Evolving Europeana's Metadata: from ESE to EDM

Boyan Bontchev[1]

[1]Chair of Software Engineering, FMI - Sofia University St Kl. Ohridski,
5, J. Bourchier Blv., Sofia, Bulgaria
bbontchev@fmi.uni-sofia.bg

Abstract. The evolution of metadata standards used for building the Semantic Web makes possible single, integrated access to regional digital libraries. The ambition of Europeana project is focused at realisation of such an integrated, secured multilingual access to digital collections of European cultural heritage artifacts distributed at various European countries. The present paper represents in brief both the goals and evolution of Europeana and its technical implementation steps. It discusses possible ways of contributing to this effort by submitting descriptive metadata according two specific notations – the flat property-value format of Europeana Semantic Elements (ESE) and the Europeana Data Model (EDM) specifying how resources as networked.

Keywords: digital content, metadata, Europeana, ESE, EDM, cultural heritage.

1 Introduction

In last decades, integrated multilingual Web access to locally developed multimedia content turns to be of key importance but still hard to become a reality for many domains like tourism and cultural heritage. Modern content tends to be more heterogeneous and distributed; it is produced using collaborative processes in various forms, formats, and languages [1]. With steadily evolving Semantic Web, worldwide content access services like semantic and optimized search and browsing increase the traffic of content providers and aggregators at national and regional level [2].

The Europeana project[1] aims at realizing an integrated multilingual access to digital collections of European cultural heritage distributed at various organizations of the European countries. Such collections contain digital objects representing art works available at museums, libraries or archives and, being a part of our multicultural and multilingual heritage, deserve to be discoverable within a unique online software environment. Besides integrated Web access to digital of European digital collections of cultural heritage artifacts, Europeana aims at building an open services platform providing facilities for management of large collections of surrogate objects representing digital content, by means of special application programming interfaces [3]. The services for access and managing of culture heritage content are available for both individual users and cultural institutions.

[1] http://pro.europeana.eu/

According to a statement of European parliament, Europeana is "a digital library that is a single, direct and multilingual access point to the European cultural heritage". The Commission's objectives for Europeana could be summarized as follows [4]:

- To create a multilingual access point to Europe's cultural and scientific heritage, which represents a public-domain
- To develop a wide range of information products and services making use of digitized cultural and scientific heritage resources
- To play a key role in expected future growth of important and promising industrial sectors such as technology based learning and tourism
- To inspire new creative enterprise and business innovation
- To promote understanding of a common European background and real sense of a European identity

Working towards implementation of these objectives, Europeana team brings added value to already existing cultural content by juxtaposing related images, texts, videos and audio collections by "'repatriating' geographically dispersed content into a single, coherent and contextual virtual space" [5]. Europeana provides multilingual integrated interface to dispersed, multi-format cultural content of Europe and, thus, enriches user experience.

The name Europeana stands for two different things: EUROPEANA foundation and EUROPEANA service. The Europeana Foundation represents the governing body of the Europeana service. The Foundation functions under the Dutch law and is geographically located at the national library of the Netherland - Koninklijke Bibliotheek. It facilitates and encourages collaboration between such organizations and museums in terms of exchanging resources and data about archives, audiovisual collections and libraries aiming at integrated access through Internet to their content through Europeana services.

The first prototype of the Europeana service was launched in November 2008 at the European Council of Ministers of Culture. This prototype provides integrated Web access to circa 2 million objects coming from about 100 content providers. Though that initial success, some warnings have been reported [5] such as the facts about majority of graphic content (77% of overall content are images) and the major part of the content (82%) coming from four countries. Thus, the major challenge for Europeana has been determined as "how to engage all cultural heritage content providers across Europe and manage to harvest, index, harmonize, enrich and make this content available in a sustainable, robust and user-friendly way to users world-wide".

For the moment, organizations from circa 30 European countries contribute to Europeana, providing metadata about content in twenty-six languages, with four types of materials, namely image, sound, video, text. It is important to note, that the Europeana service uses only harvesting of metadata describing digital objects, which represents cultural heritage objects (CHOs). Thus, the service provides a preview (thumbnail) of given described object together with a link to its location at the site of content provider or aggregator directing its users to this site in order to view details of the digital object and the object itself.

