Evolving LIDO based aggregations into Linked Data

Eleni Tsalapati[1], Nikolaos Simou[1], Nasos Drosopoulos[1] and Regine Stein[2]

[1] Department of Electrical and Computer Engineering, National Technical University of Athens, Iroon Polytexneiou 9, 15780 Zografou, Greece

etsalap ,nsimou, ndroso @image.ece.ntua.gr

[2] Deutsches Dokumentationszentrum fuer Kunstgeschichte - Bildarchiv Foto Marburg, Philipps-Universitaet Marburg, Biegenstr. 11, D-35037 Marburg, Germany

r.stein@fotomarburg.de

Abstract

During the last few years digital evolution of the Cultural Heritage field has accelerated rapidly, not least through the aggregation of cultural content into Europeana. In this process the LIDO harvesting schema has been successfully used in many EU projects (ATHENA , JUDAICA and others) due to its ability to support the full range of descriptive information about museum objects.

The next step, currently being explored in the Linked Heritage project, is the processing of LIDO metadata in order to publish it on the Web as Linked Data, and connecting it to other Linked Data resources. The aim is to provide a generally valid path for the transfer of data from LIDO XML documents to linked RDF resources.

This paper firstly discusses different possible ways that can be used for the RDF representation of LIDO metadata. From this exploration the paper draws some conclusions on the prerequisites and practical steps to be undertaken for successfully evolving LIDO based aggregations into Linked Data. It furthermore presents some preliminary experiments performed for linking resources to external data sources like DBpedia and Eurostat.

1. Introduction

During the last few years digital evolution of the Cultural Heritage field has accelerated rapidly, huge digital libraries and aggregation services are being built like the Europeana virtual library giving access to millions of books, paintings, films, museum objects and archival records that have been digitised throughout Europe. Other examples are the evolving Digital Public Library of America, or the various national portals feeding their content into Europeana: the Italian “CulturaItalia”, the French “Moteur Collections”, the Finnish “Culture Sampo”, the German “Deutsche Digitale Bibliothek”, to name just a few ones. All these services are collecting data and metadata about cultural objects, including museum objects.

As a recent development many of these services, beside presenting the cultural content in a human-readable form in web portals, focus additionally on a publication of this content in a machine-processible format, using Linked Data principles as promoted by the World Wide Web Consortium (W3C).

In this process the LIDO XML harvesting schema (Coburn, E. et al, 2010) plays a major role as it has been and is being used successfully in several EU-funded projects to aggregate cultural content mainly from museum collections for Europeana, and furthermore in various national, regional and thematic online services and research projects.

LIDO, being developed under the auspices of CIDOC, is the result of a collaborative effort of international stakeholders in the museum sector to create a common solution for contributing cultural heritage content to portals and other repositories of aggregated resources, as well as exposing, sharing and connecting data on the web. It is an application of the CIDOC Conceptual Reference Model (CRM)(Crofts, N. et al, 2010) and provides an explicit XML harvesting schema to deliver museum's object information in a standardized way.

The strength of LIDO lies in its ability to support the full range of descriptive information about museum objects. It can be used for all kinds of object, e.g. art, architecture, cultural history, history of technology, and natural history. Moreover it supports multilingual portal environments. LIDO allows for a cost-effective solution to supply museum object information originally stored in collections management systems and cataloguing databases with each one potentially being based on different descriptive metadata formats.

Compared with the most common format used in cultural heritage service environments, the Dublin Core (DC) metadata format which is also the basis of the Europeana Semantic Elements (ESE) it has to be pointed out that LIDO provides a much richer view of museum content.

In the museum community a DC derived metadata schemas is not considered as appropriate: museum metadata is 'flatten out', with most of the data going into a limited subset of elements. For example, a number of different persons and institutions are usually associated with a museum object: the creator or finder of an object, important persons who have used it, the museum currently holding it, previous owners, and so on. All this qualified information is lost in a DC based format. Moreover, the lack of structure that allows elements to be grouped according to their semantic content leads to substantial information loss.

In contrast LIDO provides sufficiently detailed and well-defined semantics while integrating this information on a reasonable level for online services.

