Making Digital Library Content Interoperable

Leonardo Candela, Donatella Castelli, and Costantino Thanos

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR, Pisa, Italy
{name.surname}@isti.cnr.it

Abstract. The demand for powerful and rich Digital Libraries able to support a large variety of interdisciplinary activities has increased the need for "building by re-use" and sharing, especially when dealing with the content space. Interoperability is a central issue to satisfy these needs. Despite its importance, and the many attempts to address it done in the past, the solutions to this problem are today, however, still very limited. Main reasons for this slow progress are lack of any systematic approach for addressing the issue and scarce knowledge of the adopted solutions. Too often these remain confined to the systems they have been designed for. In order to overcome this lack, this paper proposes an Interoperability Framework for describing and analyzing interoperability problems and solutions related to use of content resources. It also discusses the many facets content interoperability has and provides a comprehensive and annotated portfolio of existing approaches and solutions to this challenging issue.

1 Introduction

Interoperability is among the most critical issues to be faced when building systems as "collections" of independently developed constituents (systems on their own) that should cooperate and rely on each other to accomplish larger tasks.

Digital Library (DL) interoperability is an issue affecting the Digital Library domain since its beginning. It was explicitly mentioned among the challenges of the Digital Library Initiative (Challenge Four) [11] in early nineties. At that time the issue was formulated as follows: *to establish protocols and standards to facilitate the assembly of distributed digital libraries*.

Recently, the demand for powerful and rich DLs able to support a large variety of interdisciplinary activities has increased the need for resource sharing. Interoperability solutions, which lay at the core of any approach supporting such sharing, have consequently become even more important than in the past.

Despite these facts and the critical role interoperability has, there is no developed theory driving the resolution of interoperability issues when they manifest.

Actually, there is no single definition of interoperability which is accepted by the overall community nor by the Digital Library community. Wegner [34] defines interoperability as "*the ability of two or more software components to cooperate despite differences in language, interface, and execution platform. It is a scalable form of reusability, being concerned with the reuse of server resources by clients whose accessing mechanisms may be plug-incompatible with sockets of the server*". He also identifies in *interface standardization* and *interface bridging* two of the major mechanisms for interoperation. Heiler [13] defines interoperability as "*the ability to exchange services and data*

M. Agosti, F. Esposito, and C. Thanos (Eds.): IRCDL 2010, CCIS 91, pp. 13–25, 2010.

with one another. It is based on agreements between requesters and providers on, for example, message passing protocols, procedure names, error codes, and argument types". He also defines *semantic interoperability* as ensuring "*that these exchanges make sense – that the requester and the provider have a common understanding of the "meanings" of the requested services and data. Semantic interoperability is based on agreements on, for example, algorithms for computing requested values, the expected side effects of a requested procedure, or the source or accuracy of requested data elements*". Park and Ram [23] define syntactic interoperability as "*the knowledge-level interoperability that provides cooperating businesses with the ability to bridge semantic conflicts arising from differences in implicit meanings, perspectives, and assumptions, thus creating a semantically compatible information environment based on the agreed concepts between different business entities*". They also define semantic interoperability as "*the application-level interoperability that allows multiple software components to cooperate even though their implementation languages, interfaces, and execution platforms are different*" [26]. In addition to that they state that emerging standards, such as XML and Web Services based on SOAP (Simple Object Access Protocol), UDDI (Universal, Description, Discovery, and Integration), and WSDL (Web Service Description Language), can resolve many application-level interoperability problems.

As recognized by Paepcke et Al. [22] ten year ago, over the years systems designers have developed different approaches and solutions to achieve interoperability. They have put in place a pragmatic approach and started to implement solutions blending into each other by combining various ways of dealing with the issues including *standards* and *mediators*. Too often these remain confined to the systems they have been designed for and lead to "from-scratch" development and duplication of effort whenever similar interoperability scenarios occur in different contexts.

