Ontology-Based Metadata Integration in the Cultural Heritage Domain*

Thomais Stasinopoulou[1], Lina Bountouri[1], Constantia Kakali[1], Irene Lourdi[1], Christos Papatheodorou[1], Martin Doerr[2], and Manolis Gergatsoulis[1]

[1] Department of Archive and Library Sciences, Ionian University, Palea Anaktora, Plateia Eleftherias, 49100 Corfu, Greece
{boudouri,papatheodor,manolis}@ionio.gr, stasthomais@yahoo.gr, nkakal@panteion.gr, elourdi@lib.uoa.gr
[2] Institute of Computer Science Foundation for Research and Technology, Vassilika Vouton P.O.Box 1385 GR 711 10 Heraklion, Crete Greece
martin@ics.forth.gr

Abstract. In this paper, we propose an ontology-based metadata integration methodology for the cultural heritage domain. The proposed real - world approach considers an integration architecture in which CIDOC/CRM ontology acts as a mediating scheme. In this context, we present a mapping methodology from Encoded Archival Description (EAD) and Dublin Core (DC) metadata to CIDOC/CRM, and discuss the faced difficulties.

Keywords: Ontology-based Integration, Metadata interoperability, Cultural Information, CIDOC/CRM, EAD, DC.

1 Introduction

Making cultural resources accessible requires rich metadata structures, able to cover the variety of material held in memory institutions (such as archives, bibliographic and electronic material). Nowadays, it is common to find metadata sources, which may differ in various aspects, even if the resources they describe originate from the same application domain. This phenomenon is especially observed for the cultural resources, for which several metadata standards have been developed in order to cover the documentation needs and the peculiarities of every type of material.

Taking into account the metadata variety and the increasing demand for targeted global search, unified access and data exchange between heterogeneous

* The research work of the authors I. Lourdi, L. Bountouri, and M. Gergatsoulis was partially co-funded by the European Social Fund (75%) and National Resources (25%) -Operational Program for Educational and Vocational Training (EPEAEK II) and particularly by the Research Program "PYTHAGORAS II". The research work of the authors C. Kakali, Th. Stasinopoulou, C. Papatheodorou and M. Doerr was supported by DELOS Network of Excellence on Digital Libraries funded by European Commission.

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 165–175, 2007.

cultural sources, emphasis is given to matters of interoperability and integration at various levels, such as syntactic, schematic, system and also at the more complex semantic level. Data Integration has been a dynamic and challenging research area for many years. However, nowadays, research interests are moving from Data Integration to Semantic Integration in many communities and disciplines, such as e-government and cultural heritage. This movement is being seriously influenced by the new notion and philosophy that the Internet tends to acquire, the Semantic Web. Semantic Integration is the process of using a conceptual representation of the data and of their relationships to eliminate possible heterogeneities [4]. One of the main Semantic Web infrastructure elements, which are an important means in semantic integration scenarios, are ontologies. Their nature allows the sophisticated, extended and rich expression of meanings, and - at the same time - the ability of reasoning.

In this context, ontologies can be considered as an important building block for integration architectures [9], into which metadata originating from diverse sources can be semantically mapped and integrated [5,13]. They are preferred in comparison to other schemas, because of their ability to conceptualize particular domains of interest and express their rich semantics.

In our approach, we use CIDOC/CRM [2] ontology as a conceptual representation of cultural heritage domain to promote semantic integration between different metadata schemas, such as Encoded Archival Description [14] and Dublin Core [18], and eliminate their possible semantic heterogeneities. We address the problems that arise when creating semantic mappings from metadata schemas to ontological models, with the intention of achieving semantic interoperability. We document the use of those mappings in an architecture integrating different cultural metadata sources. We also present the methodology followed to create the mappings and we give an overview of the EAD elements to CIDOC mapping, extending the mappings presented in [17]. EAD is the most well known standard for archival description. It is an XML-based descriptive schema, intended to create electronic finding aids, which include the necessary information for the identification, management and interpretation of an archive. EAD has been used in order to provide archival metadata in digital libraries.

2 Problem Definition and Related Work

2.1 Mapping Metadata to Ontologies

Mapping metadata schemas to ontologies is a complicated procedure, once those two forms have many differences between them in various levels.