Looking into current EU project statistics it can be estimated that around 4,5 million object descriptions are by now delivered to Europeana in LIDO format, increasing with running projects to more than 7,5 million items from several hundred institutions across more than 20 countries and languages. Europeana-related projects using LIDO are: ATHENA[1], MIMO[2], Judaica Europeana[3], Linked Heritage[4], Digitising Contemporary Art[5], Partage Plus[6]. This makes LIDO the secondly most used format in the Europeana environment after the DC based ESE format.

Since Europeana is currently changing its data model from ESE to the new Europeana Data Model (EDM)(Isaac, A. et al 2010) which supports community standards such as LIDO by retaining their full information and integrating them on a higher semantic level, and at the same time moving on to Linked Open Data environments the evolving of the existing LIDO-based aggregations into Linked Data is an obvious necessity.

This paper, after giving in Section 2 an introduction to the Linked Open Data approach and its use in the cultural heritage field, discusses in Section 3 possible ways that can be used for the semantic representation of LIDO records as prerequisite of Linked Data: In a first step the problem of extracting identifiers (URIs) for resources from the metadata is addressed (3.1). The representation of LIDO records in RDF is then investigated firstly using the CRM ontology (3.2), secondly using the EDM ontology (3.3), and finally by introducing an additional LIDO ontology that contains missing parts from CRM and EDM (3.4). Section 4 presents some preliminary experiments for linking resources to external data sources. From these explorations in section 5 some conclusions are discussed for successfully evolving LIDO data into Linked Data.

2. Linked Open Data and Cultural Heritage

In this section a short introduction to Linked Data and its basic principles is made. In addition a use case scenario for the cultural heritage community is presented outlining the benefits from using LOD.

2.1 Linked Data Principles

During the last few years the Web has evolved from a global information space of linked documents to one where both documents and data are linked. This evolution has resulted in a set of best practices for publishing and connecting structured data on the Web known as Linked Data. In few words, Linked Data is simply about establishing typed relations between web data from a variety of sources. These may be as diverse as databases maintained by two organizations in different geographical locations, or simply heterogeneous systems within one organization that, historically, have not easily interoperated at the data level.

[1] http://www.athenaeurope.org/
[2] http://www.mimo-db.eu/
[3] http://www.judaica-europeana.eu/
[4] http://www.linkedheritage.eu/
[5] http://www.dca-project.eu/
[6] http://www.partage-plus.eu/

Technically, Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets (Bizer & Heath & Berners-Lee, 2009).

The main difference of the hypertext Web and Linked Data is that the first is based on HTML (HyperText Markup Language) documents connected by untyped hyperlinks while on the other hand Linked Data relies on documents containing data in RDF (Resource Description Framework) format (Klyne & Carroll, 2004). However, rather than simply connecting these documents, Linked Data uses RDF to make typed statements that link arbitrary things in the world. The result, or as widely known the Web of Data, may more accurately be described as a web of things in the world, described by data on the Web. Berners-Lee (2006) outlined a set of 'rules' for publishing data on the Web in a way that all published data becomes part of a single global data space:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

These have become known as the 'Linked Data principles', and provide a basic recipe for publishing and connecting data using the infrastructure of the Web while adhering to its architecture and standards.

2.2 Usage of Linked Data in the Cultural Heritage domain

Exploration of Cultural Content can be highly improved using information offered by existing digital resources and Linked Open Data. Digital resources may consist of portal aggregating cultural information as in the case of EUROPEANA, or of a network of content providers as is the case of the evolving Digital Public library of America. The metadata model elements are descriptions of the objects with basic and advanced information, starting from the answers to 'Who?', 'What?', 'When?' and 'Where?'. When a user or tourist would like to submit a query for an object or place, the respective metadata are searched and whenever a match is achieved the object or place is included in the results returned to the user, using some answer ranking scheme.

If, however, we want to let users ask complex queries and receive appropriate answers, we need a more detailed description of cultural content in the form of terminological knowledge in various domains (thematic ontologies). Whenever such knowledge is available, we can develop semantic search and semantic query answering, i.e., construct answers to queries posed by users, based not only on string matching over the digital library metadata, but also on the implicit meaning that can be extracted by reasoning using the terminological knowledge, providing details about species, categories, properties, interrelations.

In addition, Linked Open Data sources like DBpedia, Freebase and Eurostat provide more specific descriptions on resources and can also be exploited by tourists. As a consequence, the tourist will be able to obtain additional biographical information regarding on an artist or object creator, demographic information on location, or any other information relevant to the cultural heritage object available.

