A Systemic Approach for Effective Semantic Access to Cultural Content

Editor(s): Dimitrios A. Koutsomitropoulos, University of Patras, Greece; Eero Hyvönen, Aalto University, Finland; Theodore S. Papatheodorou, University of Patras, Greece
Solicited review(s): Sarantos Kapidakis, Ionian University, Greece; Werner Kuhn, University of Münster, Germany; Rainer Simon, Austrian Institute of Technology, Austria; one anonymous reviewer

Ilianna Kollia [*], Vassilis Tzouvaras, Nasos Drosopoulos and Giorgos Stamou
School of Electrical and Computer Engineering, National Technical University of Athens, Zographou Campus 15780, Athens, Greece

Abstract. A large on-going activity for digitization, dissemination and preservation of cultural heritage is taking place in Europe and the United States, which involves all types of cultural institutions, i.e., galleries, libraries, museums, archives and all types of cultural content. The development of Europeana, as a single point of access to European Cultural Heritage, has probably been the most important result of the activities in the field till now. Semantic interoperability is a key issue in these developments. This paper presents a system that provides content providers and users with the ability to map, in an effective way, their own metadata schemas to common domain standards and the Europeana (ESE, EDM) data models. Based on these mappings, semantic enrichment and query answering techniques are proposed as a means for providing effective access of users to digital cultural heritage. An experimental study is presented involving content from national and thematic content aggregators in Europeana, which illustrates the proposed system capabilities.

Keywords: cultural heritage access, metadata schema mapping, European data model, Europeana, semantic query answering, query rewriting, cultural resource discovery and enrichment

1. Introduction

Digital evolution of the Cultural Heritage Field has grown rapidly in the last few years. Following the early developments at European level and the Lund principles[1], massive digitisation and annotation activities have been taking place all over Europe and the United States. The strong involvement of companies, like Google, and the positive reaction of the European Union have led to a variety of, rather converging, actions towards multimodal and multimedia cultural content generation from all possible sources, such as galleries, libraries, archives, museums and audiovisual archives. The creation and evolution of Europeana, as a unique point of access to European Cultural Heritage, has been one of the major achievements in this procedure. More than 18 million objects, expressing the European cultural richness, are currently accessible through the Europeana portal, with the target pointing to double this number within the next five years.

As a consequence of the above, research in digital cultural heritage (DCH) is rapidly becoming data intensive, in common with the broader humanities, social science, life and physical sciences. Despite the creation of large bodies of digital material through mass digitisation programmes, only a small proportion of all cultural heritage material has been digitised to date. There is significant commitment to further digitisation at national and institutional levels across Europe [28]. An estimate of the vast amount of data (around 77 million books, 358 million photographs, 24 million hours of audiovisual material, 75 million works of art, 10,5 billion pages of archives) still to be digitized and the

[*] Corresponding author. E-mail: ilianna2@mail.ntua.gr
[1] http://www.cordis.europa.eu/pub/ist/docs/digicult/lund

related cost (about 100 billion euro) is provided in the recent European Report of the Comite' des Sages [25]. Further, substantial amounts of born-digital material are related with cultural heritage, such as data produced by scientific research and by digital analysis of cultural objects.

Due to the diversity of content types and of metadata schemas used to annotate the content, semantic interoperability plays a key role that has been identified and treated as a key issue during the last five years[2] [23]. The key in the definition of semantic interoperability is the common automatic interpretation of the meaning of the exchanged information, i.e., the ability to automatically process the information in a machine-understandable manner. The first step for achieving a certain level of common understanding is a representation language that exchanges the formal semantics of the information. Then, systems that understand these semantics, such as reasoning tools, ontology querying engines, can process the information and provide web services like cultural content searching and retrieval. Semantic Web languages and knowledge organization systems, including Resource Description Framework (RDF), Web Ontology Language (OWL), Simple Knowledge Organisation System (SKOS), ontology editing, reasoning and mapping tools [19,21] can be used to achieve this goal.

The main approach to interoperability of cultural content metadata has been the usage of well-known standards in the specific museum, archive and library sectors (Dublin Core, Cidoc-CRM, LIDO, EAD, METS)[3] [24] and their mapping to a common data model used - at the Europeana level: European Semantic Element (ESE, 2008), European Data Model (EDM, 2010) - to provide unified access to the centrally accessed, distributed all over Europe, cultural content [26] . In this framework, research in cultural heritage has to treat collections of data from many heterogeneous data sources as a continuum, overcoming linguistic, institutional, national and sectoral boundaries.[4] Moreover, semantic technologies should provide effective and efficient access to content and answer user queries in an effective, i.e., appropriate and engaging, and efficient, i.e., timely way.

On the other hand, the Web has evolved in recent years, from a global information space of linked documents to one where both documents and data are linked [27]. In this framework, effort is given to aggregating cultural content from different providers, forming unifying models (as in the Europeana case) for achieving semantic interoperability [30]. Moreover, semantic interconnections of content descriptions with rich terminological knowledge published on the web, provide the user with the ability to pose expressive queries in terms of this knowledge. However, the above procedure is not trivial, since the heterogeneity and uniqueness of the cultural content has led to metadata descriptions that differ a lot from a syntactic (based on technologies used for the representation) as well as a semantic (based on the meaning of the information provided) point of view.

The current paper presents a system that includes an ingestion mechanism, which provides users and content providers with the ability to perform, in an effective semi-automatic way, the required mapping of their own metadata schemas to common models, ESE and EDM. Moreover, the system includes a semantic enrichment and query answering part. It is shown that query answering can be used for assisting users to enrich metadata of their content, taking advantage of relevant sources, data and knowledge stores, or to link their data to relevant ones provided by other sources.
It is important to notice that the system is currently used in the framework of many European content aggregation projects (such as Athena, EU-Screen, Carare, Judaica, DCA, Linked Heritage, Europeana v1.0 and Europeana Connect[5] [26,24]) ingesting more than 4 million objects to Europeana until now.

The paper is organised as follows: Section 2 describes the architecture of the proposed system. The content ingestion workflow and the semantic enrichment parts are described in Section 3. Section 4 presents the query answering method, describing the different possible approaches based on the targeted query and ontology properties. An experimental study is presented in Section 5 which illustrates the usage of the proposed system, based on experiments with Hellenic content having been provided to Europeana through the Athena project. Section 6 summarizes the related work, while conclusions and further work are given in Section 7 of the paper.

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

[3] http://www.apenet.eu

[4] See reports of European Commission Member State Expert Group on Digitization and Digital Preservation (MSEG), available at http://ec.europa.eu/information_society/activities/digital_libraries/other_groups/mseg/index_en.htm

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

2. Semantic Cultural Content Access

The current state of the art in Cultural Heritage implements a model whereby many aggregators, content providers and projects feed their content into a national, thematic, or European portal, and this portal is then used by the end user to find cultural items. Typically, the content is described with the aid of standard sets of elements of information about resources (metadata schemas) that try to build an interoperability layer. Europeana is being developed to provide integrated access to digital objects from cultural heritage organisations, encompassing material from museums, libraries, archives and audio-visual archives as the single, direct and multilingual gateway to Europe's cultural heritage. Several cross-domain, vertical or thematic aggregators are being deployed at regional, national and international level in order to reinforce this initiative by collecting and converting metadata about existing and newly digitised resources.

The currently employed Europeana Semantic Elements (ESE) Model is a Dublin Core-based application profile providing a generic set of terms that can be applied to heterogeneous materials thereby providing a baseline to allow contributors to take advantage of their existing rich descriptions. The latter constitute a knowledge base that is constantly growing and evolving, both by newly introduced annotations and digitisation initiatives, as well as through the increased efforts and successful outcomes of the aggregators and the content providing organisations.

The new Europeana Data Model is introduced as a data structure aiming to enable the linking of data and to connect and enrich descriptions in accordance with the Semantic Web developments. Its scope and main strength is the adoption of an open, cross-domain framework in order to accommodate the growing number of rich, community-oriented standards such as LIDO for museums, EAD for archives or METS for libraries. Apart from its ability to support standards of high richness, EDM also enables source aggregation and data enrichment from a range of third party sources while clearly providing the provenance of all information.

Following ongoing efforts to investigate usage of the semantic layer as a means to improve user experience, we are facing the need to provide a more detailed semantic description of cultural content. Semantic description of cultural content, accessible through its metadata, would be of little use, if users were not in position to pose their queries in terms of a rich integrated ontological knowledge. Currently this is performed through a data storage schema, which highly limits the aim of the query. *Semantic query answering* refers to the finding of answers to queries posed by users, based not only on string matching over data that are stored in databases, but also on the implicit meaning that can be found by reasoning based on detailed *domain terminological knowledge*. In this way, content metadata can be terminologically described, semantically connected and used in conjunction with other, useful, possibly complementary content and information, independently published on the web. A semantically integrated cultural heritage knowledge, facilitating access to cultural content is, therefore, achieved. The key is to semantically connect metadata with ontological domain knowledge through appropriate mappings. It is important to notice that the requirement of sophisticated query answering is even more demanding for experienced users (professionals, researchers, educators) in a specific cultural context.

Figure 1 depicts the proposed system architecture. On the left hand side, cultural content providers (museums, libraries, archives) and aggregators wish to make their content visible to Europeana. This is performed by ingesting (usually a subset of) their content metadata descriptions to the Europeana portal. This is a rather difficult task, mainly due to the heterogeneity of the metadata storage schemas (from both technological and conceptual point of view) that need to be transformed to the EDM form. Using the proposed system, the *Metadata Ingestion* module provides users with the ability to map and transform their data to EDM elements through a graphical interface and an associated automatic procedure. The result of this module is an EDM version of the cultural content metadata. Moreover, through the *Semantic Enrichment* module, the translated metadata are represented as RDF triples, in the form of formal assertional knowledge and the Semantic Web principles, and stored in the *Semantic Repository*.