The paper presents an overview of the goals of the Europeana project, its technical implementation steps and the ways of contributing to it by submitting descriptive metadata organized according metadata notations specific for the project.

2 Europeana Aggregators

A content provider for Europeana may be any organization (and even individuals, in the future) that provides digital cultural heritage content accessible on the Web via Europeana. As stated above, Europeana stores only the institution's metadata and indexes it, however, digital objects remain stored at the content provider site (library). The principal goal of Europeana is to provide an integrated multi-lingual Web access to digital content of cultural heritage of thousands of European cultural institutions. However, there exist some obstacles for achieving this goal, such as great variation of technical infrastructures and of type and output formats of content, which is available at existing content providers. Thus, due to the great amount of work for metadata harmonization and normalization, Europeana does not practically collaborate individually with any content provider (though such opportunity does exist) but rather works with an intermediate layer of aggregators of the content providers [5].

By definition, an Europeana aggregator is "an organisation that collects metadata from a group of content providers and transmits them to Europeana" [6]. The aggregator represents a business entity that aggregates descriptive metadata from content (data) providers, usually through metadata harvesting using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), in order to make it available to Europeana. Besides simple collecting of metadata from its content provider at national or regional level, the aggregator organization has other core (or primary) functions such as standardizing file and metadata formats and providing help to the provider regarding conformance with metadata standards and process workflows of Europeana. The aggregator is supposed to help content providers with guidance and trainings and, in many cases, with administration and operation services (like discovery and search services), as well. On the other side, an aggregator has also secondary functions, namely [6]:

- Disseminate the vision and objectives of Europeana to their institutional network for a better support for and involvement with Europeana;
- Provide feedback about Europeana discussions and, as well, specific domain expertise and skills to their content providers;
- Promote standards along their content provision chain.

There exist several groups of types of aggregators regarding provisioning of public access to aggregated content, running content repository and aggregation of content belonging to single or multiple domains, as follows:

- Public versus non-public - the aggregator may run its own Web portal providing public access to locally aggregated cultural heritage content (such as national aggregators like culture.fr and bulgariana.com, or regional aggregators like erfgoedplus.be) but this is not mandatory. Aggregators without public Web portal are called 'dark' aggregators;
- Digital repository aggregators (storing digital items to a repository) versus intermediary aggregators (only collect metadata with a link to given digital item). A non-public aggregator may run its own content repository for storing digital items but usually such aggregators only collect metadata plus a thumbnail of the image of the digital item and a hyperlink to it;

- Single sector aggregator versus cross sector aggregators – while a single sector collects data from a single sector such as a local/regional/national museum, archive, library or audio-visual collection, a cross sector aggregator aggregates data from several such sectors, like national Europeana aggregators (fig. 1);
- Horizontal versus vertical aggregators – horizontal ones aggregate content across several domains, while vertical aggregators aggregate content from single domain - at international, national or regional level, like the thematic aggregator Judaica[2] collecting content about the Jewish urban culture from many sectors.

The mission of aggregators includes improvement of technical and organizational collaboration among institutions working in the cultural heritage domain. They aggregate not only cultural content and metadata about it but also communication channels, shared resources and knowledge. These issues determine great benefits for aggregators, such as growing local cultural heritage community and having as added value shared communication, business networking and synergy of efforts for launching new ideas and projects.

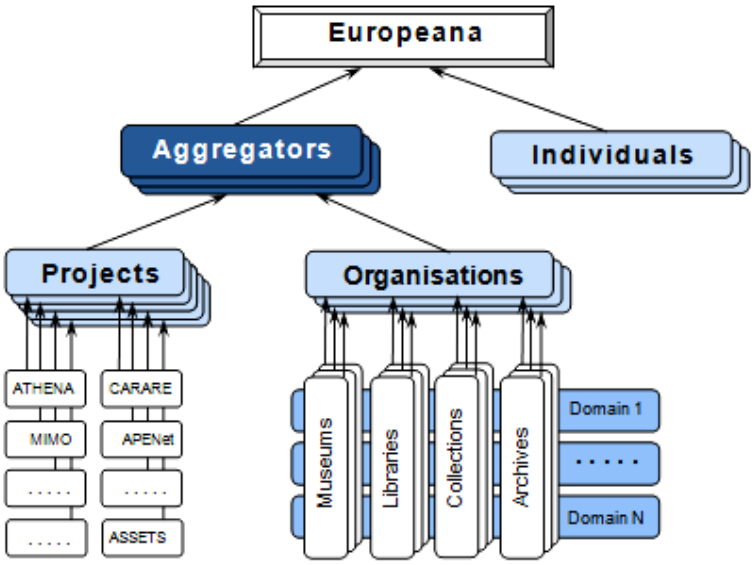

Fig. 1. Direct and indirect contributions to Europeana.

