Преглед НЦД 23 (2013), 11–21

Nataša Dakić,
Jelena Andonovski
Univerzitetska biblioteka „Svetozar Marković“

DEVELOPMENT OF A NEW FORMAT EDM FOR METADATA INGESTION IN EUROPEANA

Abstract. Europeana provides access to the digital content from libraries, museums, archives and audio-visual collections across Europe. Project called "Europeana Libraries: Aggregating digital content from Europe's libraries" was launched in January 2011, with the aim that in two years build a library-domain aggregator by which millions of new digital objects will become visible and accessible to end users. By involving in this project University library "Svetozar Markovic" in Belgrade has got the opportunity to present Europe and the rest of the world its valuable collections in digital form: the collection of Alexander the Great and the collection of oriental manuscripts.

Europeana also harvests and indexes the descriptive metadata associated with the digital objects. Format called The Europeana Semantic Elements (ESE) is the metadata set developed for the prototype version of Europeana and it is a Dublin Core-based application profile. It contains all 15 basic elements, as well as 22 qualifiers from Dublin Core. Additionally, it also contains 13 completely new elements that were created for Europeana. However, the launching of the new standard, which is under construction, is planned at the end of the 2011 because it was realized that the existing ESE schema is not semantically sufficient to describe the digital content that is presented through Europeana. This new model is named EDM (Europeana Data Model) and is based on semantic web techniques. The search interface to Europeana based on such techniques is named Thought Lab. EDM is aimed at being an integration medium for collecting, connecting and enriching the descriptions provided by Europeana content providers. To achieve this goals EDM uses set of elements which can be divided in two categories: the elements re-used from other namespaces, and the elements introduced by EDM. The paper will explain in detail the principles of the model and elements used in it.

Keywords: Europeana, digital collections, metadata, ESE, EDM, Resource Description Framework, Simple Knowledge Organization System, OAI Object Reuse and Exchange

Europeana portal, allows users to search digital resources of libraries, archives, museums and audiovisual archives across Europe. Search is possible because all institutions gathered on this site submit metadata about their digital objects, and links submitted by users are directed to the repositories of institutions where the objects can be accessed.

University library “Svetozar Markovic” in Belgrade has got the oportunity that through the project "Europeana Libraries: Aggregating digital content from Europe's libraries" present its valuable collections in digital form: a collection of Alexander the Great and a collection of oriental manuscripts. The aim of this two-year project is to build library aggregator by which 5 million new digital objects will become visible and available to end users.

Europeana, via an aggregator, also collects and indexes descriptive metadata assigned to the digital content. The metadata are currently entered in the format that was developed for a draft version of Europeana, since there was no uniform standard for the creation of metadata.[1] This format, called Europeana Semantic Elements (ESE), is based on Dublin Core format and contains all 15 essential elements and 22 additional elements taken from the Dublin Core Initiative, as well as 13 brand-new elements that were created for the Europeana.

At the end of the year 2011, it is planned to switch from ESE to a new format that is still under construction, because it was realized that the existing ESE format can not sufficiently detailed describe digital content that is presented through Europeana. The new format is called Europeana Data Model (EDM) and is based on techniques of the semantic web, and a search interface for Europeana content based on these techniques is called the Thought Lab.

All institutions, gathered around the Europeana project use different standards for data processing and existing ESE format shape them in a most general common denominator. In contrast to this approach, EDM is an attempt to overcome the barriers to exchange various information belonging to various institutions that make Europeana, such as museums, archives, audio-visual collections and libraries. Essentially, the EDM is not based on any known standard. It adopts the framework of the semantic web, which allows for integrating the distinct information perspectives and needs of the various communities providing data to Europeana and to preserve the original richness of community standards like LIDO[1] for museums, EAD[2] for archives and METS[3] for digital libraries.

EDM enables view and access to objects provided to Europeana, via the packages of digital representations submitted by Europeana providers. It also provides support for ingestion the descriptive metadata submitted by various providers, perhaps even for the same object, as well as representation of new information added by Europeana. In addition, EDM has various description paradigms for the ingested objects, and paves the way for enriching objects by connecting them to semantically enriched resources. And finally, what is crucial to make the EDM, at the same time it allows different levels of granularity in the descriptions, using semantic mapping capabilities. This allows Europeana to retain compatibility with existing description approaches, including the simpler ESE currently used for data submission at Europeana.[2]

Following the basic idea of the model to support the integration of the various models used in cultural heritage institutions, so that all original descriptions could be collected and connected through higher-level concepts, a number of requirements and principles have been formulated.