The metadata elements are represented in the semantic repository as descriptions of individuals, i.e., connections of individuals with entities of the *terminological knowledge*. This knowledge is an ontological representation of the EDM (the *EDM Ontology*), that is connected, on the one hand, to *Domain Metadata Standards* (Dublin Core, LIDO, CIDOC CRM etc) sharing terminology with them and providing the general description of ‘Who?’, ‘What?’, ‘When?’ and ‘Where?’ for every digital object and, on the other hand, to more specific terminological axioms providing details

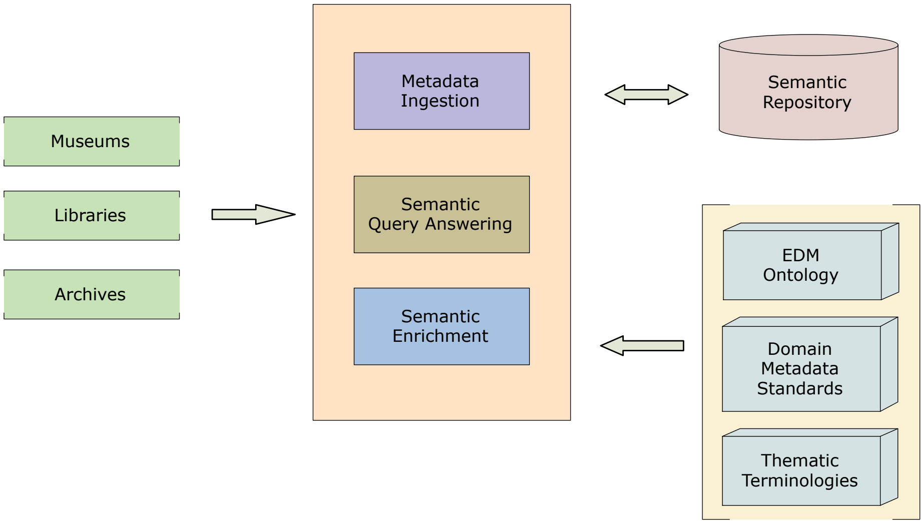


Fig. 1. The architecture of the proposed metadata aggregation and semantic enrichment system

about species, categories, properties, interrelations etc (e.g., brooches are made of copper or gold). The latter knowledge (the *Thematic Ontologies*) is developed by the providers and aggregators and can be used both for semantic enrichment of content metadata, and for reasoning in the *Semantic Query Answering* module. Thus, it provides the user with the ability to build complex queries in terms of the above terminology and access cultural content effectively.

3. Cultural Content Aggregation based on Semantic Mapping

3.1. Metadata Aggregation

The system architecture presented in Figure 1 has been implemented along with an expanding set of web services for metadata aggregation and remediation.[6] It includes ingestion of metadata from multiple sources, semantic mapping of the imported records to a well-defined machine-understandable reference model, transformation and storage of the metadata in a repository, and provision of services that consume, process and remediate these metadata. Although the design was often guided by expediency, the system has been developed using established tools and standards, embodying best practices in order to animate familiar content provider procedures in an intuitive and transparent way. The system has been customized and deployed for several European aggregators that are contributing a substantial amount of Europeana's digital heritage assets. Their diversity has guided the support for various domain metadata models and approaches, mapping cases, and consuming services such as OAI-PMH deployment for harvesting by Europeana or Lucene indexing for portal services.

The key concept behind the aggregation part of the system has been that, although 'low-barrier' standards such as Dublin Core were used in the first stages of Europeana (ESE data model) to reduce the respective effort and cost, a richer and better-defined model could reinforce the domain's conceptualization of metadata records, at least for the mainly descriptive subset of their cataloguing elements. Moreover, since the technological evolution of consuming services for cultural heritage is greater than that of most individual organizations, a richer schema would at least allow harvesting and registering of all annotation data regardless of the current technological state of the repositories or its intended (re)use.

The developed system has been deployed for several standard or specialized models such as LIDO, Dublin Core, ESE, CARARE's MIDAS-based schema, EU-Screen's EBUCore-based approach, and it is being used for the prototyping of EDM. It allows for the ingestion of semi-structured data and offers the ability to intuitively align and take advantage of a well defined,

[6] http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki

machine understandable schema. The underlying data serialization is XML while the user's mapping actions are translated into XSL transformations. The common model functions as an anchor, to which various data providers can be attached and become, at least partly, interoperable. Some of the key functionalities are:

- organization and user level access and role assignment;
- XML collection and record management;
- direct importing and validation according to a standard schema (XSD);
- OAI-PMH harvesting and publishing;
- visual mapping editing for the XSLT language;
- transformating and html previewing;
- repository deployment (XML, RDF).

In this context, the metadata aggregation workflow is illustrated in Figure 2. It consists of five steps. The first is harvesting/delivery, which refers to collection of metadata from content providers through common data delivery protocols, such as OAI-PMH, HTTP and FTP. Second is the Schema Mapping that aligns harvested metadata to the common reference model. A graphical user interface assists content providers in mapping their metadata structures and instances to a rich, well defined schema (e.g. LIDO), using an underlying machine-understandable mapping language. It supports sharing and reuse of metadata crosswalks and establishment of template transformations. The next step is Value Mapping, focusing on the alignment and transformation of a content provider's list of terms to the authority file or external source introduced by the reference model. It provides normalisation of dates, geographical locations or coordinates, country and language information or name writing conventions. Revision/Annotation, being the fourth step, enables the addition of annotations, editing of single or group of items in order to assign metadata not available in the original context and, further transformations and quality control checks (e.g. for URLs) according to the aggregation guidelines and scope. The outcome is metadata aggregation containing and/or publishing all content provider records in the reference and potential harvesting schema(s) (e.g in the case of ESE for Europeana). Finally, the Semantic Enrichment step focuses on the transformation of data to a semantic data model, the extraction and identification of resources and the subsequent deployment of an RDF repository. In the case of EDM, the output of this process is its RDF instances, as is illustrated in the EDM RDF preview of Figure 3. These RDF instances are then mapped to more specific thematic ontologies which define the knowledge that can be used in a particular domain allowing the use of reasoning techniques for the extraction of implicit knowledge. The results of this step are then saved in a semantic repository.

3.2. Mapping Editor

Metadata mapping is a crucial step of the ingestion procedure. It formalizes the notion of 'crosswalk' by hiding technical details and permitting semantic equivalences to emerge as the centrepiece. It involves a graphical, web-based environment where interoperability is achieved by letting users create mappings between input and target elements. User imports are not required to include the adopted XML schema. Moreover, the set of elements that have to be mapped are only those that are populated. As a consequence, the actual work for the user is easier, while avoiding expected inconsistencies between schema declaration and actual usage.

The structure that corresponds to a user's specific import is visualized in the mapping interface as an interactive tree that appears on the left hand side of the editor of Figure 4. The tree represents the snapshot of the XML schema that the user is using as input for the mapping process. The user is able to navigate and access element statistics for the specific import.

The interface provides the user with groups of high level elements that constitute separate semantic entities of the target schema. These are presented on the right hand side as buttons, that are then used to access the set of corresponding sub-elements. This set is visualized on the middle part of the screen as a tree structure of embedded boxes, representing the internal structure of the complex element. The user is able to interact with this structure by clicking to collapse and expand every embedded box that represents an element along with all relevant information (attributes, annotations) defined in the XML schema document. To perform an actual mapping between the input and the target schema, a user has to simply drag a source element and drop it on the respective target in the middle.

The user interface of the mapping editor is schema aware regarding the target data model and enables or restricts certain operations accordingly, based on constraints for elements in the target XSD. For example, when an element can be repeated then an appropriate button appears to indicate and implement its duplication. User's mapping actions are expressed through XSLT stylesheets, i.e. a well-formed XML document

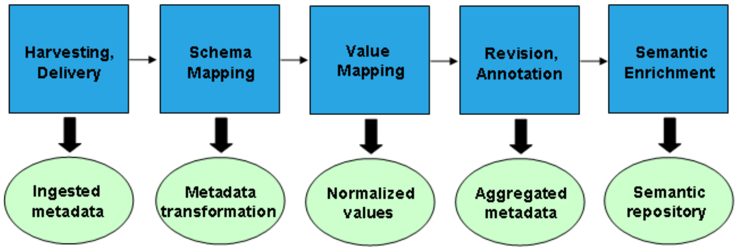
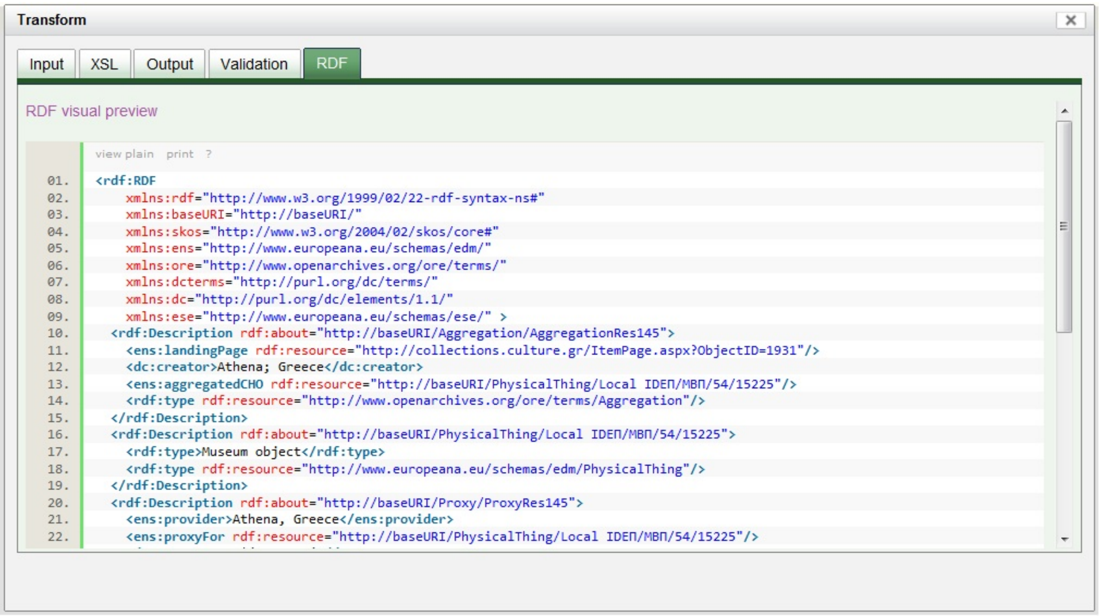


