

# Integrating Cultural Heritage Information Sources

Lina Bountouri<sup>1</sup>

Manolis Gergatsoulis<sup>1</sup>

Christos Papatheodorou<sup>1,2</sup>

<sup>1</sup>Department of Archive and Library Sciences, Ionian University, Greece

<sup>2</sup>Digital Curation Unit, Athena Research Centre, Greece  
{boudouri,manolis,papatheodor}@ionio.gr

## ABSTRACT

Our research work deals with the integration of information sources coming from Cultural Heritage (CH) institutions (such as archives, libraries and museums). In order to promote semantic interoperability in CH resources, we propose: a) an ontology-based integration architecture based on the use of CIDOC CRM ontology, b) mappings from various CH metadata schemas to the ontology, and c) metadata query transformation to queries on the ontology.

## Categories and Subject Descriptors

H.3.3 [Information Systems]: Information Search and Retrieval

## Keywords

Ontology-based Integration, Metadata interoperability, Cultural Information, CIDOC CRM, EAD, Query Transformation

## 1. INTRODUCTION

Nowadays, there is a rapidly growing number of heterogeneous resources available on the Web. The heterogeneity of these resources comes up in various levels, such as the syntactic, schematic and semantic level, often even in the same domain of interest (i.e. eGovernment, Cultural Heritage etc.). With the intention of providing integrated access to heterogeneous resources, interoperability becomes an important issue. *Interoperability* is succeeded when it is ensured that information can be searched, exchanged, transferred, used and understood by different information services and systems [8].

Given the research efforts towards the development of the Semantic Web during the last years, there is an increasing need to deal with heterogeneity issues in the semantic level. Those issues are produced by semantic conflicts arising from the fact that the meaning of the data can be expressed in different ways and with different interpretations [12]. *Se-*

*mantic integration* can be considered as a vital part of data integration oriented to solve semantic heterogeneity problems “by using conceptual representations of the data and of their relationships to eliminate possible heterogeneities” [3].

*Ontologies* are one of the main Semantic Web infrastructures promoting semantic integration needs. They have a vital role in interoperability scenarios, given that they can semantically conceptualize a domain and they can be used as an umbrella of terms and meanings expressing the same subject or concept. In this context, ontologies can be considered as an important building block for integration architectures [6]. They are preferred in regard to other schemas, due to their ability to conceptualize particular domains of interest and express their rich semantics. One of the main roles of an ontology in an interoperability scenario is to promote the semantic integration, acting as a *mediated schema* between heterogeneous information systems [1, 17, 9, 3].

Works related to ontology-based integration [13], usually put emphasis on element and structure level mappings (i.e. elements to classes, attributes to properties etc.). In [4], XML data of local sources are mapped to an RDFS local ontology, created by transforming the XML elements and attributes to RDFS classes and properties. In this method the structure of an XML local source inside the local RDFS ontology is preserved. Then, local ontologies are merged to a global ontology for unified access and semantic integration of local data sources. However, the effectiveness of those approaches in mapping really complex semantically data structures, such as metadata schemas, has not yet been tested.

In this paper we investigate the problem of the integration of information sources coming from Cultural Heritage (CH) institutions (such as archives, libraries and museums). In order to promote semantic interoperability in CH resources, we propose an ontology-based integration architecture based on the use of CIDOC CRM ontology [2]. CIDOC CRM’s main role is to facilitate information exchange and integration as the mediated schema of heterogeneous CH information sources, into which metadata originating from diverse sources can be semantically mapped and integrated. Based on that fact, we define mappings from various CH metadata schemas to the ontology and we present an example based on archival metadata commonly used in Digital Libraries. Finally, we suggest metadata query transformation to queries on the ontology.