The LOD2 large-scale Integrated Project[7] pearheads European efforts in Linked Open Data. It aims at contributing high-quality interlinked versions of public Semantic Web data sets, promoting their use in new cross-domain applications across the globe, moving towards a Web of Data. The new technologies for enabling scalable management of Linked Data collections in the many billions of triples will raise the state of the art of Semantic Web data management, both commercial and open-source, providing opportunities for new products and spin-offs, and make RDF a viable choice for organizations worldwide as a leading data management form. Europeana has also launched a substantial Pilot with Linked Open Data, containing metadata on 3.5 million texts, images, videos and sounds gathered by Europeana and belonging to 10 Europeana Content Providers, including about 300 Cultural institutions from 17 countries.

3. Semantic Representation of LIDO records

LIDO has been successfully used in many EU projects (like ATHENA, Linked Heritage, JUDAICA and others) as the harvesting schema to which metadata from various data providers has been mapped. This process has resulted in a vast amount of metadata in LIDO that need to be appropriately processed for its publication as Linked Open Data. Towards this objective the Linked Open Data principles have to be fulfilled and therefore a semantic representation of the LIDO records is required.

A possible approach for the semantic representation of LIDO XML files in RDF is to map the LIDO XML Schema to existing RDF properties and classes appropriately selected so as to be aligned with LIDO's semantics. An alternative is to introduce new properties and classes based on the elements of the LIDO XML schema for this mapping. Two models that facilitate the integration, mediation and interchange of cultural heritage information from heterogeneous resources and form possible candidates for this transformation are CIDOC Conceptual Reference Model (CRM) (Crofts, N. et al, 2010), on which the design of LIDO is based, and European Data Model (EDM) (Isaac, A. et al 2010).

CRM is a generic formal ontology that represents the underlying semantics of the database schemata and document structures used in cultural heritage and museum documentation. It provides the semantic definitions and clarifications needed to transform disparate, localised information sources into a coherent global resource. In this way, it serves as a "semantic glue" to mediate between different sources of cultural heritage information, extracted from museums, libraries and archives.

[7] http://lod2.eu/Welcome.html

EDM, on the other hand, is an integration medium for collecting, connecting and enriching the descriptions provided by Europeana content providers. Particularly, it supports the integration of the various models used in cultural heritage data, so all descriptions can be collected and mapped to higher-lever concepts. EDM adopts an open, cross-domain Semantic Web-based framework that can accommodate the range and richness of community standards like LIDO for museums, EAD for archives or MARC for digital libraries.

Since both CRM and EDM substantially constitute upper level ontologies describing the cultural heritage domain, in the following we examine their usability for the semantic representation of LIDO metadata.

3.1 URIs as names for things

A very important step in the RDF transformation process of LIDO records is the identification of the things described and the extraction, or where necessary creation, of resources with URIs for them. The main requirement for them is to be unique and consistent for every item.

The LIDO model provides repeatable identifier elements for all entities that would constitute a resource, e.g. for actors, objects, places, events, and concepts. These identifiers are obvious candidates, and particularly if there is an identifier of type URI or URL given in the data it can directly be reused, at best resolving to an already published description of the resource.

However, if the LIDO data in question does not contain reusable identifiers for the things described the creation of resources and URIs for them becomes necessary. A first step towards this aim is the identification of the resources that can be shared among the dataset. More specifically a shared resource is a resource made for a thing described in more than one LIDO record. Let’s assume for example two LIDO records describing paintings “Mona Lisa” and “John the Baptist”, both painted by Leonardo Da Vinci. In this case we need to create resources for each painting and also for the LIDO actor “Leonardo Da Vinci”. However, since both paintings were created by the same LIDO actor, there is no need for introducing the same resource two times. Therefore a method capable of determining when new resources have to be created and ensuring their uniqueness is required. Specifically, as LIDO is an event centric schema, a basic direction is to define new resources about the events and across the domains who, what, where and when.