Basic requirements are:[2]

1. distinction between “provided object” (painting, book, movie, archaeology site, archival file, etc.) and digital representation
2. distinction between object and metadata record describing an object
3. multiple records for the same object should be allowed, containing potentially contradictory statements about an object
4. support for objects that are composed of other objects
5. compatibility with different abstraction levels of description
6. EDM provides a standard metadata format that can be specialized
7. EDM provides a standard vocabulary format that can be specialized

Fundamental design principles:[2]

1. EDM allows data integration in an *open* environment: it is impossible to anticipate all data contributed

[1] Lightweight Information Describing Objects, www.lido-schema.org
[2] Encoded Archival Description, www.loc.gov/ead/
[3] Metadata Encoding and Transmission Standard, www.loc.gov/standards/mets

2. EDM allows for rich functionality, possibly via extensions
3. EDM should re-use existing models as much as possible

Elements of the EDM

Since the EDM should represent an integration medium for collecting, connecting and enriching the descriptions provided by Europeana content providers, it may be said to include any element (i.e., class or property) found in a content provider's description. Giving an account of all these elements is clearly an impossible task, since they form an open set, i.e. a set that can be extended as new providers join the Europeana information space. There is however a well-identified set of elements that EDM uses in order to carry out its task. These elements can be divided into three main categories:[3]

1. the elements re-used from other namespaces
2. the elements introduced by EDM
3. the elements taken from the ESE format

All elements are further divided into two main groups: classes and properties. Classes are then divided into classes re-used from other schemas and classes defined for EDM. Class hierarchy is shown in Figure 1.[3]

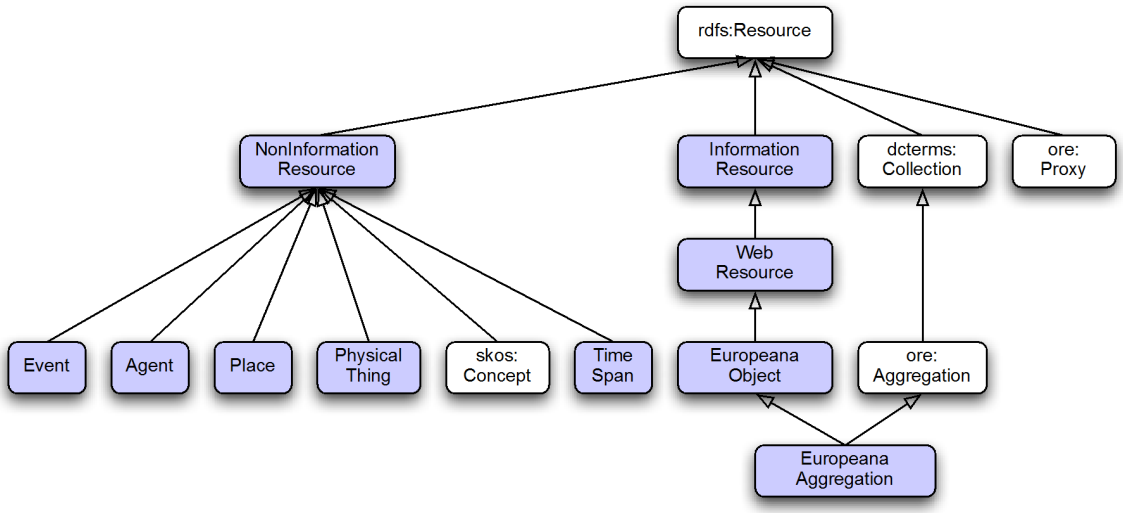

Fig. 1. The EDM Class hierarchy. The classes introduced by EDM are shown is light blue rectangles. The classes in the white rectangles are re-used from other schemas

Properties are divided on the properties re-used from other schemas, the properties defined for the EDM and ESE elements. ESE elements are included in the EDM as the properties because they allow more options for mapping to other data models and thereby enable increasing interoperability of the EDM. The hierarchy of properties is shown in Fig. 2 without entering ESE elements because they would significantly expand the display.[3]