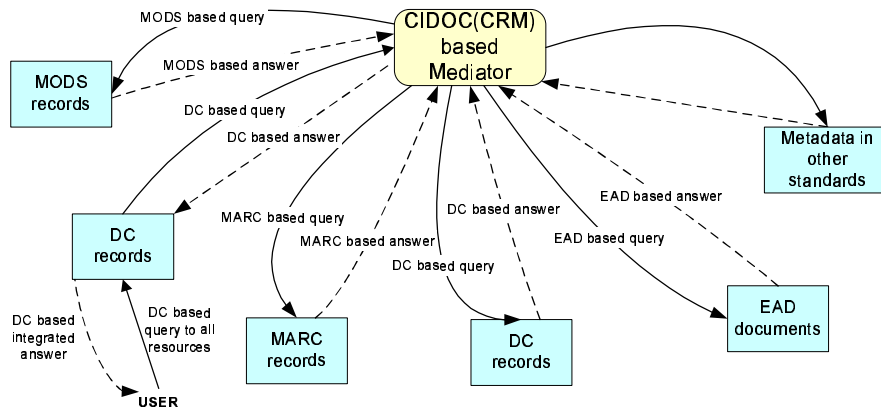


Figure 1: An Ontology-Based Mediator.

The paper is structured as follows: In Section 2, the proposed CIDOC CRM ontology-based integration architecture is analyzed and a brief view of the prerequisite mappings between archival metadata encoded in EAD standard [14] and CIDOC CRM is presented. In Section 3, we present indicative queries on EAD metadata, aiming to demonstrate the XPath adaptability and expressiveness as a query language for archival metadata. Furthermore we transform the XPath expressions to equivalent CIDOC CRM queries. Finally, conclusions and future research goals are presented in Section 4.

## 2. CIDOC CRM-BASED INTEGRATION

### 2.1 The proposed architecture

Managing heterogeneous data is a common issue in organizations that document cultural information. Those organizations dispose and develop various collections with diverse types of material (manuscripts, archival fonds, museum objects and collections, digital resources etc), which are described by different metadata schemas, according to their documentation and retrieval needs.

The integration scenario that we propose is based on the exploitation of the CIDOC CRM ontology as a mediated schema. CIDOC CRM is a conceptual model, composed of *entities* (classes), which are organized into a hierarchy and semantically related to each other with *properties*. All metadata schemas to be integrated should be mapped to the CIDOC CRM and vice versa.

In Figure 1 the proposed architecture is presented: a set of data sources exists, each of them encoded in a possibly different metadata schema (such as Dublin Core (DC) [5], Metadata Object Description Schema (MODS) [15], Encoded Archival Description (EAD) etc). All these metadata schemas are mapped to CIDOC CRM. As a result, a user can execute his queries to a local data source depending on the restrictions of the local metadata schema. The local query engine promotes the query to the mediator which translates the query to suitable forms, using the appropriate mappings, and forwards them to be answered by the other sources. Finally, the results are collected from the various sources and returned to the user. A complete analysis of our integration scenario and the mapping issues between metadata and

ontologies can be found in [17].

### 2.2 Mapping EAD to CIDOC

In this section, we shortly present the EAD to CIDOC CRM mappings as an indicative part of our proposed integration architecture.

One of the most widely implemented schemas in CH Digital Libraries is the Encoded Archival Description (EAD) [14]. EAD is the international metadata standard for encoding archival finding aids so as to make archival resources accessible to users via the Web. Archival description documents the archive, which is a complex set of materials that share common provenance [16]. The description involves a hierarchical and progressive documentation, beginning with the description of the whole, and then proceeds to define and describe its sub-components, the sub-components of sub-components, and so on. The exploitation of the flexible and tree structure based XML language allows EAD to introduce a machine readable form of the archives' multi-level structure.

EAD metadata are mainly encapsulated in two mandatory parts. The first one includes the metadata information for the archival description itself (`<eadheader>`), for instance, who, when and based on which rules someone created the archival description. The second part includes the information about the archive itself (`<archdesc>`), such as the title, the date(s) of creation and the origination of an archive.