Fig. 2. Metadata Aggregation and Semantic Enrichment Workflow



```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:baseURI="http://baseURI/"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:ens="http://www.europeana.eu/schemas/edm/"
    xmlns:ore="http://www.openarchives.org/ore/terms/"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:ese="http://www.europeana.eu/schemas/ese/" >
  <rdf:Description rdf:about="http://baseURI/Aggregation/AggregationRes145">
    <ens:landingPage rdf:resource="http://collections.culture.gr/ItemPage.aspx?ObjectID=1931"/>
    <dc:creator>Athena; Greece</dc:creator>
    <ens:aggregatedCHO rdf:resource="http://baseURI/PhysicalThing/Local IDEП/MBП/54/15225"/>
    <rdf:type rdf:resource="http://www.openarchives.org/ore/terms/Aggregation"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://baseURI/PhysicalThing/Local IDEП/MBП/54/15225">
    <rdf:type>Museum object</rdf:type>
    <rdf:type rdf:resource="http://www.europeana.eu/schemas/edm/PhysicalThing"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://baseURI/Proxy/ProxyRes145">
    <ens:provider>Athena, Greece</ens:provider>
    <ens:proxyFor rdf:resource="http://baseURI/PhysicalThing/Local IDEП/MBП/54/15225"/>
```

Fig. 3. EDM RDF preview

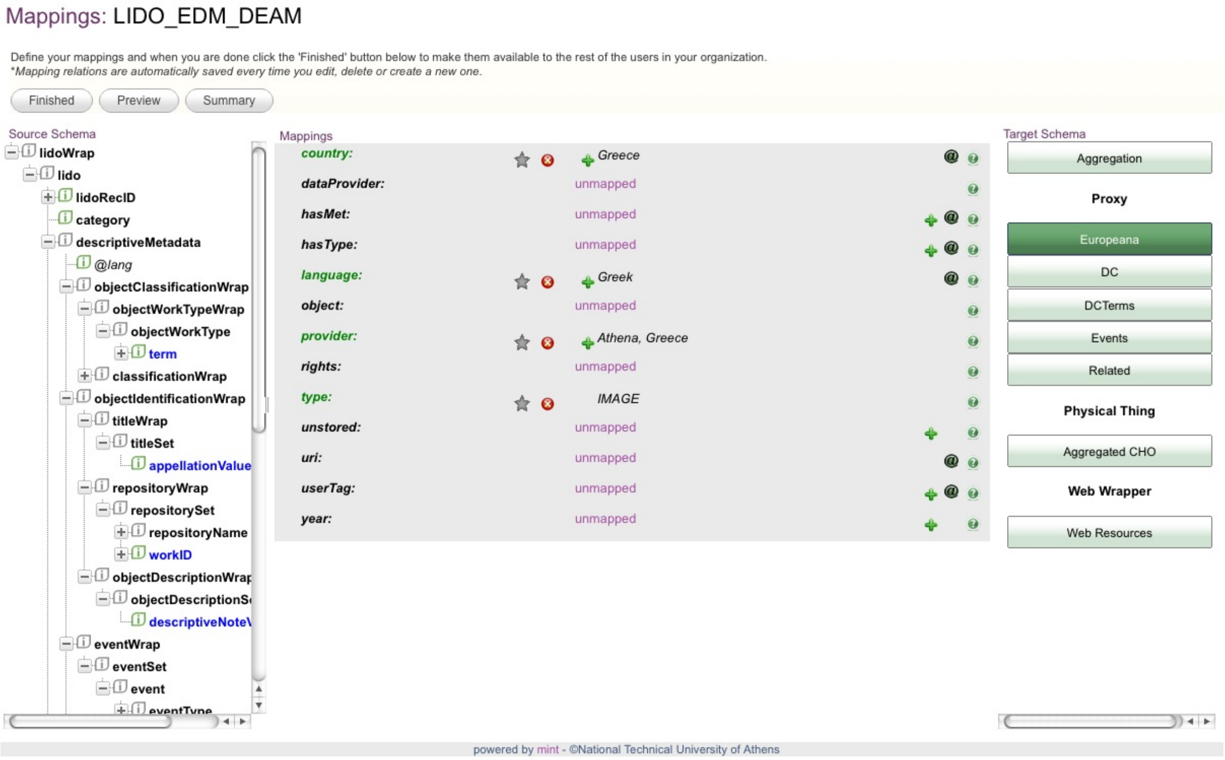
conforming to the namespaces in XML recommendation. XSLT stylesheets are stored and can be applied to any user data, can be exported and published as a well-defined, machine understandable crosswalk and, shared with other users to act as template for their mapping needs. Features of the language that are accessible to the user through actions on the interface include:

- string manipulation functions for input elements;
- 1-n mappings;
- m-1 mappings with the option between concatenation and element repetition;
- structural element mappings;
- constant or controlled value assignment;
- conditional mappings (with a complex condition editor);
- value mappings editor (for input and target element value lists).

3.3. *Semantic Representation*

One of the main points that have guided the system's development is the apparent need for preservation and alignment of as much of the original data richness as possible. The aggregation is only the first effort on the part of providers and aggregators towards the efficient mediation and reuse of their knowledge bases. The support for semantic data models such as EDM enables the repository for deployment and, most importantly, information reuse through knowledge modelling and data interoperability research activities. The aim is to support further resource linking between different collections, reconciliation across the repository and with external authorities and, enrichment of the information resources.

It should be mentioned that it is only due to the achieved metadata aggregation, validated by the con-



Mappings: LIDO_EDM_DEAM

Define your mappings and when you are done click the 'Finished' button below to make them available to the rest of the users in your organization.
*Mapping relations are automatically saved every time you edit, delete or create a new one.

country:	Greece
dataProvider:	unmapped
hasMet:	unmapped
hasType:	unmapped
language:	Greek
object:	unmapped
provider:	Athena, Greece
rights:	unmapped
type:	IMAGE
unstored:	unmapped
uri:	unmapped
userTag:	unmapped
year:	unmapped

Fig. 4. Screenshot of the mapping editor (LIDO to EDM mapping of the Hellenic Ministry of Culture/ Directorate for Archives and Monuments)

tent providers or experts themselves, that semantic enrichment and semantic answering to the queries of the experts and users is possible.

The elements of the EDM ontology are divided into two main categories, namely the elements re-used from other namespaces and the elements introduced by EDM. EDM re-uses from the following namespaces

- The Resource Description Framework (RDF) and the RDF Schema (RDFS) namespaces[7]
- The OAI Object Reuse and Exchange (ORE) namespace[8]
- The Simple Knowledge Organization System (SKOS) namespace[9]
- The Dublin Core namespaces for elements[10] (abbreviated as DC), terms[11] (abbreviated as DCTERMS) and types[12] (abbreviated as DCMITYPE).

The transformation of the data of content providers to RDF (in terms of the EDM ontology) through the schema mapping results in a set of RDF triples that are more like an attribute-value set for each object. Since the EDM ontology is a general ontology referring to metadata descriptions of each object, the usage of thematic ontologies for different domains is necessary in order to add semantically processable information to each object. This includes two steps. First, thematic ontologies are created in collaboration with field experts. These ontologies include individuals which represent the objects, concepts which define sets of objects and roles defining relationships between objects. Then the data values of the attributes of the EDM-RDF instances are transformed to individuals of the thematic ontologies. These individuals are then grouped together to form concepts as imposed by the thematic ontologies. The transformation of the data values to individuals is performed from a technical point of view by mapping the data values to URIs. After this transformation the data are stored in a semantic repository, from where they can be extracted through queries.

[7] http://www.w3.org/TR/rdf-concepts/
[8] http://www.openarchives.org/ore
[9] http://www.w3.org/TR/skos-reference/
[10] http://purl.org/dc/elements/1.1/
[11] http://purl.org/dc/terms/
[12] http://purl.org/dc/dcmitype/

4. Query Answering for Improved Resource Discovery

The result of the ingestion and semantic enrichment described in Section 3 results to a semantic repository containing millions of triples, representing the cultural content descriptions (metadata of the content ingested) in terms of the terminology defined by the EDM Ontology, the Domain Metadata Standards and the Thematic Ontologies (depending on the type of the cultural content). In this section, we present the proposed methodology that we have implemented for providing the user with rich semantic query answering over the above semantic repository.

From a technical point of view, the representation formalism used for the terminological descriptions is OWL 2 (the W3C Standard for Ontology representations on the web) [21] and for the data descriptions is RDF [19]. Actually, most of the terminological axioms do not use the full expressivity of OWL 2, and they can easily fall into the OWL 2 QL Profile [4], that is very useful in query answering. For example, the EDM Ontology is expressed in OWL 2 QL, with only one exception (an axiom that uses disjunction in the definition of the domain of some role). The use of highly expressive OWL 2 DL constructors (like disjunction, nominals, role inclusion axioms etc) is sometimes necessary in thematic ontologies that provide the user with more specific knowledge about species or sorts of cultural assets, as well as their properties and interrelations. However, even in this case most of the terminological knowledge concerns only simple taxonomic axioms, domain and range restrictions and disjoint classes, that can be easily expressed in OWL 2 QL. Concerning the query representation language, we use SPARQL (the W3C query language for RDF) [20], that is supported by most triple stores and is the standard for semantic query answering in the web. Intuitively, the queries supported in our system have the form of conjunctions of atoms that are concepts or roles of the terminologies. The answers are tuples of individuals stored in the semantic repository, satisfying the constraints expressed in the body of the query (are of the type of the specific concepts and are connected with the specific roles).

The theoretical framework underpinning the OWL 2 ontology representation language (as well as the RDF data description that we use in the construction of the semantic repository) is that of *Description Logics* (DL) [2]. Here, we assume that the reader is familiar with the basic notions and foundations of description logics. For the interested user, details can be found in [2,10,4]. Let us now recapitulate the syntax of DLs used throughout the paper.

From a theoretical point of view, we can view the semantic repository and the relevant ontologies as a DL knowledge base (KB) O=$\langle\mathcal{T},\mathcal{A}\rangle$, where $\mathcal{T}$ is the *terminology* (usually called TBox) representing the entities of the domain and $\mathcal{A}$ is the assertional knowledge (usually called ABox) describing the objects of the world in terms of the above entities. Formally, $\mathcal{T}$ is a set of terminological axioms of the form $C_1 \sqsubseteq C_2$ or $R_1 \sqsubseteq R_2$, where C_1, C_2 are $\mathcal{L}$-concept descriptions and R_1, R_2 are $\mathcal{L}$-role descriptions, where $\mathcal{L}$ is a DL language, i.e. a set of concept and role constructors connecting atomic concepts, atomic roles and individuals that are elements of the denumerable, disjoint sets $\mathbf{C}, \mathbf{R}, \mathbf{I}$, respectively. $\mathcal{T}$ describes the restrictions of the modeled domain (in our case the union of the axioms of the EDM ontology, the relevant axioms of the domain metadata standards and the axioms of the thematic ontologies). The ABox $\mathcal{A}$ is a finite set of *assertions* of the form $A(a)$ or $R(a, b)$, where $a, b \in \mathbf{I}$, $A \in \mathbf{C}$ and $R \in \mathbf{R}$. Here, the Abox $\mathcal{A}$ contains the triples of the semantic repository.