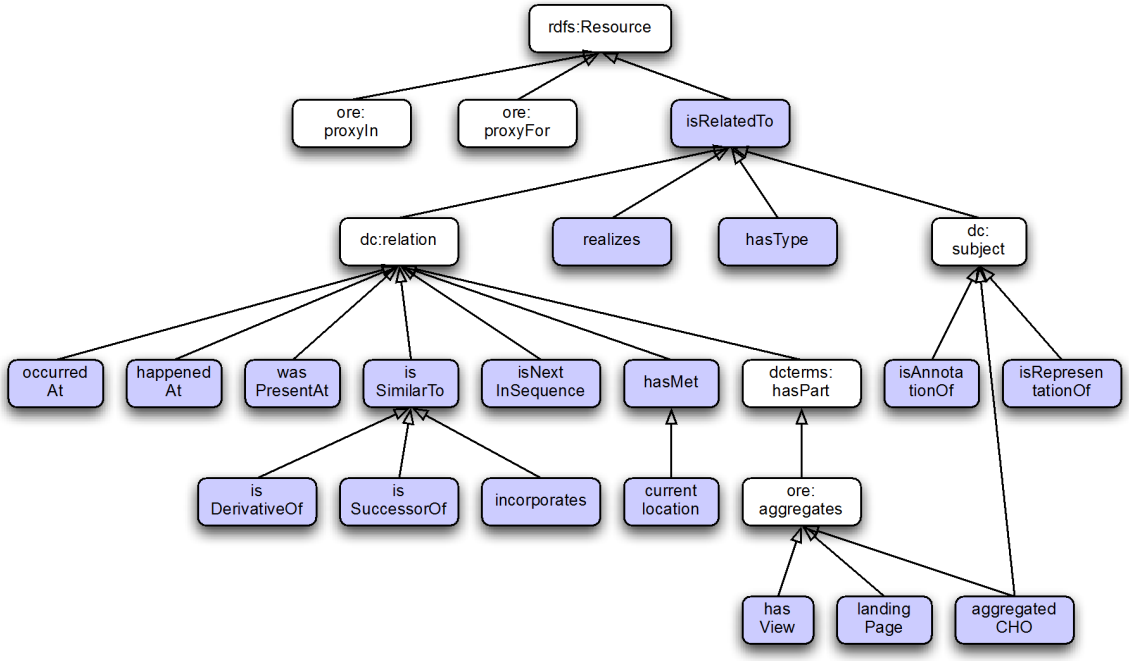

Fig. 2. The EDM property hierarchy without the properties included in ESE (for readability). The properties introduced by EDM are shown is light blue rectangles. The properties in the white rectangles are re-used from other schemas.

Aggregation in the EDM

EDM considers two basic classes of resources provided to Europeana:[2]

1. the "provided object" itself (a painting, a movie, a music score, a book…)
2. a (set of) accessible digital representation(s) of this object, some of which will be used as previews (e.g., a thumbnail of a painting's digital picture).

This allows capturing the distinction between real objects, which are expected to be the focus of users' interest, and their digital representations, which are the elements manipulated in information systems like Europeana.

Provider's aggregation

The structural modelling framework for the EDM ontology is based on OAI-ORE (Open Archives Initiative Object Reuse and Exchange)[4] specification. OAI-ORE is maintained by the Open Archive Initiative which develops interoperability standards that can describe and facilitate exchange of Web resources. OAI-ORE includes approaches for representing digital objects and facilitates access and ingest of these representations beyond the borders of hosting repositories and standardises the description of the relationship between digital objects.[4] It is build on the architecture of the World Wide Web, Semantic Web[5], Linked Open Data[6], Cool URIs[7] and RDF[8] model. Behind the ORE model are four entities:[5]

[4] http://www.openarchives.org/ore/
[5] http://www.w3.org/standards/semanticweb/
[6] http://linkeddata.org/
[7] Cool Uniform Resource Identifiers, http://www.w3.org/TR/cooluris/

– Aggregation: set of Aggregated Resources with its own URI,
– Aggregated Resource: any resource which has its own URI and is part of Aggregation,
– Resource Map (ReM): resource that describes a single Aggregation, enumerates the constituent Aggregated Resources and include additional properties about the Aggregation and Aggregated Resource. Each Resource Map has a single protocol-based URI distinct from the Aggregation URI,
– Proxy: resource that indicates an Aggregated Resource in the context of a specific Aggregation. Proxy has its own URI which must be unique to an Aggragation and to a particular Aggregated Resource of that Aggregation.