This paper tackles the interoperability problem from a different perspective. This results from the understanding that the multitude of definitions sketched above, as well as the need to have blending solutions, are a consequence of the fact that interoperability is a very multifaceted and challenging issue that is not yet fully modeled in its own. The paper focuses on *digital library content interoperability*, i.e. the problem arising whenever two or more Digital Library "systems" are willing to interoperate by exploiting each other's content resources. Different systems have to remove the barriers resulting from the different models and "ways to manage" underlying their resources. The aim is to contribute to a better understanding of this interoperability problem and the relative solutions through a systematic and organized approach. The paper refrains from introducing its own definition of content interoperability in favor of an Interoperability Framework aiming at identifying the various facets characterizing this exemplar of interoperability. By exploiting this framework, "interoperability" problems and solutions can be modeled in a multifaceted space. This study is part of a more comprehensive approach to the Digital Library interoperability problem addressed from different perspectives (content, user, functionality, policy, quality, architecture) conducted as part of DL.org[1], an EU 7th FP project.

The remainder of this paper is structured as follows. Section 2 introduces the many facets content interoperability has and proposes a systematic approach to the under-

[1] `www.dlorg.eu`

standing of this issue. Section 3 identifies which are the most important properties characterizing an information object from the interoperability point of view and briefly reviews the techniques and formalisms proposed in the literature for modeling them. Section 4 identifies significant levels of content interoperability and gives the corresponding definitions. Section 5 describes and comments existing approaches enabling interoperability. Finally, Section 6 presents concluding remarks and future plans.

2 A Content Interoperability Framework in a Nutshell

According to the DELOS Reference Model [2], *content* is one of the six domains characterizing the Digital Library universe. In particular, in this domain there are all the entities held or included in a system to represent information in all its forms. *Information Object* is the most general concept characterizing the Content Domain. An Information Object is an instance of an abstract data type and represents any unit of information managed in the Digital Library universe, including text documents, images, sound documents, multimedia documents and 3-D objects, as well as data sets and databases. Information Objects also include *composite objects* and *collections* of Information Objects.

Any interoperability scenario involves two or more "systems" and one or more "resources" about which the involved systems are willing to be interoperable. For the sake of modeling, interoperability among many systems can always be reduced to "interoperation" between pair of actors, one of which performs an operation for the other one. At any given time, one of the two actor plays the role of *provider* of the resource to be exchanged while the other plays the role of a *consumer* of this resource.

Content interoperability is a *multi-layered* and *very context-specific* concept. It encompasses different levels along a multidimensional spectrum. Therefore, rather than aiming for a single, "one-size-fits-all" definition, it seems more promising to carefully identify the properties characterizing the DL content and define different levels of interoperability supporting technical and operational aspects of the interaction between content providers and consumers.

In the setting described above, any content interoperability scenario can be characterized by a framework consisting of the following four complementary axes:

- *resource model* (cf. Section 3). Any resource is described by a set of properties that capture its essential characteristics. The larger is the set of properties producer and consumer share the same understanding of, the wider is the exploitation that the consumer can perform of the resource;
- *interoperability level* (cf. Section 4). The same understanding of a model can occur at different level of "completeness". These levels constraint the type of interoperation that can occur. Typical exemplar of interoperability levels are *syntactic*, i.e. provider and consumer agree on the representation of the resource model or part of it, and *semantic*, i.e. provider and consumer agree on the meaning of the resource model or part of it;
- *reconciliation function* (cf. Section 5). The specific interoperability can be achieved using different approaches. Reconciliation functions can vary along a multidimensional spectrum including the dimension for "unilateral" approaches to "collaborative" approaches and the dimension for "non-regulatory" approaches to

"regulatory" approaches [10]. Moreover, it should materialize in some *architectural components* implementing it and a *protocol* through which the partaking systems operate. Two notable exemplars of reconciliation functions are standards and mediators. Standards are among the most consolidated approaches to achieve interoperability, while mediators have been introduced to guarantee a high-level of autonomy for the partaking systems.

- *benchmark*. Each reconciliation function has its own strengths and weaknesses, costs and benefits. Benchmark characterize these reconciliation function features. It may include, for example, *effectiveness*, i.e. the measure of how the approach is successful in achieving the expected result, *efficiency*, i.e. the measure of the ratio between the cost of the approach and the result achieved, and *flexibility*, i.e. the measure of how much the proposed approach is change-tolerant.

