Journal of Digital Information, Vol 10, No 6 (2009)

Metadata and Semantics in Digital Object Collections: A Case-Study on CIDOC-CRM and Dublin Core and a Prototype Implementation

Dimitrios A. Koutsomitropoulos, Georgia D. Solomou and Theodore S. Papatheodorou
High Performance Information Systems Laboratory, School of Engineering
Computer Engineering and Informatics Department, University of Patras
Building B, 26500, Patras-Rio, Greece
{kotsomit, solomou, tsp}@hpclab.ceid.upatras.gr

Abstract

Digital collections often foster a large number of digital resources that need to be efficiently managed, described and disseminated. Metadata play a key role in these tasks as they offer the basis upon which more advanced services can be built. However, it is not always the case that such collections' metadata expose explicit or even well-structured semantics. Ways to bridge this "semantic gap" are increasingly being sought, as our review of the current state-of-the-art reveals. Most importantly though, in this paper we comment on two well-known metadata standards, popular in cultural heritage applications, namely CIDOC-CRM and Dublin Core; as diverse their scope may be, we nevertheless show how applications can benefit from a transition to explicit semantic structures in these domains, in a way as painless as possible and conformant to Semantic Web standards. We conclude by presenting a concrete, prototype implementation that serves as a proof-of-concept about the ideas argued for.

1. Introduction

Metadata is structured information that describes the characteristics of resources. They form, without fail, the most suitable means to facilitate resource description, integration, discovery and preservation.

A vast number of digital objects, strongly depending on the employed metadata schema for their efficient characterization and retrieval, are stored in a set of distributed, autonomous, and institution-specific repositories. An important issue emerging from the necessity to connect on-line institutions into larger digital repository networks is the integration of their underlying metadata, without losing or misinterpreting conveyed information. Therefore, uniform access to this huge amount of preserved information is required, something which is further impeded by the structural and semantic heterogeneities of the metadata schemata exploited by the source systems.

This necessity becomes also apparent for the cultural heritage field institutions, like museums, archives and libraries, which face a growing need to integrate their systems with other institutions and larger digital library organizations. In this direction, they encounter a number of problems, mainly concerning interoperability issues and loss of implicit knowledge. As a consequence, digital library networks seek for more elegant integration solutions in favor of some existing and rather weak metadata mapping approaches.

The Semantic Web along with its standard tools and processes – provided in order to describe, manipulate and convey the conceptual meaning of web-based information – naturally serves to satisfy this necessity. Additionally, the implementation of the Semantic Web facilities in popular digital repositories systems is gradually becoming a common reality. At the same time though, the process for transforming the implied metadata knowledge, originating from the digital repositories' content, into explicit Semantic Web descriptions can sometimes be problematic and is not always evident. One of these problems is utilizing expressive models, but also in a way that will keep the computational aspects of the underlying logic low.

Having these points in mind, in this article we present two well-established standards that are used as metadata models for the description of resources by many cultural heritage applications and digital repositories. We review and suggest possible upgrades of these models, which have been implemented by using Semantic Web techniques, and argue that this facilitates knowledge discovery as well as semantic interoperability.

More specifically, we discuss our work done on a global conceptual model or – as better referred in its specification – about a formal ontology specific to the cultural heritage domain, known as the CIDOC Conceptual Reference Model (CIDOC-CRM) (Crofts et al. 2003; Doerr 2003). CRM has been acknowledged as an ISO standard (21127:2006) and its primary goal is to establish metadata interoperability. What makes it considerably interesting for the characterization of digital objects is the fact that CRM doesn't behave as a simple data model but rather as a means able to transform the applications' underlying metadata schema into explicit semantics. This is shown in practice by reviewing implemented attempts to create a more expressive OWL ontology out of its formal specification.

Afterwards, we present similar work done on the Dublin Core (DC) metadata schema, which forms a widely adopted metadata standard. In particular, DC is a model suitable for the efficient description of digital objects in various domains (including the cultural heritage domain). We see how its transformation to a Semantic Web ontology can further extend DC's capabilities and applicability.

Finally, we move on to a brief presentation of a prototype implementation built upon the DSpace digital repository system. The described mechanism takes advantage of the previously mentioned DC ontology and provides several semantic-enabled facilities. DSpace serves as our basis platform, as it's a fairly popular system for managing digital collections and is increasingly being used for cultural heritage content.

The rest of this article is organized as follows: In section 2 we give an overview of how current digital repository systems and frameworks attempt to take care of content semantics and consequent interoperability issues. We especially focus on DSpace, as it constitutes the digital repository mechanism upon which the DC ontology has been applied in practice. Section 3 discusses the necessity to employ ontologies in place of simple metadata schemata for a more efficient representation of semantics, whereas sections 4 and 5 comment on work done about the CIDOC-CRM and the DC model respectively. As a conclusion, section 6 summarizes the basic objectives and proposed solutions, concerning the utilization of metadata by digital repository systems in the cultural heritage domain.

2. Digital Repository Systems for Cultural Heritage

Libraries, archives, museums, and other cultural institutions need to preserve and manage intellectual and cultural heritage in

an interoperable way. Due to this necessity, the last decade has witnessed progress towards powerful repository architectures, carrying robust metadata schemata for the characterization and retrieval of resources. Semantic-enabled techniques, widely adopted by most of these mechanisms, further improve interoperability among cultural heritage institutions that rely on digital repository systems. As a result, such systems deliver machine readable and understandable metadata, which render the digital objects available to the end-users in a more intelligent way.

2.1 Managing Cultural Heritage Metadata

The need to share knowledge and resources in the cultural heritage domain has been tackled by diverse set of projects during the last decade. For example, BRICKS (BRICKS 2007) is a project aiming at designing, developing and maintaining an open user- and service-oriented infrastructure particular to this requirement. Since it has integrated metadata and content from a number of archaeological institutions, BRICKS has attempted to integrate the CIDOC-CRM ontology in its core model. This integration has been accomplished through a mapping scenario applied between the source schemata and the CRM ontology, although a number of inconsistencies had to be resolved, mostly originating from the abstractness of some concepts definitions of the CRM (Nussbaumer and Haslhofer 2007).

SIMILE[1] is a research project focused on enhancing interoperability among several types of digital assets, schemata, ontologies, metadata, and services and deals with arbitrary metadata formats using RDF. Actually, SIMILE's primary goal is to leverage and extend DSpace – a widely adopted digital repository applied by many research, scholar and cultural institutions – thus enhancing the latter's support for arbitrary schemata and metadata, through the application of RDF and Semantic Web techniques.

Additionally, other digital repository systems (e.g., Fedora[2] , EPrints[3]) exploit popular metadata formats for the description of their resources, albeit most of them seem to opt for the DC metadata schema. By upgrading the hosted metadata schemata of these systems to more expressive and powerful ontologies they also become able to manage the semantics of their content efficiently.

