for representing information objects. When this heterogeneity is resolved we say that structural interoperability between the two entities has been achieved.
Third, heterogeneity between the "semantic universe of discourse" of the producer and consumer entities (differences in granularity, differences in scope, temporal differences, synonyms, homonyms, etc.). When this heterogeneity is resolved we say that semantic interoperability between the two entities has been achieved.
Although these three levels of interoperability, syntactic, structural, and semantic allow a meaningful exchange of information objects, they do not guarantee that "real" interoperability between the two entities has been achieved as this implies the ability of the consumer entity to use the exchanged information objects in order to perform a set of tasks.
When the sole "meaningful" exchange of information objects is sufficient to enable the consumer entity to perform a set of tasks based on the exchanged information objects we can talk about "basic" interoperability between producer and consumer entities.

3.2 Mismatching between producer information resources and consumer needs

We distinguish two kinds of interoperability: partial and full interoperability.
By "partial interoperability" between two entities we mean that the consumer entity is able to perform a limited number of tasks based on the exchanged information.
By "full interoperability" or "operational interoperability" between two entities we mean that the consumer entity is able to perform a full range of tasks based on the exchanged information.
Possible causes for inability to achieve operational interoperability between producer and consumer entities are:

- Quality mismatching;
- Data-incomplete mismatching.

Quality mismatching occurs when the quality profile associated with the exported information object does not meet the quality expectations of the consumer entity.
Data-incomplete mismatching occurs when the exported information object is lacking some useful information to enable the consumer entity to fully exploit the received information object.
In general, the meaningful exchange of information objects and therefore syntactic, structural and semantic interoperability is a necessary but not sufficient condition for achieving operational interoperability.
In fact, to achieve full operational interoperability between two entities the exchanged information must be complemented with some "descriptive" information, such as contextual, provenance/lineage, quality, security, privacy, etc. information which gives additional meaning. The descriptive information should be modelled by purpose-oriented descriptive data models /metadata models.
The use of purpose-oriented descriptive data models is of paramount importance to achieve operational interoperability.
The type of descriptive information to be provided by the producer entity depends very much on the requirements imposed by the consumer entity's tasks. For example, if the consumer entity wants to perform a data analysis task on the imported information then quality information is of paramount importance; without such information the task of data analysis cannot be performed.
Consequently, if the producer entity of an information object is willing to export/publish it, its possible uses by the potential consumer entities must be carefully taken into account and it must be endowed with appropriate descriptive information. Appropriate purpose-oriented information models /metadata models to describe the descriptive information must be chosen and used. If a multi-use of an information object is expected it could be necessary to associate different descriptive models/metadata models with it.

4. APPROACHES TO ACHIEVING INTEROPERABILITY PROBLEM

The main concept enabling the "meaningful" exchange of information objects is mediation. This concept has been used to cope with many dimensions of heterogeneity spanning data language syntaxes, information object models and semantics. The mediation concept is implemented by a mediator, which is a software device that supports a mediation schema capturing user requirements, and an intermediation function between this schema and the distributed information sources.
A key feature which characterizes a mediation process is the kind of intermediation function implemented by a mediator. There are four main functions:

- Mapping
- Matching
- Integration
- Consistency Checking.

Mapping refers to how information structures, properties, relationships are mapped from one representation scheme to another one, equivalent from the semantic point of view.
Matching refers to the action of verifying whether two strings/patterns match, or whether semantically heterogeneous data match.
Integration refers to the action of combining data residing at different sources, and providing the consumer entity with a unified view of these data.
Consistency checking refers to the action of checking the compatibility of logical relationships between functional / policy / organizational descriptions of two entities.
We can distinguish four main mediation scenarios:
Mediation of data structures: this permits data to be exchanged according to syntactic, structural and semantic matching. The functions of mapping, matching and integration are mainly adopted to implement this kind of mediation.
Mediation of functionalities: this make possible to overcome mismatching of functional descriptions of two entities that are expressed in terms of pre- and post-conditions. The functions of mapping, matching and consistency checking are mainly adopted to implement this kind of mediation.
Mediation of policies/business logics: this employs techniques to solve policy, business mismatches. The functions of mapping, matching and consistency checking are mainly adopted to implement this kind of mediation.
Mediation of protocols: this make possible to overcome behavioural mismatches among protocols run by interacting parties.
Automated mediation basically focuses on matching the producer entity information resources to the consumer entity needs. It relies on:

- Adequate modelling of structural, formatting, and encoding constraints of the producer entity information resources.
- Adequate modelling of the consumer entity needs
  - Domain-specific Ontologies;
  - Formally defined transfer and message exchange protocols.
- The definition of a matching relationship between the producer information resources and the consumer models.