As this paper aims at presenting the problems and the solutions to interoperability from the perspective of the "content" domain, in the following sections we will use the above framework to analyze interoperability scenarios in which the resources involved are Information Objects. Note, however, that the described framework (in terms of its characterizing axes) is generic enough to be easily adapted to interoperability scenarios involving other kind of resources.

3 Digital Library Content Modeling

Operating on the DL content means operating on the Information Objects that populate it. Interoperability with respect to Content is thus achieved when the provider and the consumer systems are interoperable with respect to these Information Objects.

The model of an Information Object captures its distinguishing properties. When considered from an interoperability point of view, the model should capture both (*a*) properties concurring to form the *state* of the Information Object and (*b*) properties concurring to form the *setting* of the Information Object. Among the former set of properties we discuss below the modeling of *identifier*, *type*, *metadata*, *quality* and *protection*, while among the latter set we discuss the modeling of *context* and *provenance*.

Information Object Identifier. Information Object Identifiers are tokens bound to Information Objects that distinguish them from other Information Objects within a certain scope. They play a role that is similar to that of the Uniform Resource Identifiers (URIs) in the architecture of the World Wide Web [14], i.e. they represent a cornerstone in a scenario in which any party can share information with any other since they permit to identify such a shared information.

As discussed in [30], such identifiers should be "persistent" and "actionable", i.e. they should give access to the resources and should continue to provide this access, even when the associated resources are moved to other location or even to other organizations.

Identifier interoperability is necessary for the purpose of referring the target Information Objects similarly in the provider and consumer contexts.

Identifiers are often modeled using one of the many available standards. These include the *Uniform Resource Name (URN)*, *digital object identifier (DOI)*, *persistent URL* (PURL), the *Handle system* and the *Archival Resource Key* (ARK) [30]. In addition to these standards, there are other approaches for content-based identification ("fingerprinting").

Information Object Format. An Information Object Format captures the structural (and sometimes operational) properties of the Information Objects. It is a formal and intentional characterization of all the Information Objects having such a "type" or "data model". According to the Reference Model [2], Information Objects conceptually represent DL content in terms of a "graph" of digital objects associated with each other through relationships whose "label", i.e. name, expresses the nature of their association. The Format captures these kind of structure including any constrain.

Formats interoperability is necessary for the purpose of enabling the consumer of the objects to safely and/or efficiently execute operations over it based on the structural "assumptions" declared by the associated Information Object Format.

From the modeling perspective, Information Object Formats can range from *rigid data models*, where the model basically expresses "one" Information Object model allowing for light customizations (e.g. DSpace [29], Greenstone [37], Eprints [18] data models), to *flexible models*, where the model can potentially describe "any" information object model (e.g. Fedora data model [15]). The current trend goes in the direction of data model for complex information entities as resource aggregations or compound resources [4,3,17].

Information Object Metadata. Metadata is any structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource [21].[2] Metadata is often called *data about data* or *information about information*. Because of this potentially broad coverage captured by the term "metadata", the majority of interoperability problems risk to fall in this category. In fact, many different metadata schemes are being developed in a variety of user environments and disciplines to better serve the specific needs and to capture, through metadata, the distinguishing properties of the resources that are deemed proper to the scope.

Metadata interoperability is necessary for the purpose of enabling the consumer of the object to gather / be informed on some characteristics of the Information Object the partaking systems are willing to interoperate. The wider is the set of resource properties captures through the metadata, the larger is the potential understanding the consumer might achieve and the richer is the functionality it will be able to realize by exploiting the reached understanding.

From the modeling perspectives, Information Object Metadata capture a set of properties and because of this classic data structure for representing them are exploited, e.g. key-values models. Several schemas have been produced to represent them, e.g. Dublin Core[3], MAchine-Readable Cataloging (MARC)[4], Metadata Encoding and Transmission

[2] Information Object in our terminology.