Following the ORE approach, EDM considers that the provided object, together with the digital representations that are contributed by one provider, form an aggregation. This aggregation is the result of this provider's activity and is represented using the *ore:Aggregation class*.[6]

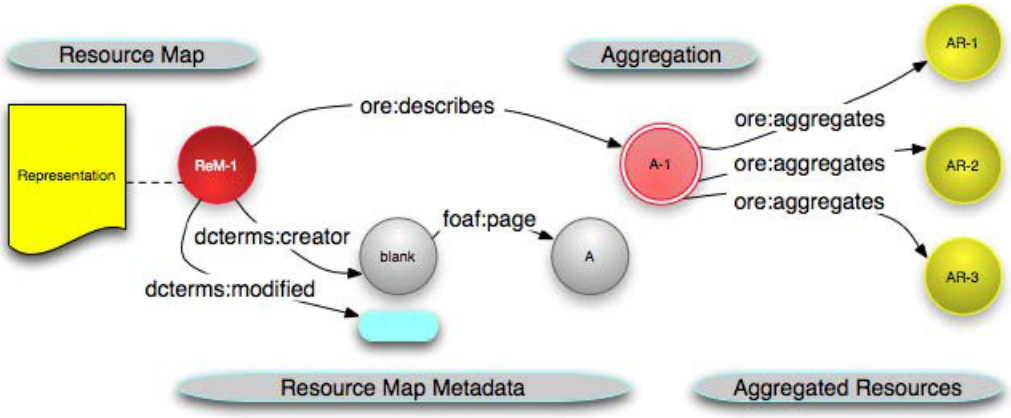

Fig. 3: The core components of the OAI-ORE Data Model

In the Europeana information space, each instance of *ore:Aggregation* is related to:[2]
- one resource that stands for the provided object, using the *edm:aggregatedCHO* property;
- one or more resources that are digital representations of the provided object, using the *edm:hasView property*.

It should be noted that both *edm:aggregatedCHO* and *edm:hasView* are sub-properties of *ore:aggregates*, representing the fact that the aggregation indeed aggregates the real object and its digital views. In the EDM exist other sub-properties which may be used to relate the aggregation to other resources. EDM itself, introduces one such property - *edm:landingPage*, which can be used to link an aggregation to some reference web page for the object. Also, descriptive metadata can be represented for the provided object, e.g., the creator, and to represent such descriptions, EDM uses properties such as *edm:hasMet, dcterms:creator* or *dcterms:title*. At the same time, it also allows use of specializations of these properties, or any other property that providers judge relevant for describing the characteristics of the object.

Although the practice has shown that there is usually one-to-one relationship between an aggregation, a provided object and a metadata record in the original provider's information system, there is no rule enforcing it. In fact, there are situations where a record can give rise to

[8] Resource Description Framework, http://www.w3.org/TR/PR-rdf-syntax/

several aggregations, as in the case of records describing complex, hierarchical digital aggregations and their linking is also foreseen in the EDM.

Europeana Aggregation

Europeana has an opportunity to create new data for the object it ingests so as to provide more value to its users. For the time being, this activity is related only to data formatted using the ESE format. Objects ingested by Europeana often use simple strings as values for the metadata field. Europeana hopes to update that information by linking objects to fully-fledged resources that are thoroughly described and are themselves connected to other resources, such as authority files for persons and thesauri for subjects. These resources enable richer functions, such as query expansion (e.g., using alternatives for a creator's name), recommendation of objects using semantic relations between them (objects created by connected artists), etc.

Europeana creates its own aggregation and proxy for the provided object. This enables the connection of new information to the original object description, while still keeping the distinction between what is provided and what is added. This new Europeana aggregation is modeled using *edm:EuropeanaAggregation*, a specific subclass of *ore:Aggregation*.[2]

Like providers' aggregations, a Europeana aggregation is linked to the provided object using *ore:aggregates*. Thereby, it can aggregate other resources, especially digital representations of the object, or a reference landing page for it, using the *edm:landingPage* property. And thanks to Europeana proxy, it is possible to keep original metadata, besides new richer information which have been generated, (e.g. from an authority file) allowing the display of one or another depending on what level of quality information user requires.