2.2 The DSpace Reality

DSpace[4] is a popular mechanism that tries to satisfy the necessity for efficient preservation and management of digital assets. It is an open-source digital repository with one of the most rapidly growing user bases worldwide. It provides a way to manage research materials, scholarly publications, scientific and cultural content as well as any other kind of digital collections in a professionally maintained repository, giving greater visibility and accessibility to its content over time. Moreover, DSpace is further enhanced by supporting several facilities in order to achieve interoperability, like the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Lagoze et al. 2002). The OAI protocol offers a means to expose the repository's metadata, rendering them available to existing service providers in an interoperable way.

DSpace exploits the Dublin Core metadata schema, based on the Dublin Core Libraries Working Group Application Profile (LAP) (DCMI-Libraries Working Group 2004), in order to characterize its content. In particular, a default DSpace installation uses the Qualified DC metadata schema, which can also be extended with additional qualifications and elements. In addition to DC, other metadata schemata can also be imported in this mechanism, thus enhancing its capabilities.

The University of Patras digital repository[5] has been built upon DSpace and serves as an institutional repository responsible for the preservation and distribution of the University's research and educational material. The University of Patras DSpace installation is the first to have deployed the semantic-enabled features of DSpace discussed in section 6.

Besides the University of Patras, other institutions can benefit from such features, especially the cultural heritage ones. For example, the digital repository *Pandektis*[6] constitutes a live DSpace installation which preserves and manages digital collections concerning Greek history and civilization. Supported by the National Documentation Centre of Greece, Pandektis is characterized as a "Digital Thesaurus of Primary Sources for Greek History and Culture". Like Pandektis, the *Doria*[7] digital object management system of the National Library of Finland and the *Digital Archive of Literacy Narratives*[8] in USA are a few examples proving how cultural institutions can easily adopt and utilize DSpace as their underlying digital repository system. The gradual increase of those installations, along with the possibility to upgrade their core schema with more cultural domain-specific and value-added metadata models, leads to the establishment of more powerful knowledge management mechanisms, in place of common digital repository systems.

3. From Metadata to Ontologies

The main requisite when exploiting metadata for the description of digital objects is to render information processable by both humans and machines in ways that promote interoperability. Generally speaking, the integration of information resources is set as a major priority in the World Wide Web, but it still remains an annoying headache, due to the vast explosion of web-based resources and the heterogeneity of their descriptive metadata.

Some significant problems that one has to address in an attempt to facilitate interoperability and legacy resource integration are the following:

- The implicit and frequently inconsistent semantics that the source information bears, along with the obvious difficulty in establishing a global standard, has led to the existence of many different metadata schemata, each usually designed in a way that better corresponds to the special needs of a particular domain of knowledge. In the cultural area alone, dozens of standards and thousands of proprietary metadata and data structures exist, as well as hundred of terminology systems (e.g., DC, VRA, CCO, FRBR, AMICO, CIMI, ARCO, SPECTRUM and many others) (Lin et al. 2008). Some of these schemata are usually implemented as semi-structured models, where possible correlations among their elements are usually lost or become useless when applied in practice. This is for example the case for the fairly popular DC metadata schema, which is often implemented as a flat aggregation of elements, as depicted in Table 1.

Table 1. A sample Dublin Core record.

Element	Content
dc.title	*Querying Distributed RDF Repositories*
dc.title.alternative	*Distributing Querying for RDF Repositories*
dc.contributor.advisor	*Papatheodorou, Theodoros*

dc.contributor.author	*Solomou, Georgia*
dc.subject	*RDF*
dc.subject.alternative	*Distributed Repositories*
dc.description.abstract	*Το RDF (Resource Description Framework), πρότυπο του W3C, είναι ένα μοντέλο δεδομένων για την αναπαράσταση πληροφορίας στον Παγκόσμιο Ιστό και αποτελεί τη θεμελίωση ενός συνόλου τεχνολογιών για τη μοντελοποίηση κατανεμημένης γνώσης στο Σημαντικό Ιστό.*
dc.description.translatedabstract	*RDF (Resource Description Framework), a W3C recommendation, is a data model for representing information in the World Wide Web and constitutes the foundation of many existent technologies for the modeling of distributed knowledge in the Semantic Web.*
dc.contributor.committee	*Papatheodorou, Theodoros*
dc.contributor.committee	*Mpouras, Christos*
dc.contributor.committee	*Likothanasis, Spiridon*

- Another important issue that encumbers further progress towards efficient resource management is the difficulty with respect to perform resourceful mappings and translations among different metadata syntaxes (Baca et al. 2000). The role of such mappings is to indicate the equivalence between concepts in different schemata, especially in the field of semantics, as well as to provide the basic guidelines for implementing a functional correspondence among their structural elements. Unfortunately, the case is not rare in which some elements and valuable data are lost when mapped to another metadata schema, as for example happens when we try to make a conversion from a rich structure to a simpler one: When the target format is more inclusive and has defined elements and sub-elements in greater detail than the source format, the value in a source metadata record may need to be broken down into smaller units, leading to the loss of information (Zeng and Xiao 2001; Zeng and Chan 2006). This is evident for example in the mapping from the more general DC Core's elements to the MARC 21 records' subfields.
- Finally, in the majority of applications, a metadata-based search facility is provided in order for users to browse their content. Consequently, for the retrieval of the desired information, the keywords used in user queries must match with the keywords used in the metadata. Possible correlations between them are left out of consideration. This implies that metadata values themselves are not adequate for allowing *semantic* queries (Buranarach and Spring 2006), i.e., queries that can take advantage of the underlying semantic knowledge, since the semantics they bear are usually only human readable.

A possible solution seems to lie in the establishment of robust metadata schemata, apparently opting for those that bear stronger semantic interpretations. The intended purpose of this approach is to enable more refined descriptions of resources, thus allowing applications to benefit from the Semantic Web facilities.

In particular, by transforming implied metadata knowledge into explicit Semantic Web descriptions – namely ontologies – we manage to further boost applications' search capabilities. Apart from information retrieval, knowledge discovery becomes feasible as well. Search services are no longer based on mere keyword matching. Instead, the efficient exploitation of the deployed semantic relations allows for reasoning and often leads to the extraction of logical conclusions that are not explicitly expressed.

Some well-established standards in the Semantic Web community that could offer considerable help towards metadata integration are mentioned below. These standards tend to form the basic objective for nearly all Semantic Web implementations:

- The Resource Description Framework (RDF) (Klyne and Carroll 2004) serves as a common data model for integrating metadata from various autonomous and heterogeneous data sources. In particular, RDF, along with its semantic extension RDF(S) (Brickley and Guha 2004), can be of great value in representing, unifying and possibly interpreting information hidden in disparate databases, information management systems and portals.
- The Web Ontology Language (OWL) (McGuinness and Harmelen 2004) standardizes the expressiveness levels of the Semantic Web and demonstrates characteristics suitable for its distributed environment. It actually facilitates greater machine interpretability of Web content than that supported by XML, RDF and RDF(S) by providing additional vocabulary along with a formal semantics. Except for the full version of OWL, a very expressive, albeit decidable sublanguage of it, which forms the main requisite for the majority of applications in the Semantic Web domain, is OWL-DL (Bechhofer et al. 2004).