The creation of unique and persistent identifiers is a subject of extensive discussion in Linked Open Data. At this point we only propose a possible way for handling resources and URIs created from LIDO metadata and not a solution that can be generally applied to all cases. The shared resources among the dataset require different handling regarding the construction of URIs from those that their uniqueness is guaranteed. Hence, assuming that all the resources created will be served under a domain (e.g. *baseURI*) together with the prefix resource used for distinguishing machine readable from human readable URIs, our proposal for shared resources is to only use the value of the described thing. In that way the URI for the resource representing an actor of a LIDO item having preferred name "Boticcelli Sandro" would be

http://baseURI/resource/Boticcelli_Sandro

while in a similar manner the resource made to represent the place of a LIDO item would be

http://baseURI/resource/Germany

In that way we are able to easily control duplicates of the same resource while the constructed URIs are descriptive enough to permit linking to them from external data sources. Since actors and places will often be shared among different datasets the detection of and linking to other published descriptions is of particular interest.

On the other hand for the unique things described in a LIDO record, like for example the cultural objects (e.g. painting) or the events that are related to them (e.g. their creation) a different type of URI is required. Although the LIDO schema offers the proper identifier elements objectPublishedID and eventID for these, they are rarely present in the actual LIDO data so far. For the cultural object described in the LIDO record the value of LIDO element lidoRecId constitutes another choice for the creation of a unique identifier, as it is mandatory and unique for every record. This is however a field filled by the content providers and for that reason its uniqueness cannot be guaranteed among different datasets. For that reason the appellation value of the described thing and a universal unique identifier constructed for every resource are used together with lidoRecId as shown in the following example.

http://baseURI /resource/MonaLisa_lidoRecID1435_3FFF000000000000

3.2 RDF Representation of LIDO records using CIDOC CRM

The design of LIDO schema was inspired by CIDOC CRM, therefore the transition from LIDO records to an RDF representation based on CRM appears to be a relatively straightforward process. Most of the elements can easily be mapped to an appropriate set of CRM concepts and properties. For instance, an example of an XML LIDO record is the following:

```
<lido:eventActor>
      <lido:actorInRole>
            <lido:actor lido:type="person">
                  <lido:nameActorSet>
                        <lido:appellationValue
                  lido:pref="preferred">
                                                  Botticelli,Sandro
                        </lido:appellationValue>
                  </lido:nameActorSet>
            </lido:actor>
      <lido:actorInRole>
```

```
<lido:eventActor>
```

More specifically, this is an except of the XML file La Primavera[8]. The definition of element eventActor is "wrapper for display and index elements for an actor with role information (participating or being present in the event)". So informally, this except states that the actor of type "person" and of preferred name "Botticelly, Sandro" participated in or was present at the described event. This information can be represented in RDF using CIDOC as demonstrated below.

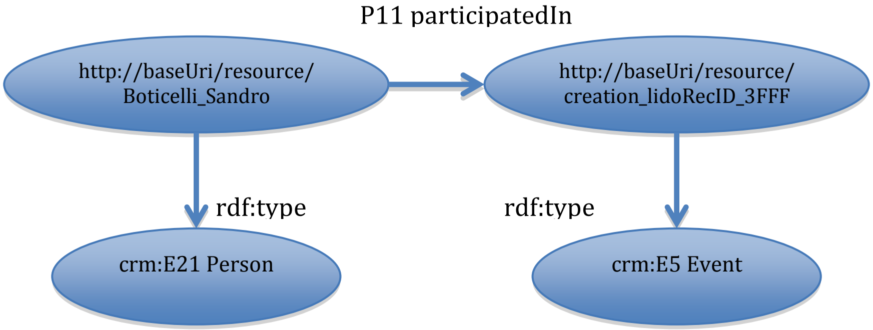
P11 participatedIn
http://baseUri/resource/Boticelli_Sandro
http://baseUri/resource/creation_lidoRecID_3FFF
rdf:type
rdf:type
crm:E21 Person
crm:E5 Event

Fig. 1 A LIDO record Representation in RDF using CRM

The resource created for the actor is an instance of the CRM class "E21 Person", the resource created for the event is an instance of the CRM class "E5 Event", while these resources are connected with the CRM property "P11 participated In".

In a similar manner, the great majority of LIDO elements can be mapped to CRM. However, there are some open conceptualization issues in CRM that do not allow the full mapping of some LIDO datasets to CRM. These limitations are being handled by other models like FRBR (Madison, O, et al 1997) or EDM (Isaac, A., et al 2010). An example of such a limitation is the case of the LIDO element "relatedWorkRelType". This element describes "the nature of the relationship between the object or work at hand and the related entity". By studying a set of LIDO records from the Linked Heritage project dataset we noticed that some candidate values for this element are: "edition", "replica of", "model", "copy of". The CRM general property "P130 shows features of" seems to be an appropriate choice to map these properties to, however with this mapping the special semantic content of these properties will be lost. For such representations it is suggested to define the property "P2 has Type" on the property P 130. For instance, the Parthenon Frieze on the Acropolis in Athens (E22) shows features of the Original Parthenon Frieze in the British museum (E22) and the property "P130 shows features of" P2 has Type copy (E55 Type). This constitutes a general approach of CRM for further specializations of the defined properties.