The mapping methodology between the metadata schemas and CIDOC CRM is based on a path-oriented approach. A mapping from a source schema to a target schema transforms each instance of the source schema into a valid instance of the target schema [11]. Hence, we interpret the metadata paths to semantically equivalent CIDOC CRM paths.

An EAD path is a sequence of EAD elements and subelements, starting from the schema root element `<ead>` separated by the slash symbol (`/`). For example, the path `/ead/archdesc/did/unittitle` documents the title of the archive.

A CIDOC CRM path is a sequence of `class`  $\rightarrow$  `property`  $\rightarrow$  `class`. For instance, the equivalent CIDOC CRM path to

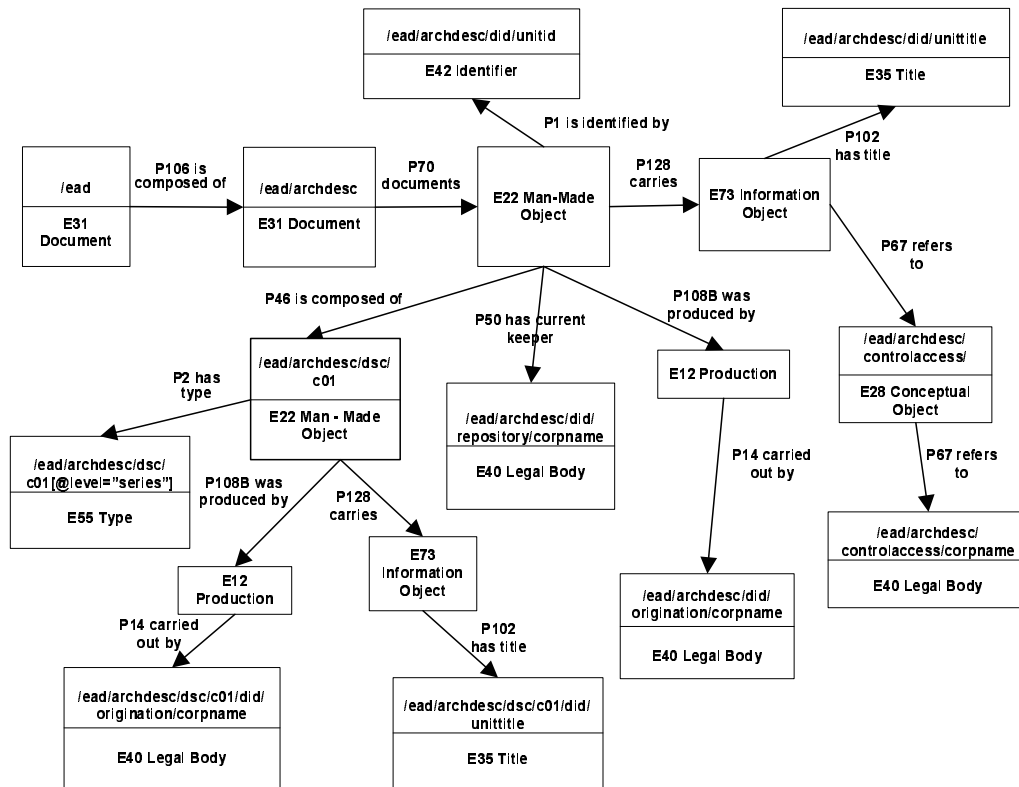


Figure 2: A fragment of EAD to CIDOC CRM mapping.

the afore-mentioned EAD path is <sup>1</sup>: E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P128 carries → E73 Information Object → P102 has title → E35 Title.

In Figure 2 a fragment of the EAD to CIDOC CRM mapping is presented. Every box represents a CIDOC CRM class. When the box is divided in two, the upper part indicates the EAD path mapped to the CIDOC CRM class included in the lower part of the box. Boxes are linked through arrows that represent the CIDOC CRM properties. In case a property is used according to its inverse property name, it is characterized by the letter “B” as part of its name (i.e. P108B was produced by).