The DL language $\mathcal{L}$ underpinning OWL 2 is $\mathcal{SROIQ}$. $\mathcal{SROIQ}$-concept expressivity employs conjunction ($C_1 \sqcap C_2$), disjunction ($C_1 \sqcup C_2$), universal and existential quantification ($\forall R.C$, $\exists R.C$), qualified number restrictions ($\geq R.C$, $\leq R.C$) and nominals ($\{a\}$), while $\mathcal{SROIQ}$-role expressivity allows for the definition of role inverse (R^-) and role compositions ($R_1 \circ R_2$) in the left part of the role inclusion axioms. On the other hand, the OWL 2 QL Profile is based on the DL language DL-Lite$_R$. A DL-Lite$_R$ concept can be either an atomic one or $\exists R.\top$. Negations of DL-Lite$_R$ concepts can be used only in the right part of subsumption axioms. A DL-Lite$_R$ role is either an atomic role $R \in \mathbf{R}$ or its inverse R^-.

The semantics of the above syntax and the definitions of the reasoning problems are standard [2]. Here, we describe only the reasoning problem of *conjunctive query answering* which is the most relevant in our case. A conjunctive query (CQ) q is of the form q : $Q(\vec{x}) \leftarrow \bigwedge_{i=1}^{n} A_i(\vec{x}, \vec{y})$, where $\vec{x}$, $\vec{y}$ are vectors of variables and $A_i(\vec{x}, \vec{y})$ are predicates, either concept or role atoms. The variables in $\vec{x}$ are called *distinguished* or *answer* variables and those in $\vec{y}$ are called *non distinguished* or *existentially quantified*. We say that q is posed over a DL knowledge base $O = \langle\mathcal{T},\mathcal{A}\rangle$ iff all the conjuncts of its body are concept or role names occurring in the ontology. A tuple of individuals $\vec{a}$ is a *certain answer* of a conjunctive query q posed over the DL KB O iff $O \cup q \models Q(\vec{a})$, considering q as a universally quantified implication under the usual first-order logic semantics. The set containing all the answers of the query q over the KB O is denoted with $\mathsf{cert}(q, O)$.

In the literature, it has been proved that the problem of query answering over OWL 2 KBs is difficult, suffering from very high worst-case complexity. The main approach for solving the problem, followed by the majority of triple store systems is to provide approximations based on the *materialisation* method [18], that introduces new triples in the semantic repository by applying the axioms of the terminology to the existing ones. Unfortunately, this approach cannot be effectively followed in OWL 2 DL, nor in OWL 2 QL, although in other clusters of OWL 2 (namely the OWL 2 RL) it has been proved to be really efficient. On the other hand, in OWL 2 QL different methods that are based on query rewriting have been efficiently applied [3,10,14,15], while for the full expressivity of OWL 2 DL, to the best of our knowledge, only approaches that try to reduce query answering to other reasoning problems have been lately implemented [11,12,13,6].

In order to decide which technique is more appropriate for a specific application scenario we need to take into account the benefits and limitations of each one of them. The rewriting approach handles scalability issues well but suffers from the fact that it cannot work with highly expressive languages such as OWL 2 DL which is useful in many practical application scenarios, since in such cases an infinite set of conjunctive or datalog queries can be created. The method that reduces query answering to traditional reasoning services is applicable to very expressive fragments of OWL such as OWL 2 DL but suffers from the fact that it cannot currently handle large amounts of data. Since in our case, we need the full expressivity of OWL 2 (used in the thematic ontologies), keeping in mind that most of the knowledge uses the OWL 2 QL, we propose a hybrid system that uses both rewriting and reduction to entailment checking.

Algorithm 1 summarises the strategy followed for the implementation of semantic query answering. The input of the system is the conjunctive query q, given by the user in SPARQL and the DL Knowledge Base $O = \langle \mathcal{T}, \mathcal{A} \rangle$, i.e., the semantic repository and the relevant knowledge from the EDM Ontology, the Domain Metadata Standards and the Thematic Ontologies. The output of the system is the set of certain answers of q over O, i.e. all the tuples of individuals of the semantic repository (the individuals of the ABox $\mathcal{A}$) that satisfy the restrictions of the query and the terminology $\mathcal{T}$. It is important to notice that, although the volume of the data stored in the semantic repository is huge, we take advantage of two important characteristics of both the data and the relevant terminologies. The first is that most of the terminological axioms can be expressed in DL-Lite$_R$. The second is that the data as well as the terminology have a highly modular form, i.e. they can be partitioned and constitute a set of much smaller independent knowledge bases. This modular character of the knowledge base is mainly a result of the different metadata origination (archives, museums etc) and the respective thematic diversity.

Let us now describe the functionality of the system. After some intialisations, the call of the procedure FindOWLQlTerm($\mathcal{T}$) results to the computation of $\mathcal{T}_{QL}$ that is the maximal subset of the terminology $\mathcal{T}$ containing only DL-Lite$_R$ axioms. Then, with the aid of a rewriting algorithm RewrQA, all the rewritings Q_r of q in terms of $\mathcal{T}_{QL}$ are computed, then executed over the ABox $\mathcal{A}$, with the aid of Execute and the set Ans of correct answers is computed and given to the user. Obviously, Ans is not the complete set if $\mathcal{T} \setminus \mathcal{T}_{QL} \neq \emptyset$, so in this case, we split the knowledge base $\langle \mathcal{T}, \mathcal{A} \rangle$ into a set $\mathcal{K}$ of smaller knowledge bases $\langle \mathcal{T}_i, \mathcal{A}_i \rangle$ (this can be done off-line, before the query answering process) and for each of them we call the query answering engine EntailQA that is based on entailment checking that finally computes all the correct answers.

Algorithm 1 The proposed query answering algorithm

```
procedure QueryAnswering(input CQ $q$, input KB $\langle \mathcal{T}, \mathcal{A} \rangle$, output $\mathsf{Ans}$)
    $\mathsf{Ans} = \emptyset$
    $Q_r = \emptyset$
    $\mathcal{T}_{QL} = $ FindOWLQlTerm($\mathcal{T}$)
    $Q_r \leftarrow Q_r \cup \{$RewrQA$(q, \mathcal{T}_{QL})\}$
    $\mathsf{Ans} \leftarrow \mathsf{Ans} \cup \{$Execute$(Q_r, \mathcal{A})\}$
    $\mathcal{K} = \{$Split$(\langle \mathcal{T}, \mathcal{A} \rangle)\}$
    if $\mathcal{T} \setminus \mathcal{T}_{QL} \neq \emptyset$ then
        for all $\langle \mathcal{T}_i, \mathcal{A}_i \rangle \in \mathcal{K}$ do
            $\mathsf{Ans} \leftarrow \mathsf{Ans} \cup \{$EntailQA$(\langle \mathcal{T}_i, \mathcal{A}_i \rangle)\}$
        end for
    end if
end procedure
```

4.1. Query answering based on query rewriting

Terminologies expressed in the OWL 2 QL Profile are appropriate for splitting the problem of query answering into two parts: the reasoning part which expands the initial query taking into account terminological knowledge provided by the ontology and the data retrieval part which retrieves the instances of the expanded query from the repository. In particular, during the first step (usually called *query rewriting*) the conjunctive query is analysed and expanded into a set of conjunctive queries, using all the constraints provided by the ontology [3,10]. Then, the resulting queries are processed with traditional query answering methods on databases or triple stores, since terminological knowledge is no longer necessary.

Let $q : Q(\vec{x}) \leftarrow \bigwedge_{i=1}^{n} A_i(\vec{x},\vec{y})$ a query posed over the terminology $\mathcal{T}$. A CQ q' is a *rewriting* of q over a TBox $\mathcal{T}$ iff $\mathsf{cert}(q',O) \subseteq \mathsf{cert}(q,O)$, with $O = \langle \mathcal{T},\mathcal{A} \rangle$ and $\mathcal{A}$ any ABox. The set of all rewritings of q over the TBox $\mathcal{T}$ is denoted with $\mathsf{rewr}(q,\mathcal{T})$. It holds that $\mathsf{cert}(q,\langle \mathcal{T},\mathcal{A} \rangle) = \bigcup_{q' \in \mathsf{rewr}(q,\mathcal{T})} \mathsf{cert}(q',\langle \emptyset,\mathcal{A} \rangle)$.

Example 1 *We now show a simple case of query rewriting via an example. Let us assume that a terminology $\mathcal{T}$ consists of the two axioms :*

$$WorkOfArt \sqsubseteq \exists madeBy.Artist \tag{1}$$

$$Painting \sqsubseteq WorkOfArt \tag{2}$$

and we ask the query

$$q : Q(x) \leftarrow madeBy(x,y) \wedge Artist(y) \tag{3}$$

The rewriting of query (3) *w.r.t. $\mathcal{T}$ consists of* (3) *and the following queries :*

$$Q(x) \leftarrow WorkOfArt(x) \tag{4}$$

$$Q(x) \leftarrow Painting(x) \tag{5}$$

Through the decoupling of the data retrieval step from the query rewriting step, users are able to build complex queries without having to know the underlying structure or technical details of the data sources but using only the terminological knowledge expressed in terms of ontologies.

The implementation that we use here is the Rapid system, a goal-oriented rewriting system developed in our Laboratory, which is a prototypical implementation of the query rewriting algorithm presented in [15].

4.2. Reduction of query answering to standard reasoning tasks

The main restriction of the method described in Section 4.1 is that it cannot be applied to terminologies expressed in very expressive clusters of OWL 2 (larger than OWL 2 QL). For these cases, we use the method described in [6] that can be applied to $\mathcal{SROIQ}$ DL KBs. This method follows a different approach translating the query answering problem to the entailment checking one, that has been solved by many reasoners in the literature.