This is also the case with the LIDO wrapper "descriptiveNoteComplexType". This complex type is a wrapper for a descriptive note, including the identifier of the description ("descriptiveNoteId"), the note itself ("descriptiveNoteValue") and

[8] http://www.lido-schema.org/documents/examples/LIDO-Example_FMobj00154983-LaPrimavera.xml

its sources (“sourceDescriptiveNote”). A physical object may be attributed with a descriptive note which “includes usually a relatively brief essay-like text that describes the content and context of the object / work”. This information can be represented with the CRM property “P 129 is about”, which connects a CRM entity with a “Propositional Object”. However, this property implies that the CRM entity is the primary subject of the propositional object, without indicating that the propositional object is a descriptive note of the object at hand. For instance, the text entitled “Reach for the sky” is about Daglas Baader, but does not constitute descriptive note about it. For capturing this information using CRM a set of sub-properties have to be defined for the property P129. A similar limitation also appears in the case of the LIDO element “roleActor”, which describes the role of a specific actor in a specific event. This information could be mapped to property “P14.1 in the role of”. This is defined by CRM as property of the property “P14 carried out by (performed)”.

As there is no way of modeling properties of properties in RDF, sub-properties of the aforementioned properties should be defined. For instance, we should introduce a set of subproperties, such as “created”, “published” etc, of the property “P14 carried out by (performed)” under a specific namespace and in our case the LIDO namespace.

Apart from the required specialization of the CRM properties, the generalization of some properties is also required in order to achieve the RDF representation of the LIDO data. A representative example of this case is again the “relatedWorkRelType” which relates the specified thing to some other thing, Since the CRM does not provide a general “related to” property between two E70 Thing it needs to be introduced when no further specification of the relationship type is given in the LIDO data. In a similar way a generalized property is needed for the representation of the element “relatedEventSet”, which relates an event that is linked in some way to the specified event. In this case, unless the type of relation is specified, e.g. “overlaps in time with”, “occurs before”, “consists of”, “is separated from”, etc, it cannot be modeled in CIDOC CRM.

There are also some instances in the LIDO datasets, for example the instances of the LIDO element genderActor, that CRM does not seem to provide the appropriate constructs to map them to. At the same time, there is a set of LIDO elements that the only way to describe them in CRM is by the use of the property “P3 has note” which has range E62 String. Some of these elements are: “displayState”, “displayEdition”, “creditLine”. It is easy to see that in this way the semantic representation of these concepts is lost and this is far from the semantic model that we seek for. Again, one way to deal with this, is by introducing a set of appropriate sub-properties of the property has Note.

Apart from the modeling limitations that we described before, there also some further practical issues that discourage us from RDFizing the LIDO records by using CRM. CRM is modeled so as to capture the full range of information provided by museum documentations and therefore it is a complex model making in some cases the representation of the LIDO data dysfunctional. A representative example is the LIDO element vitalDatesActor/earliestDate, which states the birth date of an actor (if the actor is a person). An explicit modeling in CRM would require to add a resource of type E67 Birth, which would be connected with the actor via the property “P98 brought Into life” and with the E 52 Time Span via the property P4 has time-span. While, we are interested in a more straightforward representation of the form actor has birth date some date.

3.3 RDF Representation of LIDO records using EDM

The Europeana Data Model provides constructs that support community standards of cultural heritage content. At the same time, it allows the representation of metadata of either object-centric or event-centric approach. Both CRM and LIDO follow the event-centric approach and therefore EDM can accommodate both initiatives. A representative example of the way that a LIDO record is mapped to EDM is shown below.