While mapping EAD to CIDOC CRM, the semantic richness of the ontology becomes obvious, since it allows the explicit definition of the notions implied in EAD. For example, the /ead/archdesc path is mapped to the following CRM path: E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P128 carries → E73 Information Object, denoting that the EAD document (E31 Document) comprises the following (P106 is composed of): a) identifiable immaterial items that make

<sup>1</sup>The notation Enn, Pnn corresponds to CIDOC entities and properties respectively.

propositions about reality (E31 Document) and document (P70 documents) b) the archive as a physical object created by human activity (E22 Man-Made Object) that carries (P128 carries) c) immaterial items that include human memory and do not depend on any particular physical carrier (E73 Information Object). The subelements of Archival Description are linked either to E22 Man-Made Object, when they provide information about the archive as a physical object, or to E73 Information Object, when they provide information that refer to the archive as an informational carrier.

The corporate name of the archive’s creator /ead/archdesc/did/origination/corpname, which is one of the main notions in EAD providing provenance information, is mapped to the class E40 Legal Body through the following CIDOC CRM path: E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P108B was produced by → E12 Production → P14 carried out by → E40 Legal Body. The specific CIDOC CRM path denotes that the archive as a product of human activity (E22 Man-Made Object) produced (E12 Production) by an institution or group of people that have obtained a legal recognition as a group and can act collectively (E40 Legal Body).

It is worthy of note that the class E40 Legal Body is linked to the physical object view of the archive (E22 Man-Made

Object) through the class E12 Production. The use of this class reveals one of the main CIDOC CRM's characteristics: *Event - orientation*. More analytically, the main notions of the CIDOC CRM ontology are the temporal entities and events, and the presence of classes, such as Actors, Dates, Places, Objects, etc. implies their participation to an event or an activity [7]. Therefore, in the EAD to CIDOC CRM mappings, the archive's creator and all the related information to the archive's creation surround the class E12 Production, which is a class that comprises activities that are designed to create one or more new items.

The class E40 Legal Body, as part of CIDOC CRM paths, is also mapped to various distinct EAD paths, such as:

- `/ead/archdesc/controlaccess/corpname`, which is a path denoting that the archive refers to specific corporate names (`<corpname>`) as access points (`<controlaccess>`). This path is mapped to the CIDOC CRM path: E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P128 carries → E73 Information Object → P67 refers to → E28 Conceptual Object → P67 refers to → E40 Legal Body. Since the archive as a physical object (E22 Man-Made Object) cannot contain or be related to access points, such information is related to it through its view as an information carrier (E73 Information Object). As a consequence, the information carrier view of the archive can be associated to non-material products of human minds (E28 Conceptual Object), such as the names and subjects included in access points.
- `/ead/archdesc/dsc/c01/did/origination/corpname`, which describes the creator's (`<origination>`) corporate name (`<corpname>`) of the archival component (`<c01>`) included in the archive. EAD is a multi-level standard that includes the description of the archive as a whole and the description of its components parts. This path is mapped to the CIDOC CRM path: E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P46 is composed of → E22 Man-Made Object → P108B was produced by → E12 Production → P14 carried out by → E40 Legal Body. In order to define the multi-level structure of the archival components in CIDOC CRM, the class E22 Man-Made Object representing the archive is linked to its subcomponents (E22 Man-Made Object) via the property P46 is composed of. This property allows the instances of the archive (E22 Man-Made Object) to be analyzed into sub-component elements, which are themselves instances of the class E22 Man-Made Object, hence create a hierarchy of parts.
- `/ead/archdesc/did/repository/corpname`, which defines the corporate name (`<corpname>`) of the institution (`<repository>`) which is responsible for providing intellectual access to the archive. The specific path is mapped to the CIDOC CRM path: E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P50 has current keeper → E40 Legal Body. The property P50 has current keeper identifies that an institution (E40 Legal Body) has custody of the archive (E22 Man-Made Object) at the time this property was recorded.