For the digital repository systems and all related infrastructures, responsible for managing thousands of digital objects, the introduction of semantic-intensive models – possibly expressed in the standards mentioned above – may constitute the desired solution for achieving interoperability. Therefore, the adoption of the extremely popular CIDOC-CRM and DC model for such mechanisms, along with their implementations in the aforementioned machine processable formats, could ameliorate the conditions in storing, organizing and retrieving information. The work related to the semantic enrichment of the CRM is briefly presented in section 4, whereas in subsequent section 5 we give an overview about some similar work concerning the DC standard.

4. Towards a Semantic Enrichment of CIDOC-CRM

A detailed overview about CICOC-CRM is given in the corresponding reference document (Crofts et al. 2009), released in March 2009. This specification pertains to the most recent version of CRM (namely 5.0.1.) that describes 86 classes and 137 properties and their inverses. The semantics are given as scope notes in textual form along with explanatory examples, thus providing a better understanding of the intended meaning of the described notions and concepts. A qualitative metaschema of the CRM is depicted in Figure 1.

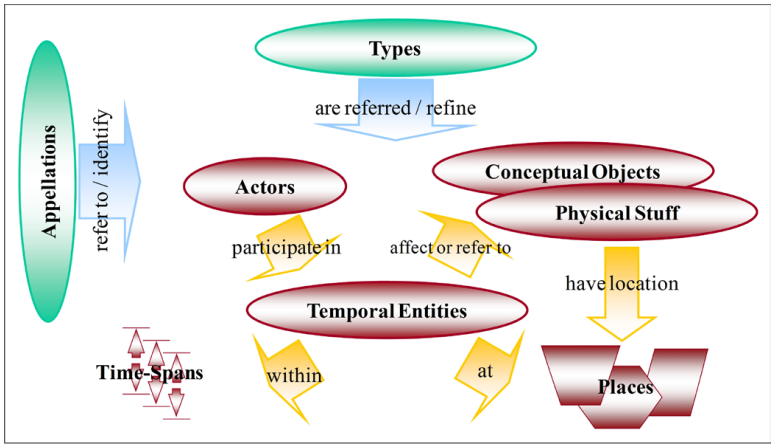


Figure 1. A qualitative metaschema of CIDOC-CRM

A RDF(S) implementation for the current version of CRM is already available[9]. In terms of expressivity, the RDF(S)-enabled structures employed by the CRM can be summarized as follows:

- Classes as well as properties are organized in hierarchies.
- For every property classes are defined, forming its domain and its range.
- For every property, its inverse is also defined, as a separate property, because RDF(S) cannot implicitly express an inversion relation between two properties.
- There is no distinction between object and datatype properties as in OWL; Rather, properties that are equivalent to datatype properties have `rdfs:Literal` as their range.

In the following we give an overview of some efforts that have been proposed in order to take the most out of the CIDOC-CRM by utilizing Semantic Web languages and explicit semantics. Then we show some specific examples that demonstrate what can be gained through such an approach and suggest its usefulness in relevant scenarios.

4.1 CRM in OWL

Based on the existing RDF(S) implementation, an early attempt to upgrade CRM up to OWL level is introduced in (Koutsomitropoulos and Papatheodorou 2007). The main incentive has been to examine the possibilities of applying Semantic Web techniques in order to enable reasoning on and discovery of cultural heritage information over distributed knowledge resources.

According to this method, a manual transformation from the RDF(S) to OWL format took place, following some simple rules:

- Common expressions between these two formats (like `rdfs:subClassOf` and `rdf:resource`) were preserved.
- Some namespace prefixes were replaced by more suitable ones (e.g., `rdfs:Class` was replaced by `owl:Class`).
- Finally, because OWL allows the distinction between object and datatype properties, the more general `rdf:Property` has been selectively replaced by `owl:ObjectProperty` and `owl:DataTypeProperty` respectively.

The proposed method led to the creation of a formal OWL document for the CRM ontology. This document was further augmented with OWL-specific structures (e.g., the notions of inverse properties and unqualified number restrictions) so that the provided ontology has finally reached the OWL-DL level. Some specific OWL-DL constructs that we made use of include the use of *nominals* in value constraints (`hasValue`) and *cardinality restrictions* (`mincardinality`, `maxCardinality`) of arbitrary values. The resulting documents are also available online[10].

The CRM augmented form, was further processed by a web-based tool that employs a reasoning module and serves as an interface for querying the ontology. Consequently, the extraction of new, useful knowledge, not previously expressed in the ontology was possible, through intelligent queries that the end-users were able to pose to the produced OWL document.

Another attempt for upgrading the CRM's semantic level has been presented by the University of Erlangen. Based on the official CRM definition and staying as close as possible to it, an OWL-DL implementation was introduced, known as the Erlangen CRM (Görz et al. 2008). The first specification document about the Erlangen CRM released in January 2008 and conforms to the CIDOC's CRM version 4.2.4. Updated versions have been published since then, coming to the most recent one (Erlangen CRM 2009), which follows the CRM's latest specification 5.0.1.

According to the Erlangen implementation, the CRM ontology is manually encoded into OWL-DL constructs, a process being accomplished through a careful evaluation of the intended meaning of each class and property defined in the specification. The scope notes, as well as the explanatory examples and help hints which augment each concept's definition, played an essential role for the successful completion of this effort.

Nevertheless, some features of the CRM weren't implemented, each for some reasons for which explanations are provided. This is, for example, the case for the so called "types". Types are expressed by a concept (i.e., class *E55 Type*) which is described as a metaclass in CRM. OWL-DL doesn't provide means to represent metaclasses, because their use leads to the loss of decidability. Consequently, an alternative implementation for this particular concept had to be proposed. Erlangen's policy of finding alternative implementations when a direct correspondence was impossible, worked as a solution for many other similar cases.

4.2 Deriving knowledge from the CIDOC-CRM ontology

The representation of CRM in an expressive Semantic Web language yields new possibilities for knowledge acquisition and discovery. In (Koutsomitropoulos et al. 2007) a specific application scenario is introduced that deals with painters and paintings. Using appropriate CRM's classes and relations some facts about the life and work of the Dutch painter Piet Mondrian are represented. Then, through a careful elaboration of the expressed facts in accordance with the OWL allowed constructs, new facts can be inferred.

For example, we state that since something is a "Painting Event" it should only create "Paintings" (in OWL, this is known as a *Value Restriction*). In Figure 2, "Creation of Mondrian's Composition" is an explicitly stated "Painting_Event" that "has_created" the "Mondrian's composition". Now, when asking "what is a painting?" it can be inferred that the "Mondrian's composition" is indeed a painting, correctly interpreting the value restriction on the "has_created" property.