As explained above, submissions of content about cultural heritage objects usually occurs indirectly, i.e. through aggregators, though individual institutions could collaborate directly provided there are not yet established any national/regional aggregators. Thus, projects and organizations accomplish contributions to Europeana usually through aggregators (Fig. 1). Until present, the European Commission has co-funded, mainly through its CIP ICT-PSP Programme (Competitiveness and Innovation Framework Programme), several European projects contributing to

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

Europeana such as ATHENA, APENet, MIMO, JUDAICA, CARARE and many more [5]. These projects have facilitated standardization of metadata describing Web resources and in particular cultural heritage objects and, therefore, have contributed to interoperability in various heritage sectors. Though many of them continue developing their own portals providing access to digital resources in a more specific context, as a whole they integrated a lot of aggregated content into Europeana and, thus, have contributed for a better integrated Web access to the cultural heritage of Europe.

3 Workflow of Contributing to Europeana

3.1 Procedure for Direct Contributions to Europeana

Data submissions to Europeana for any organization - both aggregators and content providers - have to comply with a specific workflow described in [7]. This workflow implies six organisational steps summarized below:

1. At step 1, the provider/aggregator has to a Data Provider/Aggregator Agreement with Europeana Office
2. At step 2, the organization send to Europeana Operations (i.e., to the Content ingestion team) a Data Submission Form with description of type of submission (new/update), licensing and metadata information
3. At step 3, within no more than one month the Europeana Office reviews the submission request and sends feedback and plans for the submission
4. At step 4, the provider/aggregator prepares the data sets to be submitted using the Europeana Semantic Elements (ESE) specifications [8] or Europeana Data Model (EDM) [9]. The Mapping and Normalisation Guidelines [10] may be used for facilitation of this effort. A XML Schema [11] is developed specially to validate the mapping to ESE. As well, the organization may use a Web tool intended for validation of mapping against the XML Schema - named Content Checker [12], and for local testing of ingestion operations such search, browse and display of the data in a copy of the Europeana portal. The organization is supposed to install, configure and test OAI-PMH for data transfer [13].
5. At step 5, after data are successful tested using the Content Checker tool, the Europeana Office validates the transfer with the organization.
6. At step 6, after successful validation, the Europeana Operations ingest submission data into the Europeana production environment and notifies the organization.

3.2 Technical Implementation Steps of EuropeanaLocal Project

EuropeanaLocal was a Best Practice Network project[3] funded under eContentplus from June 2008 until May 2011. The project aimed at involving into Europeana

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

libraries, museums and archives at local and regional level. In order to facilitate their contributions, the project defined several technical implementation steps to be conducted by any organization which wishes to contribute its content to EuropeanaLocal project, therefore, to Europena [14]. The steps determine a workflow of implementing OAI-PMH compliant repository solutions on top of digital collections held by local content providers and include several work tasks:

1. Constitution of digital collection(s) - the organization (content provider or aggregator) may holds one or several collections in one collection management system or, in a more complex case, may holds N collections in M collection management systems;
2. Repository installation and configuration - for running locally at the organization; if such local collection management system already exist, its repository should be configured with an OAI-PMH Web service compliant to standard requests from external harvesting repositories;
3. Metadata transformation – includes three sub-tasks:
    a. Metadata extraction from the collection management system to populate the local repository
    b. Metadata normalization - represents a process of harmonizing metadata in order to match a specific format or notations (i.e., given date format)
    c. Metadata enrichment - stays for a manual or semiautomatic process of improve the metadata quality such as multi-lingual content (e.g. by using Google Translate), temporal and/or spatial references (e.g. by using date and location extraction software services), references (through indexing and keyword extraction tools), mapping to common vocabularies such as that of Simple Knowledge Organization System (SKOS) and identification of official authorities. Many automated techniques may be faulty and could incur some errors.
4. Repository population - includes a setup of automatic metadata import into the repository, in most of the cases schedule according the frequency of updating the local digital collection;
5. Metadata harvesting - exercised by the Europeana harvesting service connecting on given schedule to the local repository and downloading either all its metadata or only the metadata changed since the previous download;
6. Usage of aggregator repositories - aggregator repositories are harvested by the Europeana service and, at the same time, may harvest metadata from local collection management systems;
7. Starting end-user services - cross-domain and cross-geography multi-lingual searches already do provide a steadily increasing Web traffic to many local repositories.