A more detailed description of EAD to CIDOC CRM mapping is presented in [17].

### 3. EAD TO CIDOC CRM QUERY TRANSFORMATION

#### 3.1 Querying EAD metadata using XPath

XPath is a language used to identify specific parts of XML documents by allowing the processing of values. XPath denotes the XML nodes by position, relative position, type, content, and several other criteria [19]. EAD is an XML based standard, therefore queries over EAD documents could be expressed in terms of XPath language. In this Section we present a part of an EAD document which includes the metadata provided for the archive itself (`<archdesc>`); to continue with, we present a representative sample of users' queries expressed in XPath.

In Example 3.1, the archival description is on the level of "fonds". Basic descriptive identification information for the archive are given inside the `<did>` element, such as the title (`<unittitle>`), the creation dates (`<unitdate>`), the identifier of the archive (`<unitid>`) and its creator (`<origination>`). Administrative and supplemental information are provided through the `<bioghist>` and `<controlaccess>` elements, while description of subordinate components is presented inside the `<dsc>` element. In detail, two subordinate components are provided through the use of `<c01>` elements. Both components are on the level of "series" and they include basic descriptive identification information for the archival series, such as the title (`<unittitle>`), the creation dates (`<unitdate>`) etc.

EXAMPLE 3.1. *In this example we present a fragment of an EAD document:*

```
<ead>
<eadheader>...</eadheader>
<archdesc level="fonds">
  <did>
    <unitid countrycode="GR" repositorycode="IU">
      ARC.14</unitid>
    <unittitle>Ionian University Archive</unittitle>
    <unitdate>1984 - 2007</unitdate>
    <origination>
      <corpname>Ionian University</corpname>
    </origination>
  </did>
  <bioghist>
    <p>The Ionian University was founded in 1984...</p>
  </bioghist>
  <controlaccess>
    <corpname>Ionian University</corpname>
  </controlaccess>
  <dsc>
    <c01 level="series">
      <did>
        <unitid countrycode="GR" repositorycode="IU">
          ARC.14/1</unitid>
        <unittitle>R.C. Archives</unittitle>
        <unitdate>1998 - 2007</unitdate>
        <origination>
          <corpname>Research Committee</corpname>
        </origination>
      </did>
    </c01>
  </dsc>
</c01>
<c01 level="series">
  <did>
```

```

<unitid countrycode="GR" repositorycode="IU">
  ARC.14/2</unitid>
<unittitle>I.U. Library Archives</unittitle>
<unitdate>1998 - 2000</unitdate>
<origination>
  <corpname>I.U. Library</corpname>
</origination>
</did>
</c01>
</dsc>
</archdesc>
</ead>

```

Representative queries on the EAD document of Example 3.1, are given in Example 3.2.

EXAMPLE 3.2. *Some XPath queries, expressed in XPath on the EAD document of Example 3.1 are:*

**Query 1:** *Find the title of the archive.*

**EAD XPath:** `/ead/archdesc/did/unittitle`

**Query 2:** *Find the creator (corporate name of the originator) of the archive.*

**EAD XPath:** `/ead/archdesc/did/origination/corpname`

**Query 3:** *Find the creator (corporate name of the originator) of the series titled "I.U. Library Archives".*

**EAD XPath:** `/ead/archdesc/dsc/c01[@level="series"]/did/unittitle="I.U. Library Archives"/origination/corpname`

**Query 4:** *Find the title of the archive which is identified as "ARC.14/2".*

**EAD XPath:** `/ead/archdesc/did[unitid="ARC.14/2"]/unittitle`

## 3.2 Querying CIDOC CRM

In this section, we present the CIDOC CRM paths produced from the EAD to CIDOC CRM mapping and their expression in a RQL-like syntax [18]. RQL [10] is adequate for the execution of queries in our mediator since it supports path expressions featuring variables on both classes (i.e. nodes) and properties (i.e. edges) of the CIDOC CRM paths.