Scope and Function: Metadata have a completely different scope and function in comparison with ontologies. Metadata are used to describe, identify, facilitate the access, usage and management of (digital) resources. Ontologies define entities in a more abstract level, with the intention of conceptualizing a domain of interest. They do not provide specific elements for the description of a resource,

but a general definition of the basic notions of a field and the relations between them.

Expression of Semantics: Metadata schemas are created for resources' identification and description and - most of the times - they do not express rich semantics. Even though the meaning of the metadata information can be processed by humans and its relationship to the described resource can be understood, for machine processing the actual relationships are frequently not obvious. In contrast to metadata schemas, ontologies provide rich constructs to express the meaning of data. For example, in DC we write that "a specific poet is the creator of a poem" by assigning a value to DC.creator. On the contrary, in CIDOC we can express general statements about the creation of poems denoting that an Actor (poet) participates in a Creation Event which produces a Linguistic Object (poem). In this way, the knowledge concerning the poem creation becomes explicit and machine "understandable"[8].

Moreover a plethora of conceptual expressions should be aligned for mapping a metadata schema to an ontology. For example, EAD carries two main semantic structures: (a) the metadata of a finding aid and (b) the encoding of a finding aid itself. On the other side, the combination of the CIDOC entities and properties generates a large number of conceptual expressions that should be studied in order to select the semantically closest of them to map the metadata elements.

2.2 Related Work

Works related to ontology-based integration, usually emphasize on element and structure level mappings and transformations (i.e. elements to classes, attributes to properties etc.). In [5], authors map the XML data of every local source to an RDFS local ontology, created by transforming the XML elements and attributes to RDFS classes and properties. An additional characteristic of their method is that they preserve the structure of an XML local source inside the local RDFS ontology. Then, local ontologies are merged to a global ontology for unified access and semantic integration of local data sources. In [13], an XML data integration approach is presented based on the Web Ontology Language (OWL). More specifically, the proposed architecture maps XML structures (such as elements and attributes) to OWL structural components (such classes, properties, etc.) and thus they convert the XML data to an OWL global ontology. In order to define mappings, mapping languages have been proposed (see [11] for example).

In [1] the intention of the work is to propose a mechanism for the cultural information sources integration. The authors map pieces of information contained in XML fragments to domain specific ontologies, such as CIDOC, defining (1) a mapping language that describes the resources by a set of rules relating XPath location paths to the concepts and roles of an ontology and (2) a query rewriting algorithm for translating user queries into queries expressed in an XML query language, which are send for evaluation to XML sources.

Even though these approaches define semantic integration formulas, they are strongly oriented to integrate XML data to RDFS and OWL ontology languages, giving emphasis to define structure mappings or model mappings between them.

However, their effectiveness in mapping really complex semantically data structures, such as metadata schemas, has not yet been tested.

3 An Ontology-Based Mediator

In our approach, we focus on the need to develop information systems able to provide access to heterogeneous data sources. We consider the existence of various cultural sources described with different metadata schemas and there is the demand that our users retrieve information from them. For this purpose, we propose to employ a mediator able to semantically integrate the various schemas. Specifically, we consider CIDOC ontology as the global schema and we define mappings from the metadata schemas to CIDOC and vice versa.

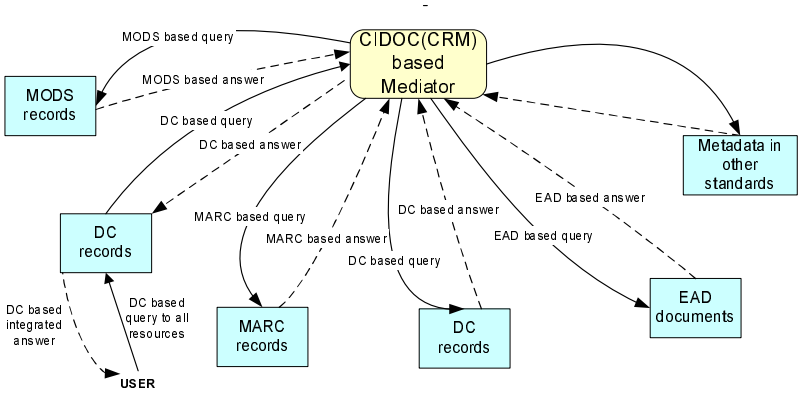

Fig. 1. An Ontology-Based Mediator

We selected CIDOC as the mediating ontology because it is a core ontology designed to be applied for the documentation, integration, mediation and exchange of heterogeneous cultural information. It is a conceptual model, composed of *entities*, which are organized into a hierarchy and semantically related to each other with *properties*. In detail, the CIDOC defines the complex interrelationships that exist between objects, actors, events, places and other concepts in the cultural heritage field [10].