[3] `http://dublincore.org/`

[4] `http://www.loc.gov/marc/`

Standard (METS)[5], Metadata Object Description Schema (MODS)[6], ISO 19115[7]. The majority of them are dedicated to capture bibliographic information including cataloguing and classification details and are encoded in XML. In addition to such schemas, others conceived to serve application specific needs are continuously defined. In order to promote the interoperability of these application specific schemas, the *application profile* [12] approach represents a good practice. Exemplars are the Darwin Core[8] and the Europeana Semantic Element Set[9].

Information Object Quality. Quality is a kind of meta-property as it describes various "characteristics" of Information Object properties and sub-properties. Information Object Quality can be pragmatically defined as "the fitness for use of the information provided". It is a multi-faceted concept to the definition of which different dimensions concur, each capturing a specific aspect of object quality. More specifically, quality dimensions or parameters can refer either to the *extension* of data, i.e. to data values, or to their *intension*, i.e. to their schema/format. In addition to that, the need for capturing the "quality of the quality", i.e. how the quality parameters values are produced or assessed, is a critical aspect to be considered. The data quality literature provides a thorough classification of data quality dimensions. By analyzing existing classifications, it is possible to define a basic set of data quality dimensions, including *accuracy*, *completeness*, *consistency*, and *timeliness* [1]. Because of the fundamental and pervasive role quality plays in any Information System, the Digital Library Reference Model [2] includes an entire domain to capture it.

Quality interoperability is necessary for the purpose of enabling the consumer to exploit every kind of Information Object in a conscious manner, i.e. being aware of the qualitative aspects of this kind of information and thus being able to put in place proper actions on top of it.

From the modeling perspectives, quality characteristics closely resemble metadata, actually they can be considered a kind of metadata. Because of this, quality perspectives might be part of a metadata record. A machine-readable Quality Profile should be associated with an Information Object containing quality assertions.

Information Object Protection. This is a highly complex problem that includes three sub-problems: *security*, *integrity*, and *privacy*. Security refers to the protection of content against accidental or intentional disclosure to unauthorized users or unauthorized uses. Integrity refers to the process of ensuring that the content remains an accurate reflection of the universe of discourse it is modeling or representing. Privacy refers to the rights of content providers to determine when, how, and to what extent their content is to be transmitted to content consumers. Protection represents one of the aspects of Policy, i.e. the set of conditions, rules, terms or regulations governing the operation of

[5] http://www.loc.gov/standards/mets/

[6] http://www.loc.gov/standards/mods/

[7] http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020

[8] http://www.tdwg.org/activities/darwincore/

[9] http://www.europeana.eu/

any Digital Library. Because of the fundamental and pervasive role played by Policy, the Digital Library Reference Model [2] includes an entire domain to capture it.

Information Object Protection becomes a concern when exchanges cross the "trust boundary"; beyond this logical line of demarcation it is rarely possible for an originating entity to assume that all potential recipients are authorized to access all information they are capable of discovering and consuming.

Information Object Protection interoperability is necessary for the purpose of enabling the consumer to be aware of the policies governing the Information Object and thus to be able to put in place proper actions on top of it.

From the modeling perspectives, this kind of policy represents an information that can be stored in the metadata attached to the Information Objects, e.g. in the "rights" element of Dublin Core. In addition to that, there are specific languages to represent policies in a declarative manner like the eXtensible Access Control Markup Language (XACML)[10].

Information Object Context. Context is the set of all "setting" information that can be used to characterize the relation between the Information Object and the "external world" [9] surrounding it. Context represents a distinguishing and complementary information that enriches the informative payload captured by the Information Object itself.

Information Object Context interoperability is necessary for the purpose of enabling the consumer of the Information Object to behave as a context-aware system, i.e. a system that is conscious of the situations surrounding the Information Object and can adapt its consumption accordingly.

From the modeling perspectives, this information closely resembles metadata, actually it can be considered a kind of metadata. Strang and Linnhoff-Popien [28] pointed out most relevant approaches for context modeling including key-value pairs, markup scheme and ontology-based ones. Najar et al. [20] recently revised this survey to capture some use of context from content adaptation to service adaptation.