Use of proxies

Europeana takes data from many providers and this data may be about the same real world resource, thus giving multiple views on the same resource. In addition, Europeana can add its own data about that resource giving yet another view on the same resource. These views will not be merged however. In such cases, it is indeed very likely that the metadata will differ, e.g., different names may be used for the same creator. So mechanisms are needed to keep the different views distinct. For now, Europeana leverages the proxy mechanism from *the Object Re-use and Exchange (ORE) model*, which is meant to enable the representation of different views on the same resource.

In the case when two data providers submit different set of digital representations, e.g., different resolutions, different file types and, of course, different locations for the same representation, each provider has to submit a proxy for the object described. This proxy is specific to a given provider, and is used to represent the description of the provided object, as seen from the perspective of that specific provider.[6] With proxies it is possible to represent different, possibly conflicting pieces information on provided objects, while still keeping track of the provenance of this information. For instance, the same old photo which represents Belgrade in the 19th century one provider can called "Image of Belgrade from 19th century"and the other "Belgrade in the 19th century".

A proxy is connected to the resource by using the *ore:proxyFor property* and by using the *ore:proxyIn property* it is connected to its provider’s aggregation.[2] Aggregation can have only one proxy per provided object that it aggregates, since it results from the activity of only one provider. Where two providers have each generated a proxy for the same real object

both proxies must be linked to a resource that represents that object independently of either description context, using the *ore:proxyFor property*.

Also, in the situation when two providers submit two different URIs for the same resource, an identification mechanism *owl:sameAs* has to be applied, as link between the two URIs, which enables the merging of these two resources.

It is difficult to predict how many cases where two providers submit data on a same object will occur, but it should be kept in mind that Europeana aggregators cannot readily know whether the providers they aggregate data for are already providing data through another aggregator. Additionally, there is always a second information source on the provided object beyond its original provider: Europeana itself. Therefore, it can be assumed that the "duplicates" will occur and therefore it is necessary to create model which will cover both these cases.

Providers are expected to make a clear distinction between the metadata that applies to the object itself, and the metadata that applies to the digital representations. And the submission of proxy-based representation would be expected in only two cases:[2]

- for Europeana aggregators who already own several records pointing to the same thing
- for providers that want to link their data submission to objects already ingested in Europeana or curated by other institutions, when they know they have records about these objects as well, and wish to help Europeana to detect this.

It is assumed that all resources have been provided (HTTP) URIs. Providers may not be expected to provide all this. A first suggestion is that providers will submit URIs for web-accessible digital representations (e.g., pictures) and for the provided objects or aggregations that already have permanent identifiers. Europeana itself would assign (or re-assign) URIs for the proxies it creates and for the aggregations that don't have URIs yet.

Descriptive metadata in EDM

EDM includes a set of descriptive and contextual properties that capture the different features of a resource, as well as relate it to the other entities in its context.

If we look at the possible approaches for metadata, it can be distinguished two basic ones: object-centric and event-centric approaches. The former focus on the object described: information comes in the form of statements that provide a direct linking between the described object and its features, weather in terms of simple strings or more complex resources denoting entities from the real world. Most of the metadata formats that are based on Dublin Core format are the application of such an approach.

Event-centric approaches, on the other hand, consider that descriptions of objects should focus on characterizing the various events in which objects have been involved. The idea is that it will lead to establishing richer networks of entities better by representing the events that constitute an object's history rather than with the object-centric approach. This approach underlies models such CIDOC-CRM[9].[6] A typical example of event-centric description, which shows how different places and actors can be unambiguously related to one object via the events these entities participated in, can be found in Figure 4.[2]