According to [3] the value of CIDOC/CRM becomes apparent when it is used as the basis for data transfer and exchange between different systems, schemas and semantics. In such a scenario, CIDOC acts as a mediated schema to which different metadata can be mapped. Given that it is a core ontology, it allows gathering all necessary cultural information in a suitable form for further reasoning [7]. Figure 1 presents an architecture in which a set of data sources exists, each of them following a possibly different metadata schema. All these schemas are mapped to CIDOC. Users can pose their queries to a local data source following the restrictions of the local metadata schema. The local query

engine returns the results from its source and 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 from each source are collected and returned to the user. Note that the queries in the DC sources might be written in a query language such as SPARQL while the queries in EAD sources might be in XQUERY.

For example, suppose that users wish to find metadata records (archival finding aids, Dublin Core records, etc.) describing documents published by a person whose name is "John Smith". In terms of Dublin Core (DC), the author is looking for records for which DC.publisher="John Smith" and DC.type="text". Suppose that users pose their (appropriately formed) query to a DC local data source. Then, the DC records from the local source matching the query are returned to the users. The query is then propagated to the mediator and transformed, using a set of mapping rules from DC to CIDOC [12], into an equivalent query in terms of CIDOC/CRM. In this query, the conditions (corresponding to the conditions of the initial query) locate the values that should be checked through appropriately formed CIDOC paths such as[1]: E33(Linguistic Object)-*P94 (has created/was created)*-E65 (Creation Event)-*P14(performed)[with subproperty P14.1 (in the role of)-E55(Type)="Publisher"]*-E39(Actor)-*P131(is identified by/identifies)*-E82(Actor Appellation)= "John Smith". The CIDOC query is then transformed to other formats and propagated to the corresponding sources. For example, a local source keeping EAD data receives a query whose condition applies on the values returned by the path /ead/eadheader/filedesc/publicationstmt/publisher/name evaluated on the EAD data. This condition compares the returned value with the string value "John Smith". If they match, the whole finding aid is returned to the mediator and then to the users (through the local client) after being transformed into DC format.

4 Mapping Metadata to CIDOC

In this section we introduce the basic methodology steps followed in order to define the mappings between metadata schemas and the CIDOC ontology. Additionally, given that mappings from different metadata schemas (DC, EAD etc.) to CIDOC are required in order to develop the ontology-based mediator described in 3, we define mappings from Dublin Core and EAD to CIDOC. Due to space reasons, we only give an overview of part of the EAD mapping. For a complete reference of DC and EAD mapping to CIDOC see [16,12].

4.1 Methodology

Path-Oriented Approach: In our methodology, 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]. For metadata to CIDOC mapping, we interpret the metadata paths to semantic equivalent CIDOC paths. We define a

[1] The notation Enn, *Pnn* corresponds to CIDOC entities and properties respectively.

CIDOC path as a chain of the form entity-property-entity, such that the entities associated by a property correspond to the property's domain and range. For example, a CIDOC path is E33(Linguistic Object)-*P94 (has created)*-E65 (Creation Event)-*P14(performed)*-E39(Actor) denoting that an author (Actor, E39) during a Creation Event (E65) generated a poem (Linguistic Object, E33). A metadata path is defined as a sequence of elements, subelements (or element refinements), encoding schemes and vocabulary terms, starting from the metadata schema root element separated by the slash symbol (/). For instance, the path /ead/eadheader/filedesc/titlestmt/author/name, is a part of the metadata of an archival description encoded in EAD and denotes the name of the author of an archival description. Moreover the path /DC/DC.Date/DC.Date.Created denotes the creation date of a resource.

It is worth mentioning that this approach is appropriate since both the metadata and the ontology participating in the integration scenario encode information via paths. An indicative example that confirms the need to follow the "path" approach is that most of the metadata schemas provide elements which, even if they have the same element name, depending on their path position declare different semantics. For instance, the EAD element <corpname> declares the organization responsible for the creation, accumulation, or assembly of the described materials before their incorporation into an archival repository, when included in the path /ead/archdesc/did/originator/corpname, and the institution or agency responsible for providing intellectual access to the archival materials being described, when included in the path /ead/archdesc/did/repository/corpname.