Information Object Provenance. Provenance, also called *lineage*, pertains to the derivation history of the Information Object starting from its original sources, that is, it describes the process that led the object to be in its current state. It is a description of the origin and/or of the descendant line of data. Keeping track of provenance has become, in the last decade, crucial for the correct exploitation of data in a wide variety of application domains.

Information Object Provenance interoperability is necessary for the purpose of enabling the consumer of the Information Object to be aware of the history leading to its current stage and thus to perform exploitation actions that take this knowledge into account.

From the modeling perspectives, a number of provenance models have been proposed ranging from generic models such as OPM [19], that aims to model any kind of provenance, to domain specific models such as the FlyWeb provenance model [38]. The more domain specific a model is, the more restrictive the domain of its provenance subjects is. These models usually materialize in XML or RDF files.

[10] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

4 Levels of Content Interoperability

As discussed above, the same understanding of a model can occur at different level of "completeness". In addition to that, the information object model comprises several characteristics (properties), thus different level of completeness can be achieved among the systems involved into an interoperability scenario with respect to the overall amount of these characteristics as well as with respect to specific characteristics.

The following levels are considered to be relevant for content interoperability.

Technical/Basic Interoperability is mainly implemented at any level of the Information Object model, i.e. with respect to any characteristic described above. Common tools and interfaces provide the consumer with a superficial uniformity of the characteristics of the provider Information Objects which allows him to access them. However, when implementing this level of interoperability abstraction the task of providing any coherence of the content relies on human intelligence.

Syntactic interoperability is concerned with ensuring that the abstract syntax of "target" Information Object characteristics, in particular the metadata and the related ones, is understandable by any other application (recipient) that was not initially developed for this purpose;

Semantic Interoperability is concerned with ensuring that the precise meaning of "target" Information Object feature is understandable by any other application (recipient) that was not initially developed for this purpose. Semantic interoperability is achieved only when Information Object producer and consumer agree on the meaning of the Information Object (actually of its properties) they exchange;

Operational Interoperability is concerned with ensuring the effective use of the "target" Information Object by the recipient in order to perform a specific task. This recipient ability is guaranteed by the fact that both originator and recipient share the same understanding with respect to the data quality property;

Secure Interoperability is concerned with ensuring secure Information Object "exchanges" between the involved systems. This must be conducted with sufficient *context* so that the purpose to which the recipient applies the received Information Object is consistent with its use as intended by the originator [6];

These levels of interoperability may be subjected to dependencies: operational/secure interoperability is only possible if semantic interoperability is ensured; semantic interoperability is only possible if syntactic interoperability is ensured; syntactic interoperability is only possible if technical interoperability is achieved.

5 Content Reconciliation Approaches

The most common solutions and approaches to Content reconciliation can be classified in two main classes: *standard-based* and *mediator-based*. Standard-based approaches rely on the usage of an agreed standard (or a combination of them) that achieves a certain amount of homogeneity between the involved systems. Mediator-based approaches are based on the development of a component specifically conceived to host the interoperability machinery, i.e. a component mediating between the the involved systems and aiming to reconcile the content heterogeneity.

However, because of the amount of heterogeneity to be reconciled, solutions properly mixing approaches belonging to the two classes can be successfully deployed. In the remainder of this section we describe the most common exploited ones while dealing with the Information Object properties previously discussed.

5.1 Standard-Based Approaches

Standards, either *de jure* or *de facto*, represent one of the most common and well recognized approach to attack interoperability issues at any level and in any domain. In this context, the term "standard" is intended with the very wide meaning of common agreed specification. Moreover, it is important to recall here that the success or failure of standards does not depend on technical merits only, social and business considerations coming into play.

Potentially, standards are everywhere, i.e. a standard can be defined to characterize every single aspect of a "system". Because of this characteristic, the list of standard reported in this section does not aim to be exhaustive nor complete with respect to the standardization initiatives.