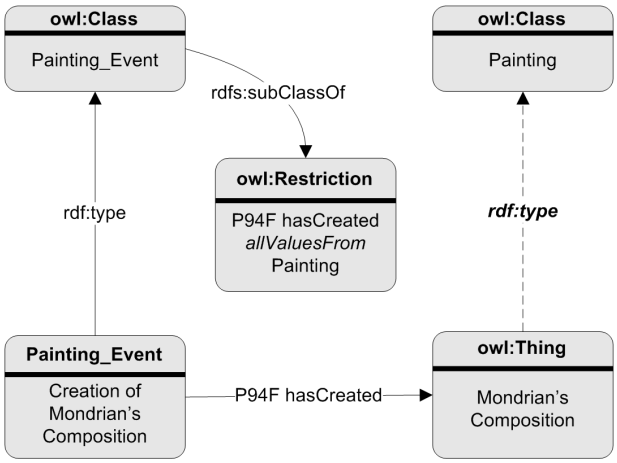


Figure 2. The CIDOC-CRM ontology: An example involving "Paintings" and "Painting Events".

Likewise, similar knowledge discovery scenarios have been suggested in (Lin et al. 2008). There, however, inference paths are being specified by the definition of SWRL rules and not through direct OWL constructs. Such a rule is for example the *hasSameStylePainting*. This rule actually correlates a *Person* with an *Image* based on the idea that a person has same style of painting with the creations of other persons that participate in the same artistic group. As a result, a rich and meaningful set of inferences that relate an artist with other artists' paintings can be generated.

Therefore, the aforementioned proposals for the encoding of CIDOC-CRM in OWL, have managed to render CRM semantically richer and closer to the Semantic Web reality. Digital repository systems, adopting such implementations, can now gain uniform access by formulating queries over a single ontology. Furthermore, these proposals form the basis for future improvements in CRM as well as for a possible OWL 2 (Hitzler et al. 2009) implementation, which is a very promising standard in the Semantic Web community and the main requisite for achieving ontology interoperability and efficient knowledge discovery.

5. Bringing Dublin Core to the Semantic Web

The DC Metadata Element Set (DCMES) (DCMI 2008) forms a simple and concise metadata schema suitable for the description of almost all kind of resources. It has been adopted by many different metadata-enabled applications because it offers a standardized and thus interoperable way to adequately describe and retrieve information. Among others, the great popularity of DCMES is attributed to its general applicability and to the fact that it forms a trade-off between expressivity and size.

Nonetheless, as pointed out in section 3, many applications implement DCMES as a flat aggregation of elements where the sub-element/qualifier relationship is not defined. Moreover, the underlying semantics of the DC metadata domain are usually misapplied, and even if they are represented, they are not always utilized to the maximum possible extent.

In the following we see the efforts towards a conceptual interpretation of the DC model that finally lead to a clear representation of its semantics in machine readable formats. In addition we show how this can be beneficial in real applications, by discussing some indicative examples.

5.1 DC in OWL

An attempt for a more concrete semantic interpretation of the DC model is reflected first in the DCMI Abstract Model (DCAM) specification (Powell et al 2007). DCAM describes an abstract model which tries to capture the inherent semantics of the DC model, hence giving it a more sound conceptual background. Further, the DCMI recommendation for expressing DC in RDF (Nilsson et al. 2008) describes how the features of the DCAM can be represented using the RDF model, virtually suggesting an ontology of DC, expressed in the well-established Semantic Web standard RDF(S).

Nevertheless, a primary requisite for the DC model in the Semantic Web community is to move a step forward and to fully utilize and enrich its semantics by taking advantage of OWL (and OWL 2) expressive strength. Having the existing RDF(S) implementation as a starting point, such an effort has been deployed in (Koutsomitropoulos et al. 2008). The main idea lying behind this work is to explicate and enrich the semantics of DC by using new constructs and refinements, some available only in OWL 2, leaving at the same time the core model intact. Table 2 provides an overview of these additions, by giving an indicative example in RDF/XML syntax along with its anticipated usage.

Table 2. OWL and OWL 2 notions in the DC ontology.

Notion	Example	Usage
Classes	<**owl:Class** rdf:about="#**community**"> <rdfs:subClassOf rdf:resource="&owl;Thing"/> <owl:disjointWith rdf:resource="#item"/> </owl:Class>	*Represent DSpace structural elements.* Here, we represent the DSpace notion of "community" as an OWL class. In addition, this class is declared as being disjoint with the "item" class.
Object Properties	<**owl:ObjectProperty** rdf:about="#**author**"> <**rdfs:subPropertyOf** rdf:resource="&terms;**contributor**"/> </owl:ObjectProperty>	*Represent DC and non-DC relations.* In this example we define the non-DC notion of "author" as being an OWL object property and relate it to the DC model by making it sub-property of dcterms:contributor.
Characteristics of	<owl:ObjectProperty rdf:about="&terms;**relation**">	*Discover relations between individuals.*

Properties	<rdf:type rdf:resource="&owl;**SymmetricProperty**"/> </owl:ObjectProperty> <owl:ObjectProperty rdf:about="&terms;**isPartOf**"> <rdf:type rdf:resource="&owl;**TransitiveProperty**"/> </owl:ObjectProperty>	Here, we define dcterms:relation as a symmetric property and dcterms:isPartOf as a transitive one.
Individual Axioms	<**owl:AllDifferent**> <owl:distinctMembers rdf:parseType="Collection"> <dspace-ont:dspacetype rdf:about="&dspace-ont;**Animation**"/> <dspace-ont:dspacetype rdf:about="&dspace-ont;**Article** "/> <dspace-ont:dspacetype rdf:about="&dspace-ont;**Book**"/> [...] </owl:distinctMembers> </owl:AllDifferent>	*Define characteristics of individuals.* In this example, owl:AllDifferent identifies all possible DSpace types as being different from each other.
Subsumption Hierarchies	<owl:Class rdf:about="#**mimetype**"> <**rdfs:subClassOf** rdf:resource="&terms;**FileFormat**"/> </owl:Class> <owl:Class rdf:about="#**application**"> <**rdfs:subClassOf** rdf:resource="#**mimetype**"/> </owl:Class> <owl:Class rdf:about="#**audio**"> <**rdfs:subClassOf** rdf:resource="#**mimetype**"/> </owl:Class>	*Organize classes in hierarchies.* In this example, the DSpace MIME type vocabulary is modeled as a partition of subclasses with dcterms:FileFormat being the top class.
Punning	<rdf:Description rdf:about="&terms;**Agent**"> <rdf:type rdf:resource="&rdfs;**Class**"/> <rdf:type rdf:resource="&terms;**AgentClass**"/> </rdf:Description>	*States that a name can be treated either as an individual or a class.* Here, dcterms:Agent is both a class on its own and an instance of dcterms:AgentClass as dictated by the DC specification.
Role Chains	<rdfs:subPropertyOf rdf:resource="#**sponsorship**"/> <owl:propertyChain rdf:parseType="Collection"> <owl:ObjectProperty> <owl:inverseOf rdf:resource="#**author**"/> </owl:ObjectProperty> <rdf:Description rdf:about="#sponsorship"/> </owl:propertyChain>	*Combine property expressions in chain-forming axioms.* In this example, we refine sponsorship so that it can be inferred that also the authors of the items (and not only the items) receive sponsorship by the same institution.