Event-Orientation: An issue we are facing while mapping metadata schemas to CIDOC core ontology is the event orientation of the specific ontology. Metadata, such as EAD and DC, are data structures oriented to describe objects and, as most of the metadata schemas, focus on the described object. On the other hand, CIDOC is event based. Its main notions are the temporal entities and events, and the presence of CIDOC entities, such as Actors, Dates, Places, Objects, etc. implies their participation to an event or an activity [7]. For instance, in order to map the DC.Creator element of a physical object in CIDOC terms, we should map this element to the CIDOC entity Actor (E39). However, in CIDOC persons perform particular roles through events, such as Period (E4), Event (E5) and any other incident valid for a certain time. As a result, the entity Actor (E39) could be interlinked to the described object only with the intermediation of an event or an activity, which indicates that the event - taking place in a particular date - resulted in the described object.

Wrapper Elements: An additional issue evoked during the mapping is that there exist metadata schemas, such as EAD, TEI and MODS, composed of many wrapper elements, which group relative information. For example, EAD Header and Archival Description - which are two basic elements of EAD - are wrapper elements. However, most wrapper elements do not have any semantics by their own, but are used to group the elements that belong to them. In fact their semantics are expressed through the semantics of the elements that they contain.

For example, the wrapper element Descriptive Identification (<did>) contains all the elements that provide the basic identification information for an archive, such as the originator (<originator>), the title (<unittitle>) and the physical location (<physloc>) of the described archive. Therefore in our approach we do not define any mappings for the "semantic - free" wrapper elements. Similarly, we do not map any formatting elements, such as tables, list of items etc.

4.2 Mapping EAD to CIDOC

EAD schema is composed of three basic elements:

- The EAD Header (<eadheader>), which is a mandatory element that includes information about the EAD finding aid (i.e. includes the metadata for the archival description and not the archival description itself).
- The Front Matter (<frontmatter>), which is an optional element that contains publication information, such as the title page information of the printed finding aid etc.
- The Archival Description (<archdesc>), which is a mandatory element that incorporates information about the archival description itself.

In our effort, we define mappings from the EAD Header and the Archival Description elements to CIDOC. We ignore Front Matter since it is extremely rarely used. Furthermore, those two groups of elements are mandatory and they constitute the core descriptive part of a finding aid.

An EAD path is a sequence of EAD elements and subelements, starting from the schema root element <ead> separated by the slash symbol (/). For instance, the path /ead/eadheader/filedesc/titlestmt/author/name, denotes the name of the author of an archival description. Specifically, this path is a part of the metadata of an archival description, since it includes the element <eadheader> and information about the name of the person who created the archival description. Therefore, we have to map the EAD paths to CIDOC paths in a way that satisfies the semantic equivalence taking into account the points mentioned in 4.1.

The Archival Description (<archdesc>) is an element that identifies the archive itself, describing its content and context of creation. From this element, we can derive the following information for an archive: (a) its description, (b) its material substance and (c) the information that it carries. In CIDOC terms this information is mapped to the following classes:

- E31 (Document) and E33 (Linguistic Object), denoting that the Archival Description is a text which describes (documents) an archive.
- E22 (Man-Made Object), declaring that the archive is a physical object created by human activity.
- E73 (Information Object) and E33 (Linguistic Object), since these classes refer to immaterial items that include human memory and do not depend on any specific physical carrier.

Given that the <ead> element is equivalent to the entities (E31 Document) and E33 (Linguistic Object), the corresponding CIDOC path to the EAD path

/ead/archdesc is: {E31 (Document), E33(Linguistic Object)}-*P106 (is composed of/forms part of)*-{E31 (Document), E33(Linguistic Object)}-*P70 (documents/is documented in)*-E22 (Man-Made Object)-*P128 (carries/is carried by)*-{E73 (Information Object), E33 (Linguistic Object)} (see Figure 2).