The standards of interest for Content Interoperability can be classified in two main non-disjoint classes: standards for content representation and standards for content exchange. Exemplars of the first class are the various formats and schemas that have been discussed in Section 3 and that are exploited to represent content features, e.g. Dublin Core for the metadata, MPEG-21 for Information Object Format as well as generic standards like XML[11] and RDF[12]. Exemplars of the second class are generic standards like RSS and Atom as well as OAI-PMH [16] and OAI-ORE [32,31], two well known approaches to interoperability in the Digital Library content domain.

In addition to these traditional standard initiatives, we include approaches like Application Profiles and Derivation in this category of approaches.

Application Profiles. Even within a particular information community, there are different user requirements and special local needs. The details provided in a particular schema may not meet the needs of all user groups. There is often no schema that meets all needs. To accommodate individual needs, an *application profile* [12] might be defined. In this approach, an existing schema is used as the basis for description in a particular digital library or repository, while individual needs are met through a set of specific application guidelines or policies or through adaptation or modification by creating an application profile for application by a particular interest group or user community.

Derivation. In this approach [5], a new schema is derived from an existing one. In a collection of digital repositories where different components have different needs and different requirements regarding description details, an existing complex schema may be used as the "source" or "model" from which new and simpler individual schemas may be derived. Specific derivation methods include adaptation, modification, expansion, partial adaptation, translation, etc. In each case, the new schema is dependent on the source schema.

[11] http://www.w3.org/XML/

[12] http://www.w3.org/RDF/

5.2 Mediator-Based Approaches

As already mentioned, a key concept enabling the content interoperation among heterogeneous systems is *mediation* [35]. This concept has been used to cope with many heterogeneity dimensions ranging from terminology to representation format, transfer protocols, semantics, etc. [27,36].

The content mediation concept is implemented by a mediator, which is a software device that supports (*a*) a mediation schema capturing user (originator and recipients) requirements, and (*b*) an intermediation function that describe how to represent the distributed information object sources in terms of the mediation schema.

A key feature which characterizes a mediation process is the kind of the reconciliation function implemented by a mediator. There are three main approaches. *Mapping* which refers to how information object structures, properties, relationships are mapped from one representation scheme/formalism to another one equivalent from the semantic point of view. *Matching* which refers to the action of verifying whether two strings/patterns match, or whether semantically heterogeneous information objects match. *Integration* which refers to the action of combining information objects residing in different heterogeneous sources and providing users with a unified view of these objects (or combining domain knowledge that is expressed in domain ontologies).

At each level of content interoperability identified in Section 4 a specific mediation process might be applied. *Technical mediation* enables the linking of the systems and services through the use of common tools, open interfaces, interconnection services, and middleware. *Syntactic mediation* is mainly implemented at the information object metadata level and makes it possible to bridge the differences of the metadata formats at the syntactic level. *Semantic mediation* enables the bridging of the differences of the exchanged information object at the semantic level allowing thus for information object to be exchanged according to semantic matching. *Operational mediation* guarantees that both the information object originator and recipient share the same quality dimensions described in the quality profile associated with the exchanged information object. *Protection mediation* enables the recipient to use the received information object without violating the security, integrity and privacy constraints associated with it and to protect it from unauthorized users.

Approaches put in place by the mediator service are based on a preliminary knowledge of the heterogeneities among the partaking entities and on how to reconcile them including the usage of a pivot schema or *lingua franca* and a series of mappings and rewriting rules. These permit to realize crosswalks and might be based on the use of some ontology.

Crosswalks. A crosswalk is a mapping of the elements, semantics, and syntax from one scheme to another. Currently, crosswalks are by far the most commonly used approach to enable interoperability between and among metadata schemes.

Ontology-based Approaches. Ontologies were developed by the the Artificial Intelligence community to facilitate knowledge sharing and reuse. They are largely used for representing domain knowledge. An ontology is a formal, explicit specification of a shared abstract model of some domain knowledge in the world that identifies that domain's relevant concepts [8].

Ontologies have been extensively used in supporting all the three content mediation approaches, i.e. mapping, matching and integration, because they provide an explicit and machine-understandable conceptualization of a domain. They have been used in one of the three following ways [33]. In the *single ontology approach*, all sources schemas are directly related to a shared global ontology that provides a uniform interface to the user [7]. In the *multiple ontology approach*, each data source 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 preceding approaches is used.