Let $q : Q(\vec{x}) \leftarrow \bigwedge_{i=1}^{n} A_i(\vec{x},\vec{y})$ a query posed over the DL KB $O = \langle \mathcal{T},\mathcal{A} \rangle$. Intuitively, the variables (both the distinguished and the non distinguished) of the query q are substituted by tuples of individuals appearing in the ABox $\mathcal{A}$ forming a boolean query q' and those tuples that result to the entailment of q' by O are kept as the answers for q. More formally, a tuple of individuals $\vec{a}$ is a certain answer of q if there is a vector of individuals (all of which appear in $\mathcal{A}$) $\vec{b}$, such that the entailments $O \models A_i(\vec{a},\vec{b})$, for $i = 1,...,n$ are valid. It should be stated that in this method non distinguished query variables have no existential meaning; they are treated like normal variables (see [6] for more details). To avoid performing m^n entailment checks (where m is the number of individuals in the ontology and n is the number of variables in the query) that would be the result of this process, optimizations can be employed to improve the running time of query answering. Such optimizations for OWL 2 DL are described in [6] in the context of the SPARQL query language. The conjuncts of the query can be evaluated sequentially and variables of subsequent conjuncts are mapped only to individuals that have resulted in the entailment of previous instantiated conjuncts.

Example 2 *Let us assume that we want to evaluate the query:*

$$Q(x,y) \leftarrow WorkOfArt(x) \wedge madeIn(x,y) \wedge Period(y)$$

over an ontology O. Let us also assume that the conjunct $WorkOfArt(x)$ is evaluated first and a set S_{1x} consisting of the individuals that satisfy the conjunct is created. Then the variable x in the second conjunct, $madeIn(x,y)$, is substituted only by the individuals in the set S_{1x} and not by all individuals appearing in O. In the same way, a set S_{1y} containing all individuals for the variable y that satisfy the first two conjuncts is created which contains individuals that can then be *used as possible substitutions for the variable y in the conjunct Period(y).*

Other optimizations refer to the use of more specialized tasks of OWL reasoners such as instance retrieval to retrieve instances of concepts instead of iterating over all individuals of the knowledge base and checking entailment of the instantiated queries obtained by substituting variables with individuals. The use of such methods greatly reduces the running time of queries.

The system that we use has been developed at the Oxford University Computing Laboratory and uses SPARQL as a language to express queries over OWL ontologies and evaluate their answers. SPARQL has currently been extended to find answers to queries under the OWL Direct Semantics Entailment relation [5].

5. Experimental Study and Evaluation of the Proposed System

Application of the proposed system has been taking place in the framework of existing European projects and initiatives. The metadata aggregation part has been largely tested and successfully evaluated in the framework of Europeana. The semantic enrichment and query answering part is to be tested in large scales within the recently started Europeana-related projects 'Linked Heritage' and 'ECLAP', as well as in the new 'Europeana v2.0' best practice network.

The experimental study presented in this section aims at illustrating the performance of the proposed system in the above frameworks. For this reason, it focuses first on the content provided to Europeana through the different projects using the metadata aggregation system described in Sections 2 and 3. Section 5.1 discusses the involvement of content providers and experts in this aggregation and the obtained evaluations. In Section 5.2 we focus on the Hellenic content in Europeana, provided through the Athena project, since it is for this content that we possess thematic knowledge. This knowledge is used to illustrate the obtained semantic enrichment and the performance of the proposed semantic query answering methodology.

5.1. Evaluation of Metadata Aggregation

The metadata aggregator of the proposed system is used and evaluated in seven European E-ContentPlus and ICT-PSP projects (Figure 6). So far, more than four million items have been aggregated to Europeana and six millions are expected to be aggregated in the forthcoming years (based on the content harvesting plan of these projects). 200 cultural organisations have registered in the system. The evaluation approach was based on questionnaires and face-to-face interviews. Evaluation reports have been produced in the form of project deliverables.

For example, in the EUscreen project, the approach to evaluation has been to assess all the available software components, examining user satisfaction with reference to design, functionality, search, navigation, and playing of content. Data feedback was gathered from a disparate set of end users, the public, academic and cultural sector, spread across different countries and languages. For this purpose a questionnaire was sent out to EUscreen consortium and further distributed by each one of the 30 partners to at least five different persons. Moreover, in face-to-face interviews with users, the interviewees were encouraged to provide continuous verbal feedback on how they found the portal. The results of the evaluation were used to improve the usability and functionality of the system. In the Athena project case, the evaluation procedure with more than 100 content providers led to a successful, validated by the content providers, aggregation of large volumes of content metadata.

5.2. Semantic Enrichment and Query Answering

The Greek Cultural Organisations that have provided content to Europeana through the Athena project include the following: the Hellenic Ministry of Culture and Tourism, with their more than 50 Ephorates, the Benaki Museum, the National Documentation Center, the Aegean Historical Archive, the National Research Foundation, the Music Library Lilian Boudouri, the Athens City Museum, the Museum of Cycladic Art, the Historic Research Centre of the Academy of Athens, the Museum of Greek Popular Art, the Hellenic National Gallery, the Marine Museum of Greece, the State Theatre of Northern Greece, the Cultural Foundation of Piraeus Bank Group, the Technical Museum of Ermoupolis, the Press Museum and other organisations aggregated by the University of Patras. This content has been transformed to LIDO (Lightweight Information Describing Objects)[13] XML format. Each of the LIDO records represents a mu-

[13] http://www.lido-schema.org

seum object (proxy instance) and is described among others by an identifier, a type, a description, the material it is made of, the museum where it can be found, the date it was created. All this information is given as data values (strings) of LIDO elements. In particular, this cultural content is classified in 55 categories (such as pottery, jewelry, stamps, wall paintings, engravings, coins) and more than 300 types, within 17 time periods from 35000 b.c. up to today. Table 2 includes a list of queries (Column 1) that can be asked by users, such as researchers, archaeologists, students, in the framework of specific uses and search scenarios (Column 2) and can be answered by the system based on the locations (Column 3) of the objects.

In the following, 40.000 of the - provided to Europeana - Hellenic objects have been included in our study, with an equivalent amount of more than one million (1.000.000) RDF triples being generated and used for metadata enrichment and query answering. Using the metadata aggregator described in Section 3, the LIDO XML records were uploaded in the proposed system and transformed in EDM RDF, being mapped to the EDM ontology. Figure 5 illustrates the RDF output of an example record. However, this mapping does not suffice for reasoning over these data, because the EDM ontology contains only general axioms about the classes and properties that describe the records. Moreover, data values - strings are used for the description of objects, which are not appropriate for reasoning.

To achieve semantic enrichment, thus providing representations that can be exploited by reasoners, we used the thematic knowledge for hellenic monuments that has been created in the framework of the Polemon and “Digitalisation of the Collections of Movable Monuments of the Hellenic Ministry of Culture” Projects of the Directorate of the National Archive of Monuments[14] and which has been included in the Polydefkis terminology Thesaurus of Archaeological Collections and Monuments [31,32,33,34]. Polydefkis is a terminology thesaurus that adopts a classification of objects according to their usage, operation, material they are made of, appearance and decoration. Based mainly on usage, a large number of objects and monument types has been accordingly classified.

In the following, we focus on the part of this knowledge referring to types of vases, since metadata and photos of vases were provided by most abovementioned Hellenic content providers to Europeana through the presented metadata aggregation system. In particular, the knowledge used contains axioms about vases in ancient Greece, i.e., class hierarchy axioms referring to the different types of vases, such as amphora, alabaster, crater, as well as axioms regarding the appearance, usage, creation period and the material vases were made of. An excerpt from this knowledge (in description logic syntax) mainly focusing on the use of vases is provided in Table 1.

Table 1

Excerpt of the used thematic ontology in description logic syntax

$Amphora \sqsubseteq BigVase \sqcap CloseVase$
$Alabaster \sqsubseteq VaseWithoutHandles$
$Crater \sqsubseteq \exists hasBase.NarrowBase$
$Pycnometer \sqsubseteq \exists hasBody.CylindricalBody$
$Amphora \neq Alabaster$
$Bowl \sqsubseteq OpenVase$
$EnclosedProduct \sqsubseteq Solid \sqcup Liquid$
$Solid \neq Liquid$
$DrinkingLiquid \sqsubseteq Liquid$
$Water \sqsubseteq DrinkingLiquid$
$Wine \sqsubseteq DrinkingLiquid$
$Oil \sqsubseteq Liquid$
$Perfume \sqsubseteq Liquid$
$Cereal \sqsubseteq Solid$
$Grain \sqsubseteq Solid$
$Usage \equiv Carrying \sqcup Storing \sqcup Drinking$
$\exists contains^{-}.\top \sqsubseteq EnclosedProduct$
$\exists isUsedFor^{-}.\top \sqsubseteq Usage$
$Alabaster \sqsubseteq \exists isUsedFor.Carrying \sqcap \exists contains(Oil \sqcup Perfume)$
$Amphora \sqsubseteq \exists isUsedFor.Carrying \sqcup \exists isUsedFor.Storing$
$Aryballos \sqsubseteq \exists isUsedFor.Storing$
$Aryballos \sqsubseteq \exists contains.Perfume$
$Cup \sqsubseteq \exists isUsedFor.Drinking$
$Lecythus \sqsubseteq \exists isUsedFor.Storing \sqcap \exists contains.(Perfume \sqcup Oil)$
$Pithos \sqsubseteq \exists isUsedFor.Storing \sqcap \exists contains.(Oil \sqcup Cereal \sqcup Grain)$
$Hydria \sqsubseteq \exists isUsedFor.Carrying \sqcap \exists contains.Water$
$Vase \sqcap \exists isUsedFor.Storing \sqsubseteq StorageVase$
$Vase \sqcap \exists madeIn.ArchaicPeriod \sqsubseteq ArchaicVase$
$ArchaicVase \sqcap Amphora \sqsubseteq ArchaicAmphora$
$\exists isUsedFor.Storing \sqcap \exists contains.Liquid \sqsubseteq LiquidStorageVase$

After the creation of the above described thematic ontology, the EDM instances were mapped to terms of this ontology. In particular, from the data values appearing in the range of some roles, individual URIs were created and after being connected (through roles) to proxy instances they were added to the ontology. These were further linked to concepts and roles of the ontology. The creation of individual URIs and