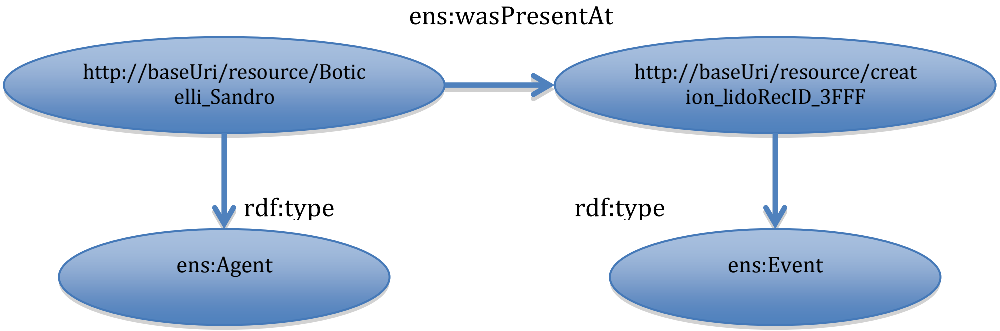
ens:wasPresentAt
http://baseUri/resource/Boticelli_Sandro
http://baseUri/resource/creation_lidoRecID_3FFF
rdf:type
rdf:type
ens:Agent
ens:Event

Fig. 2 A LIDO record Representation in RDF using EDM

Likewise, a significant part of the knowledge covered in most LIDO records can be represented through EDM. However, as the goal of EDM is to accommodate the range and richness of several community standards, not necessarily museum oriented, it constitutes a schema which is even more generic than CRM. Therefore the process of “strict” mapping from LIDO to EDM introduces several drawbacks.

An example that shows the bottlenecks rising from mapping LIDO to EDM is the representation of a relation’s type between a pair of objects. Any information about the relation’s type between objects or works can hardly be modeled with the use of EDM. Particularly, EDM defines specific kinds of relations, like “is successor of”, “is similar to”, “has part”, “incorporates”, between the described objects/works, or the general relation “ens:isRelatedTo”, omitting specific types of relation like “larger context for”, “model of”, “model for”, “study of”, “study for”, “rendering of”, “copy of” that LIDO anticipates as possible types. At the same time, EDM does not provide any specific construct to represent the type of the relation between two events.

Another issue that arose during the process of mapping LIDO records to EDM, was in our attempt to represent an actor’s role participating in an event. Unless the actor is a creator of the work at hand, publisher or contributor, his/her role cannot be specified. In addition, for any personal information concerning the actor, for instance the date and place of birth, the occupation etc, EDM suggests the creation of an explicit link between the work under study and a carefully curated resource. This resource shall stand for the actor as a person and provide all necessary information about him/her, for instance a specific VIAF authority record.

A general observation about EDM is that it provides a strong structure to represent the required knowledge about the cultural heritage object and its correlation with the events that the object participated in, the actors participating in these events, the places that happened at, the time spans that occurred at and the representing resource of the object. However, it does not support directly the representation of the information provided for these resources (actor, place, time, digital representation, etc) related to the cultural heritage object (CHO). In fact, as in the case of actor's personal information, EDM urges the ontology engineer to link the these resources with the appropriate external resources.

3.4 LIDO to LIDO

In the previous chapters we demonstrated a set of examples that show that neither CRM nor EDM can represent the full range of information contained in a LIDO record. On the one hand, although CRM is a conceptual model of very rich expressiveness, some further specializations of its classes and properties are required to cover the more specialized domain of LIDO. On the other hand, EDM is a basic model that can be seen as an anchor to which other finer-grained models, like LIDO, can be attached.

In order to avoid losing any data during the RDFization process of the LIDO XML records, we suggest the creation of a LIDO ontology that contains the further specializations missing from CRM or EDM. Following, we demonstrate some suggestions for the development of the LIDO ontology. This chapter does not constitute a full proposal, it is mostly concentrated on providing some solutions to the modeling problems that we noted in the previous chapters. Particularly, we investigate two modeling approaches: In the first approach, LIDO ontology reuses the required parts of CRM ontology and we define the missing concepts and properties with LIDO namespace (symbolized with the prefix "lido:"), and in the second approach we follow the same process but by reusing the EDM ontology. Finally, we demonstrate the pros and cons of each approach.

3.4.1 CRM- based - LIDO ontology

In this chapter we demonstrate an initial approach to the development of a LIDO ontology that reuses concepts and properties from CRM ontology. Following the structure of the previous chapters, we provide some suggestions for the modeling problems that we have indicated.