Ontology provides a framework within which the semantic matching/mapping process can be carried out by identifying and purging semantic divergence. Semantic divergence occurs where the semantic relationship between the ontology and the representation is not direct and straightforward [24].

6 Concluding Remarks

In this paper we report some preliminary results concerning the study of DL Content Interoperability. In particular, we have presented an Interoperability Framework with the aim to contribute to a better understanding of this challenging problem and its relative solutions through a systematic and organized approach. The most important properties of content, from the interoperability perspective, have been introduced and discussed. The main modeling techniques for the representation of them have been reviewed. The relevant levels of content interoperability have been identified and defined. Finally, a number of content interoperability approaches have been presented and discussed.

The proposed Interoperability Framework is currently been used in collecting interoperability requirements and in designing appropriate solutions to this problem in the context of the D4Science-II project[13]. This project aims at creating an initial ecosystem of interoperable data infrastructures and repository systems capable of exploiting each other's content resources. The type of content managed by the components of this ecosystem is very heterogeneous. This makes the process of requirement collection, analysis and design particularly complex. So far the framework has guided a systematic collection of requirements from the different actors. This systematic approach has largely facilitated the analysis phase. No particularly significant lack has been identified in the Content Framework in this initial phase. We expect a more considerable evaluation of the framework during the design phase when also the solutions will have to be modeled and described.

Acknowledgments. 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). We are grateful for many helpful suggestions from Detlev Balzer, Stefan Gradmann, C.H.J.P. Hendriks, Carlo Meghini, Luc Moreau and John Mylopoulos, members of the DL.org Content Working Group.

[13] `www.d4science.eu`

References

1. Batini, C., Scannapieco, M.: Data Quality: Concepts, methodologies and techniques. Springer, Heidelberg (2006)
2. 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.: The DELOS Digital Library Reference Model - Foundations for Digital Libraries. In: DELOS: a Network of Excellence on Digital Libraries (February 2008) ISSN 1818-8044, ISBN 2-912335-37-X
3. Candela, L., Castelli, D., Manghi, P., Mikulicic, M., Pagano, P.: On Foundations of Typed Data Models for Digital Libraries. In: Fifth Italian Research Conference on Digital Library Management Systems, IRCDL 2009 (2009)
4. Candela, L., Castelli, D., Pagano, P., Simi, M.: From Heterogeneous Information Spaces to Virtual Documents. In: Fox, E.A., Neuhold, E.J., Premsmit, P., Wuwongse, V. (eds.) ICADL 2005. LNCS, vol. 3815, pp. 11–22. Springer, Heidelberg (2005)
5. Chan, L.M., Zeng, M.L.: Metadata Interoperability and Standardization – A Study of Methodology Part I Achieving Interoperability at the Schema Level. D-Lib. Magazine 12(6) (June 2006)
6. Connors, C.L., Malloy, M.A., Masek, E.V.: Enabling Secure Interoperability Among Federated National Entities: It's a Matter of Trust. Technical report, MITRE Corporation (2007)
7. Cruz, I.F., Xiao, H.: Using a layered approach for interoperability on the semantic web, p. 221 (2003)
8. Cruz Huiyong, I., Cruz, I.F., Xiao, H.: The role of ontologies in data integration. Journal of Engineering Intelligent Systems 13, 245–252 (2005)
9. Dey, A.K.: Understanding and using context. Personal Ubiquitous Comput. 5(1), 4–7 (2001)
10. Gasser, U., Palfrey, J.: Breaking Down Digital Barriers - When and How Interoperability Drives Innovation. Berkman Publication Series (November 2007)
11. Griffin, S.M.: NSF/DARPA/NASA Digital Libraries Initiative - A Program Manager's Perspective. D-Lib. Magazine (July/August 1998)
12. Heery, R., Patel, M.: Application profiles: mixing and matching metadata schemas. Ariadne 25 (2000)
13. Heiler, S.: Semantic interoperability. ACM Comput. Surv. 27(2), 271–273 (1995)
14. Jacobs, I., Walsh, N.: Architecture of the World Wide Web, vol.1. Technical report, W3C (December 2004)
15. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: An Architecture for Complex Objects and their Relationships. Journal of Digital Libraries, Special Issue on Complex Objects (2005)
16. Lagoze, C., Van de Sompel, H.: The open archives initiative: building a low-barrier interoperability framework. In: Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 54–62. ACM Press, New York (2001)
17. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R., Warner, S.: ORE Specification - Abstract Data Model. Technical report, Open Archives Initiative (2008)
18. Millington, P., Nixon, W.J.: EPrints 3 Pre-Launch Briefing. Ariadne 50 (2007)
19. Moreau, L., Plale, B., Miles, S., Goble, C., Missier, P., Barga, R., Simmhan, Y., Futrelle, J., McGrath, R., Myers, J., Paulson, P., Bowers, S., Ludaescher, B., Kwasnikowska, N., Van den Bussche, J., Ellkvist, T., Freire, J., Groth, P.: The open provenance model (v1.01). Technical report, University of Southampton (July 2008)
20. Najar, S., Saidani, O., Kirsch-Pinheiro, M., Souveyet, C., Nurcan, S.: Semantic representation of context models: a framework for analyzing and understanding. In: CIAO 2009: Proceedings of the 1st Workshop on Context, Information and Ontologies, pp. 1–10. ACM, New York (2009)