[14] http://nam.culture.gr

Table 2

User queries and associated context

Query	**Scope**	**Location of objects**
Pottery of Mycenaean period found in museums of Peloponnese, Crete, Aegean islands	Research for findings while designing organization of an archaeological (physical and virtual) demonstration	Such items can be found in the HMCT portal and in Europeana coming from the archaeological museums of Kalamata Peloponnese, Heraklion Crete, Ierapetra Crete, Sitia Crete, Kea and Chios in Aegean
Minoan pottery with sea pace decoration	Research for publishing findings from excavation	Items from the archaeological museums of Heraklion and Sitia, Crete
Jewellery of Hellenistic period	Collection of content for museological educational programs	Items from the archaeological museums of Thessaloniki, Kalamata, Larissa, Athens, Pella
Molyvdovoula (king's stamps) of the Middle and Late Byzantine period	Presentation of characteristic archaeological objects in a University course	Items from the Museum of Byzantine Culture and the Numismatic Museum Athens
Minoan and Mycenaean Wall Paintings	Organisation of content for archaeological tours	Items from the Archaeological Museums of Thiva and Heraklion
Figurines from the Geometric up to the Early Classical period	Electronic aggregation of findings, from a single excavation, that are scattered in different locations or Departments	Items from the National Archaeological Museum and the Museums of Thiva and Samos
Engravings and paintings of the 19th century	Search for materials in order to create a thematic portal of archaeological content	Items from the Museum of Byzantine Culture, the Byzantine and Christian Museum, the Rethymno Preveli Monastery and the Pyrgos Picoulaki Museum in Aeropoli
Coins of the late Byzantine period	Preparation of a publication or organization of an exhibition	Items from the Museum of Byzantine Culture and the Nomismatic Museum Athens
Individual inscriptions of the Roman period	Providing additional educational content to courses (e.g., history) of the primary or secondary education	Items from the Epigrafic Museum
Copies of Byzantine paintings of the 20th century	Organising touristic visits for educational or training purposes	Items from the Byzantine and Christian Museum

their mapping to the thematic ontology was done using string matching and stemming on the fields of the EDM ontology regarding the type, creation date, material and museum that proxy instances are found. The OWL API has been used for the creation of the thematic ontology and for the parsing and processing of the EDM RDF data. For some data values, proxy instances were directly assigned to concepts of the ontology. For example, each proxy has been put as an instance of one vase type. As far as the creation date of objects is concerned, time was split to periods of particular interest and each proxy instance was assigned to one of these periods according to the value in the appropriate field of the EDM RDF data.

The resulting tuples of this procedure were then added in a Sesame[15] repository.

Using the above described ontologies and data sets, we applied the methodology described in Section 4 to generate queries and provide semantic answers to them, as described below. All experiments were performed on a Windows 7 machine with a double core 2.53GHz Intel x86 64 bit processor and Java 1.6 allowing 1GB of Java heap space.

A sample of the tested queries are shown in Table 3, where the times needed to answer them are shown. The first column after the Query column refers to the running times of the *RewrQA* and *Execute* methods of Section 4.1, while the second column refers to the running time of the method *EntailQA* for all ABoxes A_i that the initial ABox is split into in Section 4.2. Table 3 does not show the total running time of our system, since it progressively provides the results as they are computed by methods 1 and 2.

The queries start with nearly database/triple store queries that do not need any reasoning to get answered but involve only a retrieval task from the repository and continue with queries that make use of knowledge that is expressible in OWL 2 DL. In particular, Query 1 is matched to triples that are explicilty found in the triple store without any reasoning taking place. Query 2 asking for the clay vases made in the Copper period again does not require any reasoning to get answered apart from the definition of the Copper period; it is more restrictive than Query 1 since it poses more

[15] http://www.openrdf.org/

Table 3

Response times (ms) of the Query Answering (1) and (2) methods and System Results

Query	Running time (1)	Running time (2)	Results of our system	Results without reasoning	Precision(%)	Recall(%)
1. $Q(x) \leftarrow Amphora(x)$	147	3828	118	118	100	100
2. $Q(x) \leftarrow Vase(x) \wedge madeBy(x,y) \wedge Clay(y) \wedge madeIn(x,z) \wedge CopperPeriod(z)$	295	15911	348	348	99.4	98.9
3. $Q(x) \leftarrow ArchaicAmphora(x)$	132	13302	23	0	95.7	95.7
4. $Q(x) \leftarrow Vase(x) \wedge isUsedFor(x,y) \wedge Storing(y)$	223	22887	322	0	100	100
5. $Q(x) \leftarrow OpenVase(x)$	165	13080	404	0	94.8	95.3
6. $Q(x) \leftarrow VaseWithTwoHandles(x)$	189	11939	248	0	92.7	92

constraints on the vases that are matched to variable x, needing therefore more time to get answered. The precision and recall values are lower than those of Query 1, because some slight variations exist in the duration of the Copper period used by different cultural content organizations. Queries 3,4,5,6 all require the use of reasoning which is done either in DL-Lite$_R$ (Queries 4,5,6) or in OWL 2 DL (Query 3). For Queries 4,5,6 we can take all the answers from the query rewriting technique. Query 3 uses some OWL 2 DL axioms of the created thematic ontology. In this case if we want complete answers, we need to use the technique of Section 4.2. The query rewriting technique returns no answers in this case. This happens because the axioms $\exists madeIn.ArchaicPeriod \sqsubseteq ArchaicVase$ and $ArchaicVase \sqcap Amphora \sqsubseteq ArchaicAmphora$ that should be used in the reasoning process to find the answers to Query 3 are disregarded by Rapid (they are not expressed in DL-Lite$_R$). The precision and recall values of Query 3 are about 96% due to the fact that the creation date of a couple of items is given as a range that partly belongs to the Archaic period and partly to the Classical period. Query 4 has precision and recall values of 100% since the knowledge that is used exactly defines the types of vases used for storage, such as amphora, jar, pelike. In Query 5 the knowledge used for the definition of open vases accounts for an error of approximately 5%. Similarly in Query 6 the knowledge used for the definition of vases with two handles is valid for approximately 92% of all vases. In all cases both precision and recall values are very high, illustrating the capabilities of the proposed approach to model well the associated problems and answer the related queries. Looking at the time it takes to answer the queries, it is evident that the query rewriting technique scales much better for larger amounts of data. It is important to notice that without the use of the thematic ontology and the proposed semantic query answering system much fewer results would be obtained, as shown in Table 3 (Results without reasoning). Figure 7 shows an example of a close and an open vase while Figure 8 shows examples of vases with zero, one and two handles. All examples shown can be found in the website of the Hellenic Ministry of Culture and Tourism[16]

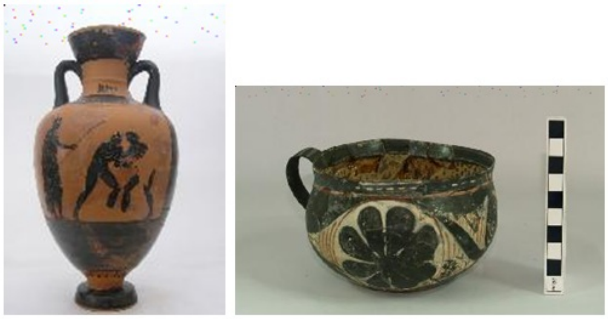


Fig. 7. A close vase (on the left) and an open vase (on the right); the latter is included in the results of Query 5 of Table 3

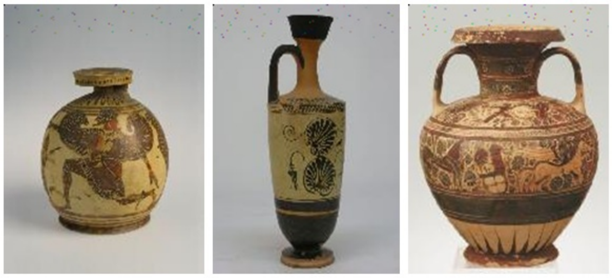


Fig. 8. A vase without handles (left), with one handle (middle) and with two handles (right); the latter is included in the results of Query 6 of Table 3

[16] http://collections.culture.gr/

6. Related Work

A number of systems have been implemented that provide harvesting, mapping, repository and retrieval services, the most important of which are Dspace[17], Fedora[18], Driver[19] and Repox[20].

DSpace is a platform that allows capturing of items in forms of text, video, audio with the purpose of distributing them over the web. It is typically used as an institutional repository supporting ingestion of content, accessing it both by listing and searching, and preserving it. The Fedora digital object repository management system is based on the Flexible Extensible Digital Object and Repository Architecture. Its interface provides administration of the repository, including operations necessary for clients to create and maintain digital objects, discovery and dissemination of objects in the repository. The DRIVER platform constitutes a framework for creating and managing a network of existing repositories. The Driver network-Evolution-Toolkit is already released under the Apache open source license to the public including a repository network administration software and end user services (search, browsing, profiling).

Repox is a framework to manage metadata spaces. It is the system that falls into the same category with the one presented in this paper. It comprises channels to import metadata from data providers, services to transform metadata between different schemas according to user specified rules, and services to expose the results. It has been designed mainly focusing on the Library sector, assisting the Libraries' TEL project partners to import, convert and expose their bibliographic data via OAI-PMH. Repox currently supports MARC21, UNIMARC, MarcXchange and MARCXML schemas out of the box and encodings in ISO 2709. In its current state, Repox is limited to support only the exposure of metadata transformed in the format defined and supported by the TEL project and Dublin Core.

Providing web search engines with semantic capabilities is a target related to the approach presented in this paper. This is the direction followed by the collaboration of Microsoft with Powerset targeting to enhance (in 2012) the 'Bing'[21] capabilities with the developments of the Powerset natural language based search engine. The latter is a tool that extracts semantic relations in queries / phrases, based on natural language processing of their content, working on Wikipedia pages. This is complementary to our system which can be extended to also include natural language processing of users' queries while exploiting the available knowledge as described in the former sections.

The need for developing structured querying facilities, coupled with text retrieval capabilities, has been recognized in recent works, such as [22], where an entity structured scheme called Shallow Semantic Query is presented. This captures entity properties and relationships through shallow syntax requirements implied by user specified predicates at query time; enabling users to issue structured entity-centric queries with typed entity variables and selection/relation predicates. However, this scheme, on the one hand, does not take into account any (existing) knowledge, and on the other hand its effectiveness relies on users' capability to provide proper predicates. In all cases, it can be considered as complementary or of additional value to our system.