Amphora of Tuthmosis III
Identifier: Λ2409
Classification: Amphora

[9] CIDOC Conceptual Reference Model, http://www.cidoc-crm.org/

Event: Type: Excavation
Agent: Stylianos Alexiou
Date: 1951, October
Place: Katsampas, Tomb of the ”blue coffin”, Heraklion
Event: Type: Deposition
Place: Katsampas, Tomb of the “blue coffin”, Heraklion
Period: LMIII A1 (14th century BC)
Event: Type: Production
Place: Egypt
Period: 18th Dynasty, reign of Tuthmosis III (15th century BC)
Current Location: Archaeological Museum of Heraklion Crete
Current Owner: Archaeological Museum of Heraklion Crete
Description: Intact, veined, Egyptian alabaster jar. It has a piriform body, short neck, flat everted rim, foot of biconcave profile, defined by a ring with hollow underside, imitating a slightly asymmetrical base. Two vertical strap handles separate the shoulder from the top of the belly. On one side of the belly is a rectangular frame enclosing a hieroglyphic inscription with the name of Tuthmosis in two cartouches. The inscription reads:
“1.The virtuous god
2. Men-Heper-Re
3. Son of the Sun
4. Tuthmosis, the Fair One in the transformations
5. Blessed with eternal life”.
This imported Egyptian vase of the 18th Dynasty was found at Katsampas, in the tomb of the “blue coffin”, together with other Egyptian stone vessels. The name Men-Heper-Re refers to the pharaoh of the dynasty of Tuthmosis III, who reigned from about the beginning to the middle of the 15th century BC. The vase was probably imported to Crete in the years when Egypt was strongest at sea. […]

Fig. 4.: Event-centric description of an object in EDM

Properties in the object-centric approach

In this approach most frequent properties are: *edm:isRelatedTo*, *edm:hasMet* i *edm:hasType.*

- *edm:isRelatedTo* can be used to link an object to virtually any entity that belongs to its “context”: agents involved in its life cycle, places it has been associated with, subjects it is about, etc.
- *edm:hasMet* is used to relate more precisely a given object to the various things (persons, places, etc.) that have participated to the same events as this object. For example, the creator of an object is an agent that participated in the creation event of that object. Also, the current location of an object can be expressed using the specific *edm:currentLocation property*, which is a sub-property of *edm:hasMet*
- *edm:hasType* connects an object to a concept from a type system to which that object belongs

Properties in the event-centric approach

Object linking with other entities is performed by using the three following properties:

- *edm:wasPresentAt*, represents the connection between any resource and an event it is involved in;
- *edm:happenedAt*, represents the connection between an event and a place;

- *edm:occurredAt*, represents the connection between events and the time spans during which they occurred.

Finally it should be noted that EDM perfectly allows both object-centric and event-centric approaches to co-exist seamlessly for the same object.

Use of the classes for representation of contextual entities

Each type of semantic enrichment leads to significant improvements in the search processes. Therefore, the EDM introduces different classes related to the presentation of contextual entities that can provide more complete descriptions:[2]

- *edm:Agent*, class designed to be used for representing persons or organizations
- *edm:Place*, class designed for spatial entities
- *edm:TimeSpan*, class designed for time periods or dates
- *skos:Concept*, class designed for all entities from knowledge organization systems like thesauri, classification schemes…

These features allow bringing in more information to enhance access to the original objects, They can also enable a complete change of paradigm in the way these objects are accessed, by allowing the user to browse through a semantic space of contextual entities before getting to the actual objects.

Data mapping

As previously shown, EDM provides a number of constructs (classes and properties) that can be used by providers when submitting metadata to Europeana However, it is expected that in most cases these constructs will be used indirectly, via RDF assertions using more specialized constructs. It is expected that providers, while submitting data to Europeana, will submit descriptions that fit their own specific level of interest. The key to ensure interoperability at the semantic level is mapping.

However, the practical details on how to organize the submission of precise metadata together with its mappings are still being elaborate and have been already the subject of workshops organized with representatives from archives, audiovisual archives, libraries and museums. All representatives have provided typical sample data from their collections. The aim was to find out how well the various community standards could be mapped to the EDM.

The archives delivered example files of finding aids for archival material encoded in EAD. The distinct feature of such archival descriptions is the deep hierarchical structure and strong focus on fine grained and contextual description. The EDM properties for part decomposition and incorporation demonstrated its ability to handle descriptions of collections which contain several levels of finer grained sub-descriptions where each intermediate level contains contextual information.