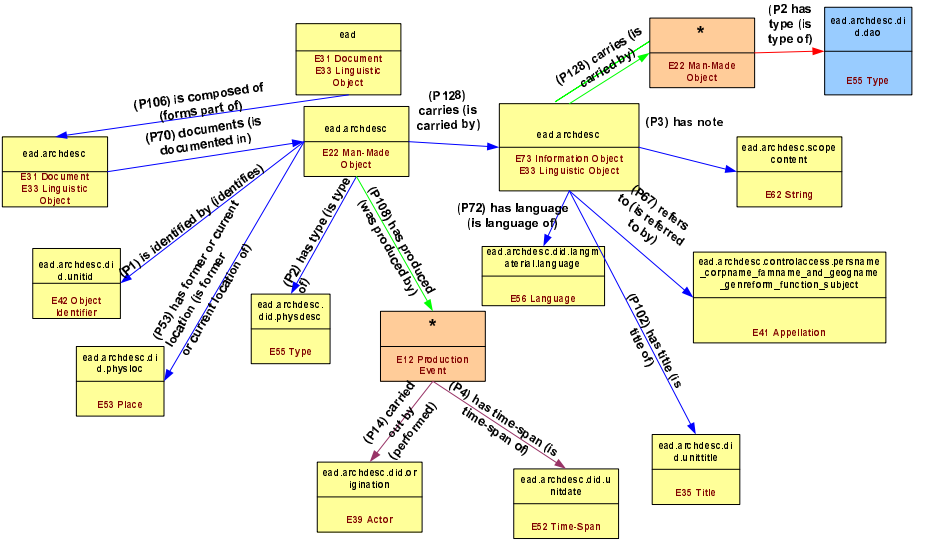

Fig. 2. An indicative part of the Archival Description mapping

According to all the above, 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) and E33 (Linguistic Object), when they provide information that refer to the archive as an informational carrier, or to E31 (Document) and E33 (Linguistic Object), when the provide descriptive information about the archive.

Most of the information contained in the subelements of Descriptive Identification wrapper element (<did>) identifies core information about the described materials and it is linked to the E22 (Man-Made Object) entity dealing with material substance of the archive. For instance consider the path (See Figure 2): /ead/archdesc/did/origination which corresponds to the originator of the archive. The <origination> is an Actor (E39) and we have to link the archive production event with this person. Since CIDOC does not provide a direct link from an object to the person that created it, we consider the archive Production Event (E12) and we link it with the archive using the property P108 (has produced/was produced by). The originator is linked with the production event using the property P14 (carried out/performed). Hence, the CIDOC path is:

{E31 (Document), E33(Linguistic Object)}-*P106 (is composed of/forms part of)*-{E31 (Document), E33(Linguistic Object)}-*P70 (documents/is documented in)*-E22 (Man Made Object)-*P108 (has produced/was produced by)*-E12 (Production Event)-*P14 (carried out by/performed)*-E39 (Actor).

Similar mappings could be generated for the elements Physical Location (<physloc>), Physical Description (<physdesc>), etc. Moreover, there are many EAD elements which are related to the archive as an information carrier. For instance, the element <controlaccess> contains the thematic metadata of an archive. The mapping for the <controlaccess> and its subelements denoting access points, such as /ead/archdesc/controlaccess/persname, to CIDOC is (See Figure 2): {E31 (Document), E33(Linguistic Object)}-*P106 (is composed of/forms part of)*-{E31 (Document), E33(Linguistic Object)}-*P70 (documents/is documented in)*-E22 (Man Made Object)-*P128(carries/is carried by)*-{E73 (Information Object), E33(Linguistic Object)}-*P67 (refers to/is referred to by)*-E41 (Appellation). A similar mapping is followed for the element Title of the Unit (<unititle>), Scope and Content (<scopecontent>), Abstract (<abstract>), etc.

For the development of digital libraries consisting of archival material, elements referring to the digital version of the archive are significant. Those elements are Digital Archival Object (<dao>), Digital Archival Object Group (<daogrp>), Digital Archival Object Location (<daoloc>) and Digital Archival Object Description <daodesc>. All of them are linked with the entity Information Object (E73) since they carry information about the digitized form of the archive. For instance, the element Digital Archival Object (<dao>) provides information about the digital representation of an archive and its components parts (e.g. its URI).

5 Conclusion

Metadata semantic interoperability in the cultural heritage domain is one of the main issues in the digital environment. In our attempt to accomplish that goal, we proposed a semantic integration mechanism, so as to provide unified access to collections of heterogeneous material. In this context, we described an ontology-based integration architecture and addressed the issues of mapping metadata schemas to ontologies. What is more, we presented part of the necessary mappings from cultural heritage metadata to CIDOC mediated ontology.