Other smaller efforts have targeted towards including criteria and information structures in searching for specific content types. For example, CatScan[22] is a tool which searches article categories (and subcategories) to find articles, stubs, images. Such tools are rather restricted and of limited interest in the framework of the proposed approach.

Let us now refer to the complexity of the proposed approach. As was stated in Section 4 the problem of answering conjunctive queries in terms of ontologies represented in description logics (the underlying framework of the W3C's Web Ontology Language - OWL) has been proved to be difficult, suffering from very high worst-case complexity (higher than other standard reasoning problems) that is not relaxed in practice [7]. This is the reason that methods targeting the development of practical systems mainly follow two distinct directions. The first suggests reduction of the ontology language expressivity used for the representation of conjunctive queries vocabulary, while the second sacrifices completeness of the query answering process, providing as much expressivity as it is needed.

Systems following the first direction focus on the query rewriting approach described in Section 4, i.e., the use of terminological knowledge provided by the ontology to rewrite a user's query and the consequent

[17] www.dspace.org

[18] www.fedora-commons.org

[19] www.driver-repository.eu

[20] repox.ist.utl.pt

[21] http://www.bing.com

[22] http://toolserver.org/~daniel/WikiSense/CategoryIntersect.php/

execution of the rewritten query over a database or a triple store. The main objective is to reduce the expressivity of the ontology language until the point that the procedure guarantees completeness. Late research in the area, introduced the DL-Lite family of description logics, underpinning W3C's OWL 2 QL Profile [8,4], in which the CQ answering problem can be solved in polynomial (over the data) time (actually its complexity is AC0). The main restriction is that in the presence of large terminologies, the algorithm becomes rather impractical, since the exponential behaviour (caused by the exponential query complexity) and the big number of query rewritings affect the efficiency of the system.

Systems following the second direction use approximate reasoning over ontologies expressed in larger fragments of OWL in order to achieve scalability. Approximate reasoning usually implies unsoundness and/or incompleteness. However in the case of semantic query answering most systems are sound. Typical examples of incomplete query answering systems are the well-known triple stores (Jena, Sesame, OWLIM, Virtuoso, AllegroGraph, Mulgara etc).

7. Conclusions and Future Work

Digital Cultural Heritage has been one of the most ambitious and most promising scopes at international level. All over the world, cultural institutions have been digitizing their collections of books, manuscripts, newspapers, maps, museum mobile and immobile objects, archives, audio and visual material, photographs, and are making them available online. Searching for information over all available spaces and semantically interpreting the available cultural content has been one of the main targets of activities performed in national, European and international levels. Different metadata schemas are used to annotate the digitized material and make its access feasible for citizens. Europeana, as well as national and thematic content aggregators provide access to the distributed content through collection of contributing metadata schemas. In this framework, semantic interoperability has been identified as one of the main targets of these developments. Recent results in the Semantic Web and the Linked Open Data fields can be used to achieve these goals. Moreover, user engagement and involvement in evaluating and contributing to the aggregated content and the provided services has been recognized as one of the most critical issues for the development of the field in the following years.

The current paper proposes a system for metadata aggregation and semantic enrichment of cultural content, implementing, in a simple, semi-automatic, user-friendly way, the required mappings and data transformations. Using this system, different users' metadata schemas can be mapped, e.g., to the European Data Model , and expressed in RDF and OWL. As a consequence, they can be used by reasoning and other explorative techniques, in which data from various sources and formats are combined and are appropriately presented to the users so as to cover their needs. In this framework, we propose semantic query answering as the technical approach which can assist content providers and users to enrich their data, to get effective answers meeting the semantics of their queries.

The computational cost of semantic query answering is currently affordable when dealing with normal sized knowledge bases and content sources. Nevertheless, as is indicated in Section 5 (Table 3) the computational load can become excessive when data and inferences are made at very large scales, e.g., at the Europeana level. This holds, even in cases in which the expressivity of the used ontologies is low. For this reason our future work includes investigating scalability of the query answering system. In particular, we will consider algorithms that combine materialization techniques with query rewriting methodologies [17] trying to improve scalability and make the system more efficient. Interweaving query answering with linked (open) data - that are currently widely considered as an important technology for cultural content search [29] - constitutes another important future task that will reduce the computational load of semantic analysis of data and improve scalability. Involving user characteristics, profiles and behaviours can further reduce the computational load and match performances to the context of interaction.

Various interesting results can be obtained by applying the semantic technologies proposed in the paper to the aggregated content. Following the aggregation of content by the Athena project, a study has been performed identifying the different ways used in this content to refer to goddess Athena/Minerva. All information related to her birth and life, as represented on coins, sculptures, vases and paintings has been manually searched and used to create a virtual exhibition, including interactive knowledge tests and games [35]. Extending the results by combining manual search with the semantic query answering method proposed in this paper is a topic we are currently examining for providing users of our system with rich and powerful capabilities when creating services based on the aggregated cultural content.

Acknowledgments The authors wish to thank the Hellenic Ministry of Culture and specifically Ms Metaxia Tsipopoulou, Director of the HMCT Directorate of National Archive and Monuments and Mr Kostas Chatzixristos, Director of the Informatics Division of HMCT for their assistance in working with the cultural content of the www.collections.culture.gr. We also thank Miss Effie Pasatzie for assisting with the mapping of the HMCT metadata schema to EDM through the NTUA ingestion tool.

References

[1] S. Abiteboul, R. Hull, and V. Vianu, (1995). *Foundations of Databases*. Addison Wesley Publ. Co

[2] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications.* Cambridge University Press (2007)

[3] Perez-Urbina, H., Motik, B., Horrocks, I.: Tractable query answering and rewriting under description logic constraints. *J. of Applied Logic*, 8(2), 186-209 (2010)

[4] B. Motik et al (eds.): OWL 2 Web Ontology Language Profiles. W3C Recommendation, (27 October 2009), available at http://www.w3.org/TR/owl2-profiles/

[5] B. Glimm, M. Krötzsch: SPARQL Beyond Subgraph Matching. In: *Proceedings of the 9th International Semantic Web Conference (ISWC 2010)*. LNCS, vol. 6496. Springer Verlag (2010)

[6] I. Kollia, B. Glimm and I. Horrocks: SPARQL Query Answering over OWL Ontologies. In: *Proceedings of the 8th Extended Semantic Web Conference* (ESWC 2011). LNCS, vol. 6643, 382-396. Springer Verlag (2011)

[7] B. Glimm, I. Horrocks, C. Lutz, and U. Sattler. Conjunctive query answering for the description logic SHIQ. *J. of Artificial Intelligence Research*, 31:157–204 (2008)

[8] A. Artale, D. Calvanese, R. Kontchakov, and M. Zakharyaschev. The DL-Lite family and relations. *Journal of Artificial Intelligence Research*, pp. 36–69 (2009)

[9] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. *J. on Data Semantics*, pp. 133–173 (2008)

[10] Diego Calvanese, Giuseppe de Giacomo, Domenico Lembo, Maurizio Lenzerini and Riccardo Rosati. Tractable Reasoning and Efficient Query Answering in Description Logics: The DL-Lite Family. *J. of Automated Reasoning*, 39(3):385–429, (2007)

[11] E. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur and Y. Katz, Pellet: A practical OWL-DL reasoner, *Journal of Web Semantics*, 5(2), 51-53, (2007)

[12] Rob Shearer, Boris Motik and Rob Shearer and Ian Horrocks, HermiT: A Highly-Efficient OWL Reasoner, *Proc. of the 5th Int. Workshop on OWL: Experiences and Directions* (OWLED 2008 EU) (2008)

[13] E. Sirin, B. Parsia: Optimizations for answering conjunctive abox queries: First results. In: *Proc. of the Int. Description Logics Workshop DL* (2006)

[14] H. Perez-Urbina, I. Horrocks, and B. Motik. Efficient query answering for OWL 2. In: *8th International Semantic Web Conference* (ISWC 2009), vol. 5823 of Lecture Notes in Computer Science, pp. 489–504. Springer (2009)

[15] A. Chortaras, D. Trivela and G. Stamou. Optimised query answering in OWL 2 QL. In: *23th Conference on Automated Deduction* (2011)

[16] R. Rosati, A. Almatelli. Improving Query Answering over DL-Lite Ontologies. *In Procs of KR 2010*, pp. 290–300, (2010)

[17] R. Kontchakov, C. Lutz, D. Toman, F. Wolter, M. Zakharyaschev, M. The combined approach to ontology-based data access. In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence IJCAI 2011*, (2011)

[18] H. J. Horst. Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary. *Journal of Web Semantics*, 3(2-3):79-115, (2005)

[19] Frank Manola and Eric Miller, editors. *Resource Description Framework (RDF): Primer*. W3C Recommendation (2004), available at http://www.w3.org/TR/rdf-primer/

[20] Eric Prud'hommeaux, Andy Seaborne, editors. *SPARQL Query Language for RDF*. W3C Recommendation (2008), available at http://www.w3.org/TR/rdf-sparql-query/

[21] Boris Motik, Peter F. Patel-Schneider and Bijan Parsia, editors. *OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax*. W3C Recommendation (2009) , available at http://www.w3.org/TR/owl2-syntax/

[22] Xiaonan Li, Chengkai Li and Cong Yu. Structured querying of annotation-rich web text with shallow semantics. Technical report, CSE Department, UT-Arlington, (2010)

[23] SIEDL: First Workshop on Semantic Interoperability in the European Digital Library, 5th European Semantic Web Conference, Tenerife, Spain, June 2, 2008.