Another issue was to retain decidability of the produced ontology, a problem arising from the self-descriptive nature and metamodeling capabilities of the DC schema. To this end, the notion of *punning*, introduced with OWL 2 (Golbreich and Wallace, 2009), is found to be useful (see also Table 2); however, the problem is circumvented by the definition of separate namespaces between new and legacy DC elements.

The whole process has finally led to an OWL-2-specific ontology which spans three separate OWL 2 documents that import each other: *dcterms.rdf*[11] , which is actually the original DC-in-RDF implementation, *dc-ont.owl*[12] , containing DC refinements using OWL constructs, and finally *dspace-ont.owl*[13] which further refines DC, based on its use by DSpace, utilizing OWL and OWL 2 constructs. The class hierarchy of the resulting ontology is depicted in Figure 3.

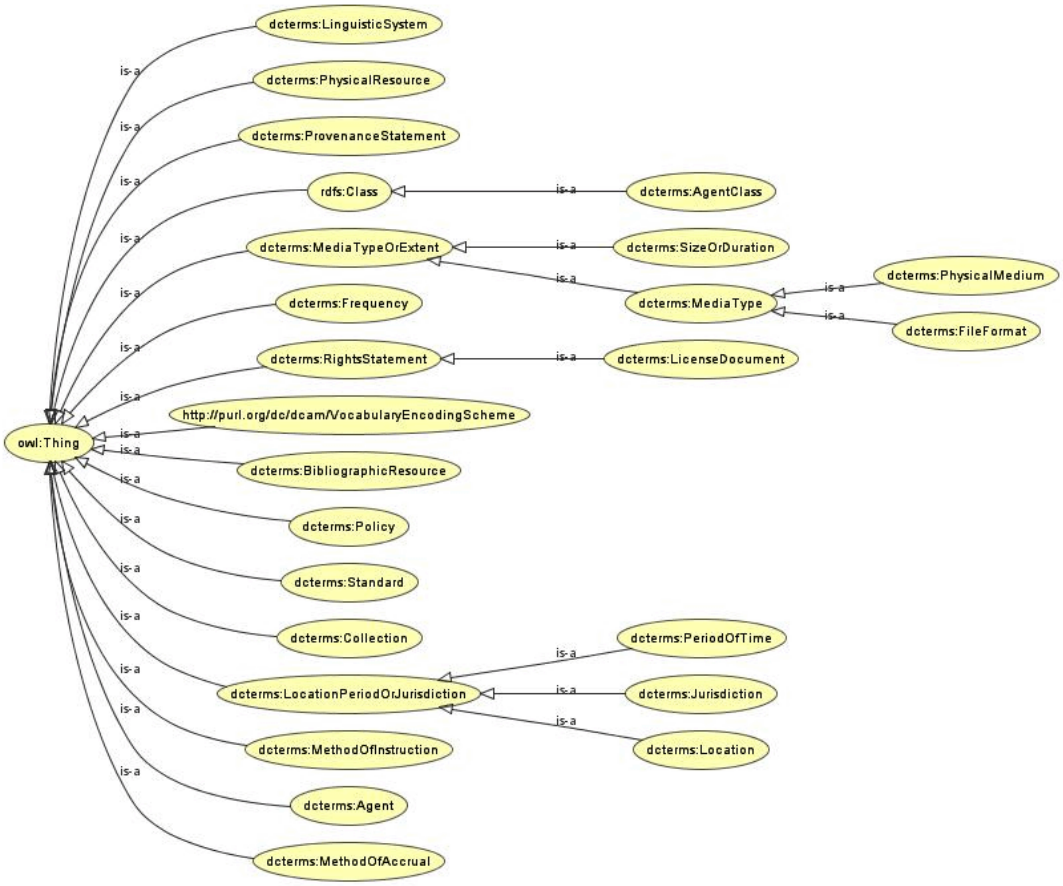


Figure 3. Class hierarchy of the DC ontology, as shown by the OWLViz Tab of Protégé 4.0.

5.2 Semantic gains

Arguably, the most important gain from an OWL ontology of DC is the transition from element-value pairs to subject-predicate-object triples. DC elements, which usually assign a literal value to a resource, are now full-fledged properties that relate compact entities with one another. For example examine the following fragment:

```
<ex:Item rdf:ID="item1">
  <dcterms:type rdf:resource="book"/>
  <dcterms:format rdf:resource="msword"/>
  <dcterms:contributor rdf:resource="authorX"/>
  <dcterms:contributor rdf:resource="authorY"/>
</ex:Item>
```

One can notice that traditional literal values can now be referred to as entities themselves, a fact that allows consistent grouping of resources along semantics axes or facets (e.g., "*Find all books*" or "*Find all items that have an MS Word format*").

Out of the many additional implications this may have (types as classes, language attribution, definition of datatypes) an interesting effect is the ability to specify new, knowledge-bearing axioms that would allow the discovery of hidden relations. In addition to what is summarized in Table 2, a common example is the notion of *co-author*: Using an OWL 2 role-chain, we can express the fact that a contributor of a specific resource, say *item1*, is related to every other entity that *item1* is inversely related with through the `dcterms:contributor` property. If we call this new relation `ex:co-author` then, from the fragment above we can safely deduce that `authorX ex:co-author authorY` and vice-versa (see Figure 4).

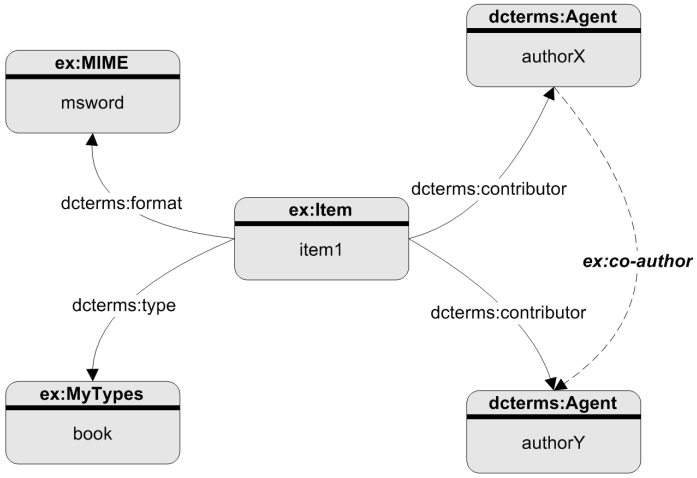


Figure 4. The DC ontology: The "co-author" example.