The mapping definition between the metadata schemas and CIDOC was complex enough. One of the difficulties encountered was the absence of ontology concepts semantically equivalent to metadata fields. In this case, our research team proposed the creation of new classes and properties [12]. An additional issue was the event-based logic that CIDOC implements. Due to that fact, we had to make use of intermediate CIDOC activity and event entities to represent the relationships expressed in metadata between objects (i.e. the archive) and persons (i.e. the creator of the archive). In case of EAD, the mapping difficulties were empowered because of the two different - but related - semantic structures it includes: the metadata of the finding aid (<eadheader>) and the

finding aid itself (<archdesc>). In order to evaluate the handling of the specific difficulties, our future work is to define the inverse mapping from CIDOC to metadata schemas and implement the metadata-CIDOC-metadata query engine.

To conclude, the mapping defined can be encoded using automated tools, such as OWL editors and XML technologies [11]. However, human intervention is necessary in order to define the semantic mapping, given that it is a deep conceptual work.

References

1. Amann, B., Fundulaki, I., Scholl, M., Beeri, C., Vercoustre, A.M.: Mapping XML Fragments to Community Web Ontologies. In: WebDB, pp. 97–102 (2001)
2. CIDOC Documentation Standards Working Group and CIDOC CRM SIG. The CIDOC Conceptual Reference Model, http://cidoc.ics.forth.gr/
3. Crofts, N., Doerr, M., Gill, T.: The CIDOC Conceptual Reference Model: a Standard for Communicating Cultural Contents. In: Cultivate Interactive, vol. (9) (February 2003)
4. Cruz, I.F., Xiao, H.: The Role of Ontologies in Data Integration. Journal of Engineering Intelligent Systems 13(4), 245–252 (2005)
5. Cruz, I.F., Xiao, H., Hsu, F.: An Ontology-Based Framework for XML Semantic Integration. In: Proceedings of the 8th International Database Engineering and Applications Symposium (IDEAS 2004), Coimbra, Portugal, July 7-9 (2004)
6. Doerr, M.: Mapping of the Dublin Core Metadata Element Set to the CIDOC CRM. Technical Report 274 (July 2000)
7. Doerr, M.: The CIDOC CRM An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine 24, 75–92 (2003)
8. Doerr, M., Hunter, J., Lagoze, C.: Towards a Core Ontology for Information Integration. Journal of Digital Information 4(1) (2003)
9. Doerr, M., Lagoze, C., Hunter, J., Baker, T.: Building Core Ontologies: a White Paper of the DELOS Working Group on Ontology Harmonization. White paper, DELOS Network of Excellence on Digital Libraries (2002)
10. ISO. Information and documentation: A reference ontology for the interchange of cultural heritage information, ISO 21127 (2006)
11. Kondylakis, H., Doerr, M., Plexousakis, D.: Mapping Language for Information Integration. Technical Report 385 (December 2006)
12. Kakali, C., Lourdi, I., Stasinopoulou, T., Bountouri, L., Papatheodorou, C., Doerr, M., Gergatsoulis, M.: Integrating Dublin Core metadata for cultural heritage collections using ontologies. In: Proc. Int'l Conf. on Dublin Core and Metadata Applications (DC 2007), Singapore, 27 - 31 August, pp. 128–139 (2007)
13. Lehti, P., Fankhauser, P.: XML Data Integration with OWL: experiences and challenges. In: Proc. of SAINT 2004, pp. 160–170. IEEE Computer Society Press, Los Alamitos (2004)
14. Library of Congress. Encoded Archival Description (2002), http://www.loc.gov/ead/
15. Partridge, C.: The Role of Ontology in Integrating Semantically Heterogeneous Databases. Technical Report 05/02, LADSEB-CNR (June 2002)
16. Stasinopoulou, T., Doerr, M., Papatheodorou, C., Kakali, K.: WP5 - Task 5.5.: EAD mapping to CIDOC/CRM. Report, DELOS-WP5 - Task 5.5 Ontology-driven Interoperability (2007)

17. Theodoridou, M., Doerr, M.: Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM. Technical Report 289 (June 2001), http://cidoc.ics.forth.gr/docs/ead.rtf
18. DCMI Usage Board. DCMI Metadata Terms (2006), http://dublincore.org/documents/dcmi-terms/
19. DCMI Usage Board. DCMI Type Vocabulary 2006, http://dublincore.org/documents/dcmi-type-vocabulary/
20. Uschold, M., Gruninger, M.: Ontologies: principles, methods and applications. Knowledge Engineering Review 11(2), 93–155 (1996)