[24] G. McKenna, C. D. Loof. Existing standards applied by European Museums. Report, (2009), available at http://www.athenaeurope.org/index.php?en/149/athena-deliverables-and-documents

[25] The New Renaissance. Report of the 'Comite Des Sages', European Reflection Group on Digital Libraries, January 10, 2011, available at http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/final_report_cds.pdf

[26] Europeana Data Model, available at http://www.version1.europeana.eu/web/europeana-project/technicaldocuments/

[27] C. Bizer, T. Heath and T. Berners-Lee. Linked Data - The Story So Far. *Journal on Semantic Web and Information Systems*, 5(3):1–22, (2009)

[28] Numeric Study Final Report, available at http://cordis.europa.eu/fp7/ict/telearn−digicult/numeric−study_en.pdf

[29] M. Zeinstra and P. Keller. Open Linked Data and Europeana, 2011, http://www.version1.europeana.eu/c/document_library

[30] E. Bermes. Linked Data and Europeana: Perspectives and issues. Europeana Plenary Conference, The Hague, The Netherlands, September 14, 2009

[31] Ch. Bekiari, Ch. Gritzapi and D. Kalomirakis. POLEMON : A Federated Database Management System for the Documenta-

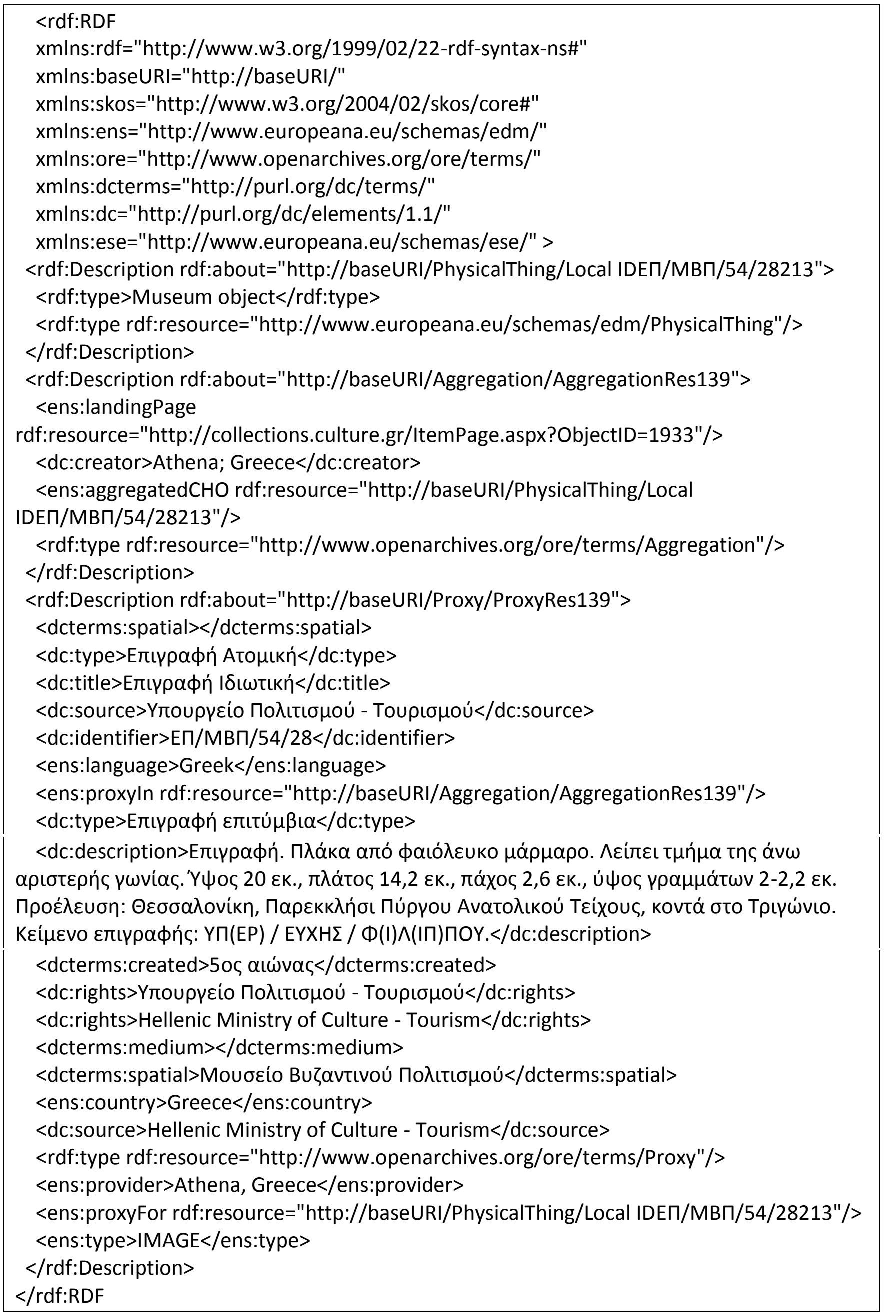
tion, Management and Promotion of Cultural Heritage. In Proceedings of the *26th Conference on Computer Applications in Archaeology*, March 24-28, 1998, Barcelona

[32] M. Doer, D. Kalomirakis. A Metastructure for Thesauri in Archaeology. Computing Archaeology for Understanding the Past. In Proceedings of the *of the 28th Conference*, Lublijana, April 2000, BAR International Series 931, 200, p.117-126

[33] D. Kalomirakis, A. Alexandri. Deploying the POLEMON system for the National Monuments Record of Greece: experience and outlook. In: *Computer Applications and Quantitative Methods*, Archaeology Conference, Heraklion, 2-6 April, 2002

[34] D. Kalomirakis, A. Kalatzopoulou. Polydefkis: A Terminology Thesauri for Monuments. In: *Applications of Advanced Technology in Archaelogical Research and Spilling of its Results* Rethumno, 2000

[35] S. Hazan. A Virtual Exhibition: A Voyage with Gods: the Godess Athena. In Proceedings of the *ATHENA Conference 'Cultural Institutions Online'*, Rome, 28, April 2011



```
   <rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:baseURI="http://baseURI/"
   xmlns:skos="http://www.w3.org/2004/02/skos/core#"
   xmlns:ens="http://www.europeana.eu/schemas/edm/"
   xmlns:ore="http://www.openarchives.org/ore/terms/"
   xmlns:dcterms="http://purl.org/dc/terms/"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:ese="http://www.europeana.eu/schemas/ese/" >
  <rdf:Description rdf:about="http://baseURI/PhysicalThing/Local IDΕΠ/ΜΒΠ/54/28213">
   <rdf:type>Museum object</rdf:type>
   <rdf:type rdf:resource="http://www.europeana.eu/schemas/edm/PhysicalThing"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://baseURI/Aggregation/AggregationRes139">
   <ens:landingPage
rdf:resource="http://collections.culture.gr/ItemPage.aspx?ObjectID=1933"/>
   <dc:creator>Athena; Greece</dc:creator>
   <ens:aggregatedCHO rdf:resource="http://baseURI/PhysicalThing/Local
IDΕΠ/ΜΒΠ/54/28213"/>
   <rdf:type rdf:resource="http://www.openarchives.org/ore/terms/Aggregation"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://baseURI/Proxy/ProxyRes139">
   <dcterms:spatial></dcterms:spatial>
   <dc:type>Επιγραφή Ατομική</dc:type>
   <dc:title>Επιγραφή Ιδιωτική</dc:title>
   <dc:source>Υπουργείο Πολιτισμού - Τουρισμού</dc:source>
   <dc:identifier>ΕΠ/ΜΒΠ/54/28</dc:identifier>
   <ens:language>Greek</ens:language>
   <ens:proxyIn rdf:resource="http://baseURI/Aggregation/AggregationRes139"/>
   <dc:type>Επιγραφή επιτύμβια</dc:type>
   <dc:description>Επιγραφή. Πλάκα από φαιόλευκο μάρμαρο. Λείπει τμήμα της άνω
αριστερής γωνίας. Ύψος 20 εκ., πλάτος 14,2 εκ., πάχος 2,6 εκ., ύψος γραμμάτων 2-2,2 εκ.
Προέλευση: Θεσσαλονίκη, Παρεκκλήσι Πύργου Ανατολικού Τείχους, κοντά στο Τριγώνιο.
Κείμενο επιγραφής: ΥΠ(ΕΡ) / ΕΥΧΗΣ / Φ(Ι)Λ(ΙΠ)ΠΟΥ.</dc:description>
   <dcterms:created>5ος αιώνας</dcterms:created>
   <dc:rights>Υπουργείο Πολιτισμού - Τουρισμού</dc:rights>
   <dc:rights>Hellenic Ministry of Culture - Tourism</dc:rights>
   <dcterms:medium></dcterms:medium>
   <dcterms:spatial>Μουσείο Βυζαντινού Πολιτισμού</dcterms:spatial>
   <ens:country>Greece</ens:country>
   <dc:source>Hellenic Ministry of Culture - Tourism</dc:source>
   <rdf:type rdf:resource="http://www.openarchives.org/ore/terms/Proxy"/>
   <ens:provider>Athena, Greece</ens:provider>
   <ens:proxyFor rdf:resource="http://baseURI/PhysicalThing/Local IDΕΠ/ΜΒΠ/54/28213"/>
   <ens:type>IMAGE</ens:type>
  </rdf:Description>
</rdf:RDF
```

Fig. 5. Example

Project	Content	Metadata Harvesting Standard	Items for Europeana	Evaluated?	Approach	Results	URL (Project, tool)
ATHENA	Museums, Archives	LIDO	4 million	yes	Questionnaire	conditional mappings, element concatenation, constant values, data reports, Europeana preview	http://athenaeurope.org http://oreo.image.ece.ntua.gr:8080/athena/
EUSCREEN	Audiovisual, Television Archives	EBUcore	40 thousand	yes	Questionnaire, Interviews	value mappings, annotation tool, elements statistics	http://euscreen.image.ntua.gr/euscreen/ http://euscreen.image.ntua.gr/euscreen/
CARARE	Archaeological, Architectural	CARARE	2 million	yes	Questionnaire	semantic relations, repository services, EDM preview	http://carare.eu http://carare.image.ntua.gr/carare/
ECLAP	Performing Arts	DC	1 million	yes	Questionnaire, Interviews	string manipulation functions, element annotation, EDM graph visualisation	http://www.eclap.eu/drupal/ http://oreo.image.ece.ntua.gr:9990/eclap/
JUDAICA	Museums, Libraries Archives	LIDO, EAD	500 thousand	no			http://www.judaica-europeana.eu/ http://oreo.image.ece.ntua.gr:9990/judaica/
LINKED HERITAGE	Museums, Archives	LIDO	3 million	Not yet			http://www.linkedheritage.org/
DCA	Contemporary Art	LIDO	500 thousand	no			http://www.dca-project.eu/ http://oreo.image.ece.ntua.gr:9990/dca/

Fig. 6. Use and evaluation of the metadata aggregation system