There are several approaches to implementing the mediation process; of particular relevance are approaches based on standard data modelling formalisms, formal transfer protocols, and ontologies.

4.1 Ontology-based Mediation Approaches

Ontologies were initially developed by the Artificial Intelligence community to facilitate knowledge sharing and reuse. An ontology consists of a set of concepts, axioms, and relationships that describes a domain of interest.

Ontologies have been extensively used to support all the mediation functions, i.e. mapping, matching, integration, and consistency checking because they provide an explicit and machine-understandable conceptualization of a domain.
They have been used in one of the three following ways.
In the single ontology approach, all source schemas (producers entities schemas) are directly related to a shared global ontology that provides a uniform interface to the consumer entity.
In the multiple ontology approach, each source schema (producer entity' schema) is described by its own (local) ontology separately. Instead of using a common ontology, local ontologies are mapped to each other.
In the hybrid ontology approach, a combination of the two above approaches is used.
Ontologies provide a framework within which the semantic matching/mapping/integration/consistency checking processes can be carried out by identifying and purging semantic divergences. Semantic divergences occur where the semantic relationship between the ontology and the representation is not direct and straightforward

4.2 The Role of Standards in the Mediation Process

The role of formal models and standards in the development of mediation technologies is of paramount importance. In fact, automatic mediation requires the existence of formal models and standards for representing information objects, behaviour, functionality, policy, and protocols.

4.2.1 Information Modelling: Modelling represents basic technologies for modelling, organizing and exchanging information objects.
An information object is composed of a digital Data Object and the Descriptive Information which allows for the full interpretation of the data into meaningful information (see OAIS Reference Model (CCSDS, 2002)).
Several formal information models and languages have been defined and developed for representing, organizing and exchanging information objects (for example, RDF, XML, etc.).
Several discipline-specific standard models have been proposed and developed for representing discipline-specific descriptive information (discipline-specific metadata models) which greatly support the mediation process.
Logic-based and ontology-based models have been defined for representing behaviour, functionality, and policy (for example, OWL-S).
An important role in the mediation process is played by ontologies. Several domain-specific ontologies are being developed (e.g., CIDOC).

5. TYPES OF INTEROPERABILITY

Sometimes operational interoperability between two entities can be characterized by the type of actions the consumer entity is enabled to perform on the exchanged information. For example, if the consumer entity applies a preservation action on the exchanged information we characterize this type of interoperability as "temporal interoperability", as it guarantees access to the exchanged information over time.
Another type of operational interoperability occurs when the consumer entity is obliged to observe security, integrity, confidentiality / privacy, etc. constraints when performing tasks on the exchanged information object. In this case we characterize this type of interoperability as "secure interoperability".
When the consumer entity (usually a middleware component) is searching for a producer entity (another middleware component) that provides a given or complementary functionality, we characterize this type of interoperability as "functional interoperability".
When, in order to perform some tasks the consumer entity (usually an organization) needs to check the consistency of its behaviour / policies / business logics with those of a producer entity (another organization), then we characterize these types of interoperability as "behavioural interoperability", "organizational interoperability" and "business interoperability" respectively.
Therefore, temporal, secure, behavioural, functional, organizational, etc. are specializations of the operational interoperability.