The RQL-like syntax provides a *select-from-where* set of clauses to construct queries and introduce variables. In the *select* clause, the variables to be answered are inserted. In the *from* clause, data path expressions are used based on the triple syntax of CIDOC CRM paths (*class* → *property* → *class*). For every triple, in the *from* clause we use a data path expression with the domain and range classes of the triple, their data variables which are introduced respectively to the classes and the property of the triple. The reuse of a particular variable in more than one data path expressions introduces joins between the triples. For data filtering, our RQL-like syntax relies on the *where* clause for string pattern matching.

**Query 1:** *Find the title of the archive.*

**Corresponding CIDOC CRM path:**

E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P128 carries → E73 Information Object → P102 has title → E35 Title

**Query 1 in RQL-like syntax:**

```
select X5
```

```

from
{X1;E31_Document}P106_is_composed_of{X2;E31_Document},
{X2;E31_Document}P70_documents{X3;E22_Man-Made_Object},
{X3;E22_Man-Made_Object}P128_carries
  {X4;E73_Information_Object},
{X4;E73_Information_Object}P102_has_title{X5;E35_Title}

```

**Query 2:** *Find the creator (corporate name of the originator) of the archive.*

**Corresponding CIDOC CRM path:**

E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P108B was produced by → E12 Production → P14 carried out by → E40 Legal Body

**Query 2 in RQL-like syntax:**

```

select X5 from
{X1;E31_Document}P106_is_composed_of{X2;E31_Document},
{X2;E31_Document}P70_documents{X3;E22_Man-Made_Object},
{X3;E22_Man-Made_Object}P108B_was_produced_by
  {X4;E12_Production},
{X4;E12_Production}P14_carried_out_by{X5;E40_Legal_Body}

```

**Query 3:** *Find the creator (corporate name of the originator) of the series titled "I.U. Library Archives".*

**Corresponding CIDOC CRM path:**

E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P46 is composed of → E22 Man-Made Object (P2 has type → E55 Type="series") (P128 carries → E73 Information Object → P102 has title → E35 Title="I.U. Library Archives") → P108B was produced by → E12 Production → P14 carried out by → E40 Legal Body

**Query 3 in RQL-like syntax:**

```

select X9 from
{X1;E31_Document}P106_is_composed_of{X2;E31_Document},
{X2;E31_Document}P70_documents{X3;E22_Man-Made_Object},
{X3;E22_Man-Made_Object}P46_is_composed_of
  {X4;E22_Man-Made_Object},
{X4;E22_Man-Made_Object}P2_has_type{X5;E55_Type},
{X4;E22_Man-Made_Object}P128_carries
  {X6;E73_Information_Object},
{X6;E73_Information_Object}P102_has_title{X7;E35_Title},
{X4;E22_Man-Made_Object}P108B_was_produced_by
  {X8;E12_Production},
{X8;E12_Production}P14_carried_out_by{X9;E40_Legal_Body}
where X5='series'
where X7='I.U. Library Archives'

```

**Query 4:** *Find the title of the archive which is identified as "ARC.14/2".*

**Corresponding CIDOC CRM path:**

E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object (P1 is identified by → E42 Identifier="ARC.14/2") → P128 carries → E73 Information Object → P102 has title → E35 Title.

**Query 4 in RQL-like syntax:**

```

select X6
from
{X1;E31_Document}P106_is_composed_of{X2;E31_Document},
{X2;E31_Document}P70_documents{X3;E22_Man-Made_Object},
{X3;E22_Man-Made_Object}P1_is_identified_by
  {X4;E42_Identifier},
{X3;E22_Man-Made_Object}P128_carries
  {X5;E73_Information_Object},
{X5;E73_Information_Object}P102_has_title{X6;E35_Title}
where X4='ARC.14/2'

```

## 4. CONCLUSIONS

The creation of mappings as part of our proposed integration scenario requires deep conceptual work by CH metadata specialists. However, the semantic richness of CIDOC CRM provides a stable point of reference for heterogenous data. Our current research work focusses on creating mappings new metadata standards (i.e. VRA, MODS etc) to CIDOC CRM. Besides, we are currently exploring a PROLOG implementation for the execution of queries in CIDOC CRM.