21. National Information Standards Organization. Understanding Metadata. NISO Press (2004)
22. Paepcke, A., Chang, C.-C.K., Winograd, T., García-Molina, H.: Interoperability for Digital Libraries Worldwide. Communications of the ACM 41(4), 33–42 (1998)
23. Park, J., Ram, S.: Information Systems Interoperability: What Lies Beneath? ACM Trans. Inf. Syst. 22(4), 595–632 (2004)
24. Partridge, C.: The role of ontology in integrating semantically heterogeneous databases. Technical Report 05/02, LADSEB-CNR (2002)
25. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal 10(4), 334–350 (2001)
26. Ram, S., Park, J., Lee, D.: Digital libraries for the next millennium: Challenges and research directions. Information Systems Frontiers 1(1), 75–94 (1999)
27. Spalazzese, R., Inverardi, P., Issarny, V.: 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., CONNECT (2009)
28. Strang, T., Linnhoff-Popien, C.: A context modeling survey. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) UbiComp 2004. LNCS, vol. 3205, Springer, Heidelberg (2004)
29. Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., Smith, M.: The DSpace Institutional Digital Repository System: current functionality. In: Proceedings of the third ACM/IEEE-CS joint conference on Digital libraries, pp. 87–97. IEEE Computer Society, Los Alamitos (2003)
30. Tonkin, E.: Persistent Identifiers: Considering the Options. Ariadne 56 (2008)
31. Van de Sompel, H., Lagoze, C., Bekaert, J., Liu, X., Payette, S., Warner, S.: An Interoperable Fabric for Scholarly Value Chains. D-Lib. Magazine 12(10) (October 2006)
32. Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., Warner, S.: Rethinking Scholarly Communication - Building the System that Scholars Deserve. D-Lib. Magazine 10(9) (September 2004)
33. Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S.: Ontology-based integration of information — a survey of existing approaches. In: Stuckenschmidt, H. (ed.) IJCAI 2001 Workshop: Ontologies and Information Sharing, pp. 108–117 (2001)
34. Wegner, P.: Interoperability. ACM Comput. Surv. 28(1), 285–287 (1996)
35. Wiederhold, G.: Mediators in the Architecture of Future Information Systems. Computer 25(3), 38–49 (1992)
36. Wiederhold, G., Genesereth, M.: The conceptual basis for mediation services. IEEE Expert: Intelligent Systems and Their Applications 12(5), 38–47 (1997)
37. Witten, I., Bainbridge, D., Boddie, S.: Greenstone - Open-Source Digital Library Software. D-Lib. Magazine 7(10) (October 2001)
38. Zhao, J., Miles, A., Klyne, G., Shotton, D.: Linked data and provenance in biological data webs. Briefings in bioinformatics 10(2), 139–152 (2009)