As a result, this upgrade of DC from a mere metadata schema to a semantically structured ontology appears to contribute to the solution of the metadata integration problem. This updated model can enable more robust search capabilities in the underlying digital collection, apart from the common and fairly weak keyword-based search. This becomes true for systems that have already been equipped with the necessary mechanisms, thus allowing for reasoning-based queries (as for example presented in section 6 about the DSpace system). Consequently, the updated version of DC could offer more intelligent organization, characterization and retrieval of resources.

6. A Prototype Implementation

The OWL 2 version of the DC model can easily apply to a DSpace installation through slight modifications in its programmatic components and in a way that leaves the underlying system's architecture intact. In particular, the proposed technique has been deployed on top of the University of Patras live DSpace installation, in an effort to fully exploit these semantic enhancements, through the implementation of an extensible semantic search and navigation facility.

This facility has been designed so as to be independent of the underlying system architecture, following a "plug-in" philosophy. The two main services, around which it mostly revolves, are the following:

- Semantic-enabled search of DSpace content and
- Semantic navigation amongst the repository's instances.

A simple interface, augmented by the appropriate inference engine, provides access to the semantic search service through which the construction, submission and evaluation of semantic queries becomes possible. This service is actually aimed at those users that are more familiar with Semantic Web technologies. Queries are typed as mere text in the provided text-area field using the Manchester OWL Syntax (Horridge and Patel-Schneider 2009). The latter, in contrast to other, more verbose OWL syntaxes, is simpler to write and read and has replaced mathematical symbols with keywords.

As an example, suppose we want to retrieve all repository items that are of type "Book". The corresponding query in Manchester syntax is expressed as follows:

```
dcterms:type value dspace-ont:Book
```

Since users don't know in advance the contents of the ontology, an auto-complete facility is also provided for their convenience. With this facility matching entities' names, included in the ontology, are suggested to the user, as the query is typed in. Retrieved results are clickable and are displayed in this same interface in the form of a list, further organized in pages (see Figure 5).

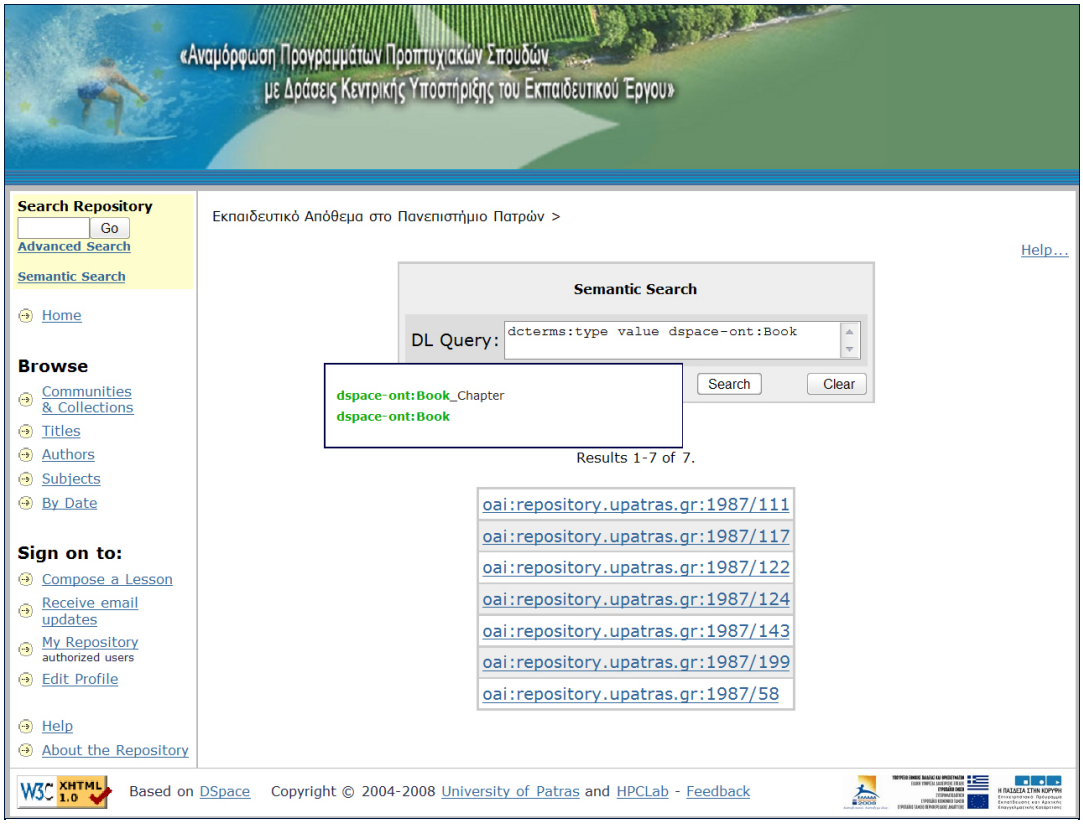
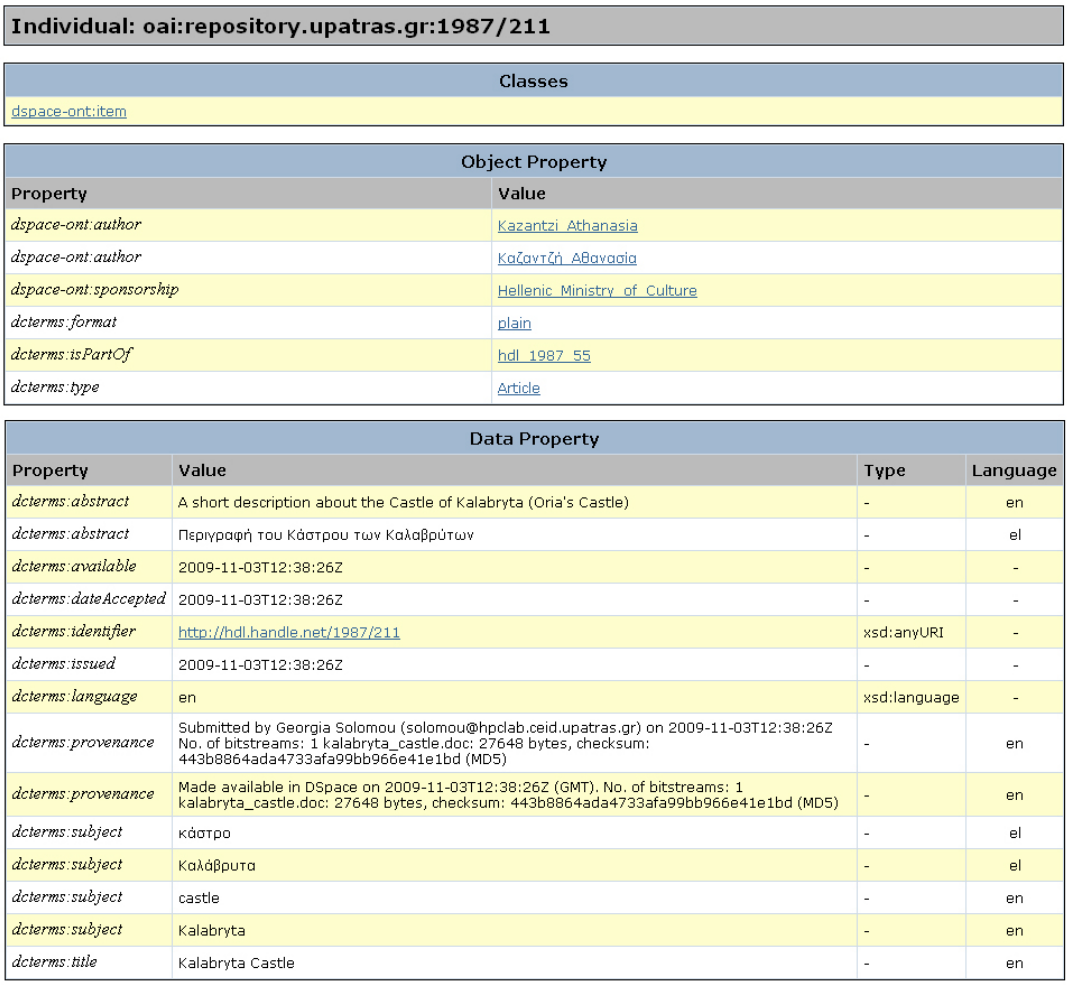