As it is already stated, it is impossible to represent solely in CRM the information that the object / work at hand is related to another work, object or a collection, if the type of this relation is unspecified. To overcome this issue we suggest the introduction of the property "lido:isRelatedToThing", which has as domain and range the CRM concept "E70 Thing". As the same problem also applies to the case of two interrelated events, we recommend the introduction of the property "lido:

isRelatedToEvent", with domain and range the CRM concept "E5 Event", i.e. that an instance of type Event can be related to another instance of type Event.

Also, CRM does not provide some types of relations found in the LIDO datasets, such as "edition of", "replica of", etc. Therefore, we suggest the introduction of a set of new properties, such as "lido:isEditionOf", "lido:isReplicaOf", that will be sub-properties of the property "lido:isRelatedToThing", with domain and range the CRM concept "E70 Thing". At this point a new issue arises, as these types of properties are not specified by LIDO, however it could be addressed by LIDO terminology. Similarly, sub-properties of the property "lido: isRelatedToEvent", could also be defined in order to capture the full range of types of relations between a pair of events.

Concerning the modeling issues of connecting a physical thing to a descriptive note of the object, we can introduce the LIDO property "lido: hasDescriptiveNote" as subproperty of CRM property "P129 is about". Its range is the class "lido: DescriptiveNote", which is subclass of "E73 Information Object". The CRM concept "E73 Information Object" is of type "E1 CRM Entity", and therefore it can be attributed with the CRM properties: "P1 is identified by" and "P 70 documents (is documented in)". In this way, the representation of the identification and the source of the descriptive note referring to the described physical object is accomplished. With this modeling solution, the information of and about a descriptive note referring to an event is also easily represented.

Regarding the LIDO elements roleActor, attributionQualifierActor and extentActor, which refer to the specific role of an actor at an event, CRM suggests the definition of a set of sub-properties of the property "P14 carried out by (performed)". Therefore we can introduce an appropriate set of sub-properties, such as "lido: executed", "lido: designed" which again could be extracted by the respective LIDO terminology, in this case e.g. starting from the LIDO terminology for event types. Concerning the limitation of CRM to model the gender of an actor, we can introduce the subclasses lido:Male and lido:Female of the class E21 Person. Finally, elements like "displayEdition", "creditLine" could be represented by defining the subproperties "lido:displayEdition", "lido:creditLine" of the CRM property "P3:hasNote", with domain the class "CRM Entity" and range "E62 String", which, as CRM suggests, is practically modeled in RDF, as a datatype property.

However, the high complexity of the structure of CRM and the special naming conventions that has adopted remain open issues for web oriented tasks, such LOD publication services and resources or linking and cleaning services that are already developed in the europeana project cluster.

3.4.2 EDM- based - LIDO ontology

In this chapter, we demonstrate how the LIDO ontology based on EDM would solve the modeling problems that we dealt with in the chapter LIDO to EDM. To achieve this we introduce an appropriate set of LIDO concepts and properties.

As in the case of CRM, the expressiveness of EDM does not suffice to represent the full range of types of relation between two objects/works. To overcome this problem we can define the set of possible sub-properties, such as “lido: is Replica Of”, “lido: is Edition Of” of the EDM property “ens:is Related To”, as they are defined from the LIDO profiles. Similarly, a set of properties defining the type of relation between two events could also be introduced.

The LIDO elements “roleActor”, “attributionQualifierActor” and “extentActor” can again be represented by introducing an appropriate set of subproperties, such as “lido: executed”, “lido: designed”, of the EDM property “ens: was Present At”. Concerning the personal information of the participating actor as it is provided by LIDO, a possible solution would be to create the respective properties or classes. For instance, the element “nationalityActor” could be represented with the property “lido:hasNationality”, with domain the class “ens:Agent” and range a literal value or a resource. Concerning the modeling of the gender of the actor, we could introduce the classes lido:Male and lido:Female. Finally, the LIDO element “extentSubject” could be represented by introducing the property “lido:hasExtentSubject” as subproperty of dc:subject with domain the “ens:Proxy” class and range a literal value.

4. Linking cultural heritage resources

Linked Data is simply about using the Web to create typed links between data from different sources, therefore after the RDF representation of the LIDO metadata, links to other resources have to be created. For that purpose we have developed a linking algorithm that has been examined within INDICATE[9] project by using the MICHAEL[10] dataset.