6. CONCLUDING REMARKS

Achieving a full interoperability among Digital Libraries requires that all the Digital Library resources, i.e., content, user, functionality, and policy are interoperable. This means that different types of interoperability must be achieved, i.e. content interoperability, functional interoperability, policy interoperability, etc. and all of them concur to the achievement of a full Digital Library interoperability. This study has been conducted within DL.org, an EU 7th FP project (www.dlorg.eu).
We have emphasized that the use of purpose-oriented descriptive data models is of paramount importance to achieve operational interoperability.
Consequently, if the producer entity of an information object is willing to export / publish it, its possible uses by the potential consumer entities must be carefully taken into account and it must be endowed with appropriate descriptive information. Appropriate purpose-oriented information models / metadata models to describe the descriptive information must be chosen and used. If a multi-use of an information object is expected it could be necessary to associate different descriptive models/metadata models with it.

REFERENCES

Arms, W.Y., 2000. *Digital Libraries*. MIT Press, Cambridge, Massachusetts.

Borgman, C.L., 2000. *From Gutenberg to the global information infrastructure: access to information in the networked world*. MIT Press, Cambridge, Massachusetts.

Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., Schuldt, H., 2008. *The DELOS Digital Library Reference Model - Foundations for Digital Libraries*. DELOS: a Network of Excellence on Digital Libraries.

Candela, L., Castelli, D., Thanos, C., 2010. Making Digital Library Content Interoperable. In: *Post-proceedings of the 6th Italian Research Conference on Digital Libraries*, 91 (to appear).

Consultative Committee for Space Data Systems, 2002. Reference Model for an Open Archival Information System.

Cruz, I.F., Xiao, H., 2005. The role of ontologies in data integration. *Journal of Engineering Intelligent Systems*, 13(4), pp. 245–252.

European Commission, 2003. European Interoperability Framework. http://ec.europa.eu/idabc/en/document/3473 (accessed 23 August 2010).

Fox, E., Marchionini, G., 1998. Toward a Worldwide Digital Library. *Communications of the ACM*, 38(4), pp. 23–28.

Gill, T., Miller, P., 2002. Re-inventing the Wheel? Standards, Interoperability and Digital Cultural Content. *D-Lib Magazine*, 8(1).

Heiler, S., 1995. Semantic interoperability. *ACM Computing Survey*, 27(2), pp. 271–273.

Lagoze, C., Van de Sompel, H., 2001. The open archives initiative: building a low-barrier interoperability framework. In: *Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM Press, pp. 54–62.

Lagoze, C., Van de Sompel, H., 2008. Open Archives Initiative Object Reuse and Exchange User Guide – Primer. http://www.openarchives.org/ore/1.0/primer (accessed 23 August 2010).

Lenzerini, M., 2002. Data Integration: A Theoretical Perspective. In: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM, pp. 233–246.

Paepcke, A., Chang, C.-C. K., Winograd, T., García-Molina, H., 1998. Interoperability for Digital Libraries Worldwide. *Communications of the ACM*, 41(4), pp. 33–42.

Park, J., Ram, S., 2004. Information Systems Interoperability: What Lies Beneath? *ACM Transactions on Information Systems*, 22, pp. 595–632.

Rahm, E., Bernstein, P. A., 2001. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4), pp. 334–350.

Spalazzese, R., Inverardi, P., Issarny, V., 2009. Towards a Formalization of Mediating Connectors for on the Fly Interoperability. In: *Joint Working IEEE/IFIP Conference on Software Architecture 2009 & European Conference on Software Architecture 2009*, Cambridge Royaume-Uni.

Van de Sompel, H., Lagoze, C., Bekaert, J., Liu, X., Payette, S., Warner, S., 2006. An Interoperable Fabric for Scholarly Value Chains. *D-Lib Magazine*, 12(10).

Wegner, P., 1996. Interoperability. *ACM Computing Survey*, 28(1), pp. 285–287.

Wiederhold, G., 1992. Mediators in the Architecture of Future Information Systems. *Computer*, 25(3), pp. 38–49.

Wiederhold, G., Genesereth, M., 1997. The conceptual basis for mediation services. *IEEE Expert: Intelligent Systems and Their Applications*, 12(5), pp. 38–47.

ACKNOWLEDGEMENTS

The work reported has been partially supported by the DL.org Coordination and Support Action, within FP7 of the European Commission, ICT-2007.4.3, Contract No. 231551.