## 5. REFERENCES

- [1] L. Bountouri, C. Papatheodorou, V. Soulikias, and M. Stratis. Metadata Interoperability in Public Sector Information. *Journal of Information Science*, 35(2):204–231, April 2009.
- [2] CIDOC CRM Special Interest Group. Definition of the CIDOC Conceptual Reference Model. Technical report, December 2008.
- [3] I. F. Cruz and H. Xiao. The Role of Ontologies in Data Integration. *Journal of Engineering Intelligent Systems*, 13(4):245–252, 2005.
- [4] I. F. Cruz, H. Xiao, and F. Hsu. An Ontology-Based Framework for XML Semantic Integration. In *Proceedings of the 8th International Database Engineering and Applications Symposium (IDEAS 2004), July 7-9, Coimbra, Portugal*.
- [5] DCMI Usage Board. DCMI Metadata Terms, 2006. <http://dublincore.org/documents/dcmi-terms/>.
- [6] DELOS. Building Core Ontologies: a White Paper of the DELOS Working Group on Ontology Harmonization. White paper, DELOS Network of Excellence on Digital Libraries, 2002.
- [7] M. Doerr. The CIDOC CRM An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, 24:75–92, 2003.
- [8] G. Hodge. Metadata for electronic information resources: from variety to interoperability. *Information Services and Use*, 25(1):35–45, 2005.
- [9] C. Kakali, I. Lourdi, T. Stasinopoulou, L. Bountouri, C. Papatheodorou, M. Doerr, and M. Gergatsoulis. Integrating Dublin Core metadata for cultural heritage collections using ontologies. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2007), Singapore, 27 - 31 August*, pages 128–139, 2007.
- [10] G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis, and M. Scholl. RQL: A Declarative Query Language for RDF. In *Proceedings : Sheraton Waikiki, Honolulu, Hawaii, 7 - 11 May 2002*, 2002.
- [11] H. Kondylakis, M. Doerr, and D. Plexousakis. Mapping Language for Information Integration. Technical Report 385, December 2006.
- [12] G. Koutrika. Heterogeneity in Digital Libraries: Two Sides of the Same Coin, June 2005. Delos Newsletter Issue 3.
- [13] P. Lehti and P. Fankhauser. XML Data Integration with OWL: experiences and challenges. In *Proceedings of the SAINT*, pages 160–170. IEEE Computer Society, 2004.
- [14] L. of Congress. Encoded Archival Description, Version 2002, 2002. <http://www.loc.gov/ead/>.
- [15] L. of Congress. Metadata Object Description Schema (MODS), 2008. <http://www.loc.gov/standards/mods/>.
- [16] I. C. on Archives. *ISAD(G): General International Standard Archival Description*. International Council on Archives, Ottawa, 2000.
- [17] T. Stasinopoulou, L. Bountouri, C. Kakali, I. Lourdi, C. Papatheodorou, M. Doerr, and M. Gergatsoulis. Ontology-Based Metadata Integration in the Cultural Heritage Domain. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, volume 4822 of *LNCS*, pages 165–175. Springer Berlin/Heidelberg, 2007.
- [18] M. Theodoridou, Y. Tzitzikas, M. Doerr, Y. Marketakis, and V. Melessanakis. Modeling and Querying Provenance using CIDOC CRM. Technical Report Draft 0.94, Institute of Computer Science, FORTH-ICS, December 2008.
- [19] W3C. XML Path Language (XPath) 2.0, January. <http://www.w3.org/TR/xpath20/>.