Figure 5. Semantic search interface in DSpace.

Semantic navigation amongst the repository's instances is implemented through the semantic navigation interface, where detailed ontological information about a selected entity (individual) is presented. This interface gives detailed reference to the individual's ontological information, regarding the classes to which it belongs, as well as its properties (i.e., *object*, *datatype* and *annotation* properties) (Figure 6).



Individual: oai:repository.upatras.gr:1987/211

Classes
dspace-ont:item

Object Property	
Property	**Value**
dspace-ont:author	Kazantzi Athanasia
dspace-ont:author	Καζαντζή Αθανασία
dspace-ont:sponsorship	Hellenic Ministry of Culture
dcterms:format	plain
dcterms:isPartOf	hdl 1987 55
dcterms:type	Article

Data Property			
Property	**Value**	**Type**	**Language**
dcterms:abstract	A short description about the Castle of Kalabryta (Oria's Castle)	-	en
dcterms:abstract	Περιγραφή του Κάστρου των Καλαβρύτων	-	el
dcterms:available	2009-11-03T12:38:26Z	-	-
dcterms:dateAccepted	2009-11-03T12:38:26Z	-	-
dcterms:identifier	http://hdl.handle.net/1987/211	xsd:anyURI	-
dcterms:issued	2009-11-03T12:38:26Z	-	-
dcterms:language	en	xsd:language	-
dcterms:provenance	Submitted by Georgia Solomou (solomou@hpclab.ceid.upatras.gr) on 2009-11-03T12:38:26Z No. of bitstreams: 1 kalabryta_castle.doc: 27648 bytes, checksum: 443b8864ada4733afa99bb966e41e1bd (MD5)	-	en
dcterms:provenance	Made available in DSpace on 2009-11-03T12:38:26Z (GMT). No. of bitstreams: 1 kalabryta_castle.doc: 27648 bytes, checksum: 443b8864ada4733afa99bb966e41e1bd (MD5)	-	en
dcterms:subject	κάστρο	-	el
dcterms:subject	Καλάβρυτα	-	el
dcterms:subject	castle	-	en
dcterms:subject	Kalabryta	-	en
dcterms:title	Kalabryta Castle	-	en

Figure 6. Semantic navigation service in DSpace.

The ontology that the aforementioned services act upon is constructed and populated on-the-fly, by automatically transforming the DC metadata. These metadata are consumed through the OAI-PMH interface and the resulting ontology is then silently fed to a reasoner (either FaCT++ or Pellet). In fact, by passing the ontology's URI to the reasoner as an HTTP parameter, we render our implementation totally independent of the specific ontological model.

This "plug-in" architecture is further enhanced by the fact that, in the overall implementation, we avoid any references to DSpace-specific code or any direct access to DSpace's database. In this way we ensure that our services are also system-independent. Moreover, the fact that the repository's metadata are accessed only indirectly, through the supported OAI-PMH interface, further helps maintain interoperability. This means that all OAI-compliant digital repositories, capable of exposing their metadata in this manner, could upgrade their core schema, following a similar process. Apart from DSpace, also Fedora, EPrints and many other digital repository systems offer support for OAI-PMH, meaning that also these mechanisms could benefit from such an approach.

Additionally, it is worth noting that our prototype is not restricted to any specific ontology: other ontologies can be exploited, apart from the initial updated version of the DC model. New ontologies can be loaded by just defining their URI.

A direct consequence is that, just like DC, the OWL version of CIDOC-CRM could also form a base ontology for similar, semantically enabled facilities. By just adapting the implementation process to the specific CRM's constructs and needs, the population of an ontology that is specific to the cultural heritage domain becomes feasible.

Our prototype is described in more detail in (Koutsomitropoulos et al. 2009). Its source code is available online[14].

7. Conclusions

As metadata standards usually tend to bear weak semantic interpretations, possible extensions and modifications to their core schemata seem to be valuable in practice. These semantic enhancements become even more necessary when those models are exploited in digital collections. Digital repositories manipulate vast numbers of objects and utilize diverse metadata schemata. Therefore, they seek not only for expressive representations, but also for ways to integrate these metadata in order to be able to provide uniform access to their stored resources.

Semantic Web is a key aspect to this interoperability issue and can offer a new and challenging dimension in the way information is managed and manipulated by applications. Well-known Semantic Web standards, like RDF and OWL, play an important role for the construction of practical and expressive ontologies. This situation, for example, becomes obvious in the case of two fairly popular models, namely CIDOC-CRM and DC.

The semantic enhancement of both CIDOC-CRM and DC and their upgrade to OWL ontologies, has revealed several gains towards our attempt to achieve knowledge discovery. In the case of CRM we have seen how rich and meaningful inferences can now be produced. Similarly, the semantic upgrade of DC appears to enable the specification of knowledge constructs that allow the discovery of hidden relations.

The created DC ontology, along with an already implemented semantic search and navigation facility, has been applied to a real digital repository system. This process has led to this system's improvement, as it caused its conversion into a more intelligent collection management mechanism. What is more, as the semantic search and navigation facilities are made

independent of the underlying system's architecture or the ontology model, many collection management systems can also benefit from this approach.