Specific information served in the examined dataset such as names of countries, persons and languages, are compared with the values of resources served by DBpedia using SPARQL and, if a match is found, a link is created. Thus, country England is linked to http://dbpedia.org/resource/England and so on. Similarly, person’s names appearing as value of several elements in several items are identified by the corresponding DBpedia resource.

The outcome of this process is very important for the retrieval of the content since the linking to an external source brings to our disposal all additional information served by the source. In the following table we illustrate some evaluation results concerning data enrichment. The first column indicates the number of instances found in Michael dataset and the second one the number of DBpedia resources that were mapped to the corresponding instance. We observed a significantly high percentage of countries and language instances included in Michael data that was linked to DBpedia.

[9] http://83.212.97.67:8080/IndicatePilot/
[10] http://www.michael-culture.org/en/about/project

	Total	Found	Percentage
Countries	16429	15987	97.3%
Languages	11090	11032	99.5%
Persons	6442	3632	56.4%

Table 1: Linking results in MICHAEL dataset

The main difficulty with the linking of person resources is that there is no guarantee that the resource can be discovered in the external source. In other words, the fact that a person is found in the dataset does not necessarily make it a resource served by external sources, while on the other hand countries and languages resources exist in more than one external sources making their discovery easier.

5. Conclusion

The main objective of this paper was to investigate the way for moving from LIDO aggregated metadata that is the outcome of many EU projects to Linked Open Data. For that purpose a sufficient representation of LIDO records in RDF is firstly required. The use of the CIDOC CRM and the EDM data models that facilitate the integration, mediation and interchange of cultural heritage information from heterogeneous resources was examined. CRM on one hand is a very expressive data model on which implementation of LIDO was based and hence it is an ideal candidate for its RDFization. By using CIDOC CRM and the rich vocabulary that it offers enables us to create a semantic representation of LIDO supporting inference services. However CRM in some cases lacks the ability to effectively represent some of the LIDO information or it requires complex concept modelling and extensive instatiation that is not suggested for Linked Open Datasets. On the other hand EDM, as LIDO, is a web oriented model that can efficiently represent LIDO records. In this case the produced RDFized flavour of LIDO would be limited regarding the inference services but most appropriate for the transition to Linked Open Data, since EDM is biased towards this aim. In both cases however the introduction of a LIDO namespace and terminology is necessary for the complete transition of a LIDO record and the representation of its information to RDF. In addition some preliminary results of a linking algorithm examined in the cultural metadata of the MICHAEL project are presented illustrating very good performance.

6. References

Berners-Lee, T. (2006). "Linked Data - Design Issues", Consulted January 31, 2012. Available: *http://www.w3.org/DesignIssues/LinkedData.htm*

Rickley, D. & Guha, R. V.(2004) "RDF Vocabulary Description Language 1.0: RDF Schema - W3C Recommendation", Consulted January 31, 2012. Available: *http://www.w3.org/TR/rdf-schema/*

Bizer, C. & Heath, T. & Berners-Lee, T. (2009) "Linked Data - The Story So Far", International Journal on Semantic Web and Information Systems, Volume: 5, Issue: 3, Publisher: Elsevier, Pages: 1-22 ISSN: 15526283, DOI: 10.4018/jswis.2009081901

Coburn, E. & Light, R. & McKenna, G. & Stein, R. & Vitzthum, A. (2010) "LIDO – Lightweight Information Describing Objects Version 1.0". Available: http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf

Crofts, N. & Doerr, M. & Gill, T. & Stead, S. & Stiff, M. (2011) "Definition of the CIDOC Conceptual Reference Model". Available: http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf

Isaac, A. (2010) "Europeana Data Model Primer". Availabe: http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5

Madison, O. & Byrum, J. & Jouguelet, S. &Mc Garry, D. & Williamson, N. & Witt, M. , IFLA Study Group (1997) "Functional Requirements for Bibliographic Records", September 1997. Available: http://www.ifla.org/files/cataloguing/frbr/frbr.pdf

Prud'hommeaux, E. & Seaborne, A. (2008) "SPARQL Query Language for RDF - W3C Recommendation", Consulted January 31, 2012. Available: *http://www.w3.org/TR/rdf-sparql-query/*

Klyne, G. & Carroll, J. (2004). "Resource Description Framework (RDF): Concepts and Abstract Syntax" - W3C Recommendation. Retrieved June 14, 2009, Consulted January 31, 2012. Available: *http://www.w3.org/TR/rdfconcepts/*