The museums mainly provided examples encoded in LIDO. The strong event-based approach of LIDO format fits very well into EDM. The provided classes and event-centric properties offered modelling possibilities which were flexible enough to integrate the rich event-centric descriptions of LIDO by means of typing events and creating sub-classes and sub-properties as specializations of EDM ones. The museum community, however, suggested to replace EDM classes and properties with CRM ones wherever possible and to use CRM entities without a counterpart in the EDM part of a museum application profile.[6]

The audiovisual archives constitute a very heterogeneous community which provides very diverse objects and applies a variety of different encoding standards. In these archives can be found the highest number of digital born object, and although there is no agreed standard for this type of material as LIDO for museums or EAD for archives, EDM proofed to be able to integrate the diversity and richness of the audiovisual provided examples.

The library representatives provided, among other, a number of very complex examples, all of which were successfully mapped to the EDM. It was evident, however, that modelling librarian data in the EDM would benefit a lot from an extension of the model including the FRBR[10] categories. The librarian experts agreed that the introduction of RDA, once operational, would substantially intensify the need to include the FRBR categories, eventually as part of a community application profile.[6]

Conclusion

Workshop results were encouraging, and EDM has proven to be very flexible and stable, a model that can satisfy the needs of the expert users, and at the same time preserves a richness of standards such as LIDO, CIDOC CRM, MARC or EAD.

The basic purpose of the EDM is to offer greater expressivity and flexibility than the existing ESE format. In comparison with prior data models EDM realizes a very high level of abstraction. It is the most radical generalization of metadata properties in the cultural heritage area so far and it does not bind the representation of ingested metadata to one common schema. The EDM carefully integrates well-established ontologies like SKOS[11], Dublin Core, and FOAF[12] in order to allow for rich and interoperable descriptions of Europeana objects.

Also, EDM uses RDF(S) as its meta-model and URIs to identify structured information about cultural heritage objects. The structural modelling framework for the EDM ontology is provided by the OAI Object Reuse & Exchange (OAI-ORE) specifications. This open architecture of the EDM makes Europeana compatible with the Semantic Web paradigm and enables it to become part of the emerging Linked Open Data community.

Bibliography

[1] Sally Chambers and Wouter Schallier, «Bringing research libraries into Europeana: establishing a library-domain aggregator», *Liber Quarterly* 20, 1 (2010), http://liber.library.uu.nl/index.php/lq-/article/view/7980/8291 (retrieved 17.09. 2011.)

[2] Europeana Data Model Primer. http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5 (retrieved 02.10. 2011.)

[3] Definition of the Europeana Data Model elements: version 5.2.3. http://pro.europeana.eu-/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22 (retrieved 02.10. 2012.)

[4] David Tarrant, Ben O'Steen, Tim Brody, Steve Hitchcock, Neil Jefferies and Leslie Carr, "Using OAI-ORE to Transform Digital Repositories into Interoperable Storage and Services Applications", *Code4Lib Journal* 6 (2009), http://journal.code4lib.org/articles/1062?utm_source=feedburner-&utm_medium=feed&utm_campaign=Feed%3A+c4lj+(The+Code4Lib+Journal) (retrieved 02.10.2012.)

[10] Functional Requirements for Bibliographic Records, http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[11] Simple Knowledge Organization System, http://www.w3.org/2004/02/skos/

[12] Friend of a Friend, www.foaf-project.org

[5] ORE Specification - Abstract Data Model. http://www.openarchives.org/ore/1.0/datamodel#Entities (retrieved 03.10.2012.)

[6] Martin Doerr, Stefan Gradmann, Steffen Hennicke, Antoine Isaac, Carlo Meghini and Herbert van de Sompel, «The Europeana Data Model", Work presented at World library and information congress: 76th IFLA general conference and assembly Gothenburg, Sweden, 10-15. August 2010, http://www.ifla.org/files/hq/papers/ifla76/149-doerr-en.pdf (reviewed 17.09. 2011.)

[7] Open Archives Initiative Protocol – Object Exchange and Reuse. http://www.openarchives.org/ore/ (retrieved 02.10.2012.)

[8] Sanja Antonic, Jelena Mitrovic and Adam Sofronijevic, «Fostering Open Access usage by creation of the library aggregator for Europeana: project Europeana libraries», INFORUM 2011: 17th Conference on Professional Information Resources, Prague, May 24-26. http://www.inforum.cz/pdf/2011/antonic-sanja.pdf (retrieved 15.09.2011.)

dakic@unilib.bg.ac.rs andonovski@unilib.bg.ac.rs