References

- Baca, M., Gill, T., Gilliland, A.J., and Woodley, M.S. (2000) "Introduction to metadata: pathway to digital information". http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html
- Bechhofer, S., Harmelen, V.F. , Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P., and Stein, L. (2004) "OWL Web Ontology Language: Reference". W3C Recommendation. http://www.w3.org/TR/owl-ref/
- BRICKS (2007) "Building Resources for Integrated Cultural Knowledge Services". http://www.brickscommunity.org/
- Brickley, D., and Guha, R.V., (eds) (2004) "RDF Vocabulary Description Language 1.0: RDF Schema". W3C Recommendation. http://www.w3.org/TR/rdf-schema/
- Buranarach, M., and Spring, M.B. (2006) "Metadata and Semantics: a Case Study on Semantic Searching in Web System". In *Proceedings of the IFIP International Conference on Research and Practical Issues of Enterprise Information Systems (CONFENIS 2006)*, Austria.
- Crofts, N., Doerr, M., and Gill, T. (2003) "The CIDOC Conceptual Reference Model: A standard for communicating cultural contents". *Cultivate Interactive*, Issue 9. http://www.cultivate-int.org/issue9/chios/
- Crofts, N., Doerr, M., Gill, T., Stead, S., and Stiff, M. (editors) (2009) "Definition of the CIDOC Conceptual Reference Model". http://cidoc.ics.forth.gr/docs/cidoc_crm_version_5.0.1_Mar09.pdf
- DCMI-Libraries Working Group (2004). Library Application Profile. DCMI Working Draft. http://dublincore.org/documents/2004/09/10/library-application-profile/
- DCMI (2008) "DCMI Metadata Terms. Dublin Core Metadata Element Set, Version 1.1". DCMI Recommendation. http://dublincore.org/documents/dces/
- Doerr, M. (2003) "The CIDOC conceptual reference model: an ontological approach to semantic interoperability of metadata". *AI Magazine*, Vol. 24, No. 3, 75-92
- Erlangen CRM Version 2009-03-30, based on CIDOC-CRM 5.0.1 (2009) http://www8.informatik.uni-erlangen.de/IMMD8/Services/cidoc-crm/erlangen-crm_090330_5_0_1.owl#
- Golbreich, C., amd Wallace, E. (2009). "OWL 2 Web Ontology Language: New Features and Rationale". W3C Recommendation. http://www.w3.org/TR/owl2-new-features/
- Görz, G., Schiemann, B., and Oischinger, M. (2008) "An implementation of the CIDOC Conceptual Reference Model (4.2.4) in OWL-DL". In *Proceedings of CIDOC 2008*, Greece. http://www8.informatik.uni-erlangen.de/IMMD8/staff/Goerz/crm_owl_cidoc2008.pdf
- Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P., and Rudolph, S. (2009). "OWL 2 Web Ontology Language: Primer". W3C Recommendation. http://www.w3.org/TR/owl2-primer/
- Horridge, M., and Patel-Schneider, P. F. (2009) "OWL 2 Web Ontology Language: Manchester Syntax". W3C Working Group Note. http://www.w3.org/TR/owl2-manchester-syntax/
- Klyne, G., and Carroll, J. J., (eds) (2004) "Resource Description Framework (RDF):Concepts and Abstract Syntax". W3C Recommendation. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
- Koutsomitropoulos, D., and Papatheodorou, T. (2007) "Expressive Reasoning about Cultural Heritage Knowledge Using Web Ontologies" In *Proceedings of 3d International Conference on Web Information Systems and Technologies (WEBIST 2007)*. WIA track, pp.276-281.
- Koutsomitropoulos, D., Paloukis, G., and Papatheodorou, T. (2007) "From Metadata Application Profiles to Semantic Profiling: Ontology Refinement and Profiling to Strengthen Inference-based Queries on the Semantic Web. *International Journal on Metadata, Semantics and Ontologies*, Vol. 2, No. 4, 268-280
- Koutsomitropoulos, D., Solomou, G., and Papatheodorou, T. (2008) "Semantic Interoperability of Dublin Core Metadata in Digital Repositories". In *Proceedings of 5th International Conference on Innovations in Information Technology (Innovations 2008)*, UAE. pp. 233-237.
- Koutsomitropoulos, D., Solomou, G., and Papatheodorou, T. (2009) "Knowledge Management and Acquisition in Digital Repositories: A Semantic Web Perspective". In *Proceedings of the Int. Conference on Knowledge Management and Information Sharing (KMIS 2009)*.
- Lin, C., Hong, J., and Doerr, M. (2008) "Issues in an inference platform for generating deductive knowledge: a case study in cultural heritage digital libraries using the CIDOC CRM". *International Journal on Digital Libraries*, Vol. 8, No. 2, 115-132.
- Lagoze, C., Van de Sompel, H., Nelson, M., and Warner, S. (2002). "The Open Archive Initiative Protocol for Metadata Harvesting". http://www.openarchives.org/OAI/openarchivesprotocol.html
- McGuinness, D., and Harmelen, V.F. (2004) "OWL Web Ontology Language Overview ". W3C Recommendation. http://www.w3.org/TR/owl-features/
- Nilsson, M., Powell, A., Johnston, P., and Naeve, A. (2008) "Expressing Dublin Core metadata using the Resource Description Framework (RDF)". http://dublincore.org/documents/dc-rdf/
- Nussbaumer, P., and Haslhofer B. (2007) "Putting the CIDOC CRM into Practice - Experiences and Challenges". Technical Report, University of Vienna. http://www.cs.univie.ac.at/publication.php?pid=2965
- Powell, A., Nilsson, M., Naeve, A., Johnston, P., and Baker, T. (2007) "DCMI Abstract Model. DCMI Recommendation". http://dublincore.org/documents/abstract-model/
- Zeng, M.L., and Xiao, L. (2001) "Mapping metadata elements of different format". In *Proceedings of E-Libraries 2001*, pp. 91-99
- Zeng, M., and Chan, L. (2006) "Metadata Interoperability and Standardization - A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels". *D-Lib Magazine*, Vol. 12, No. 6. http://dlib.ejournal.ascc.net /dlib/june06/zeng/06zeng.html

Notes

1. http://simile.mit.edu/
2. http://www.fedora.info/
3. http://www.eprints.org/
4. http://www.dspace.org/
5. http://repository.upatras.gr/dspace
6. http://pandektis.ekt.gr/dspace/
7. http://www.nationallibrary.fi/libraries/doria.html
8. http://daln.osu.edu/handle/2374.DALN/1
9. http://cidoc.ics.forth.gr/rdfs/cidoc_crm_v5.0.1.rdfs
10. http://swig.hpclab.ceid.upatras.gr/SWIGroupPapers/CRMinOWL
11. http://repository.upatras.gr/dspace/dc-ont/dcterms.rdf
12. http://repository.upatras.gr/dspace/dc-ont/dc-ont.owl
13. http://repository.upatras.gr/dspace/dc-ont/dspace-ont.owl
14. http://www.dspace.org/add-ons-and-extensions/#semantic