As the reader may note, metadata transformation is a rather complex process including extracting, mapping, normalizing and enriching metadata. It sticks either to the Europeana Semantic Elements (ESE) notation – the only one supported by Europeana until the end of 2011, or by the relatively new Europeana Data Model (EDM) being still under development. Next section discusses these two specifications and their benefits and shortcomings.

4 Notations for Descriptive Metadata

The data model of Europeana enables integrated multi-lingual search and discovery of digital cultural heritage objects (CHOs) distributed at local European collections. These services are available thanks to a common central index of CHOs metadata maintained by Europeana. As far as Europeana does not store provided CHOs but their metadata, there is generated a description and a thumbnail (preview) of any found CHO with a link to the content provider or aggregator Web side. Thus, a content provider or an aggregator is supposed do provide to Europeana three types of data:

- Metadata describing given provided CHO
- A preview (thumbnail) of the provided CHO – a thumbnail image or audio/moving image previews. Audio/moving image preview usually is a short extract of audio/video content with lower resolution
- Active and stable links to the provided CHO on the provider/aggregator Web site

Descriptive metadata is mapped to either the ESE or the EDM notation.

4.1 Europeana Semantic Elements

To the present moment, the last version of the specification of Europeana Semantic Elements (ESE V3.4) dates back from the end of March 2011 [8]. This metadata set was developed for the first prototype version of Europeana which is operational since November 2008. It is formed as an enriched application profile of Dublin Core metadata and provides a generic set of terms appropriate for to heterogeneous digital objects.

The ESE specification was developed specially for the first prototype of Europeana in 2008 as a Dublin Core (DC) application profile and therefore incorporates 37 DC terms from both the `dc` and `dcterms` namespaces. As well, there are defined 12 ESE-specific terms using the Europeana namespace, specially purposed for supporting portal functionality. The full list of elements is separated into four groups building the four columns in Table 1 [10].

Table 1. Elements of ESE version 3.4 [10].

Mandatory elements	Recommended elements	Additional elements	Elements supplied by Europeana
dc:title or dc:description dc:language europeana:dataProvider europeana:isShownAt or europeana:isShownBy europeana:provider dc:subject or dc:type or dc:coverage or dcterms:spatial	dcterms:alternative dc:creator dc:contributor dc:date dcterms:created dcterms:issued dcterms:temporal dc:publisher dc:source	dc:format dcterms:extent dcterms:medium dc:identifier dc:rights dcterms:provenance dc:relation dcterms:conformsTo dcterms:hasFormat	europeana:country europeana:language europeana:uri europeana:usertag europeana:year

europeana:rights europeana:type	dcterms:isPartOf	dcterms:isFormatOf dcterms:hasVersion dcterms:isVersionOf dcterms:hasPart dcterms:isReferencedBy dcterms:references dcterms:isReplacedBy dcterms:replaces dcterms:isRequiredBy dcterms:requires dcterms:tableOfContents europeana:unstored	

The mandatory elements are grouped together in order to accommodate different types of CHOs and metadata practices. Some of them are separated in subgroups, where one of them is required to be provided. The elements named `type`, `language` and `rights` are contained in both `dc` and `europeana` namespaces but have different meanings. The other three groups are built by recommended elements, additional elements and elements supplied by Europeana.

The ESE mapping guidelines [10] in general instruct to answer questions about provided CHOs such as "who, what, where and when?". Thus, the organisation should provide names, types, places and dates relevant to the provided CHOs by mapping as many as possible of original source metadata to the specified ESE elements. As far as many of the elements are refinements of other elements (e.g., `dcterms:created` and `dcterms:issued` are refinements of `dc:date`), the provider must always use the more specific `dcterms` refinements when possible (e.g., to use `dcterms:spatial` or `dcterms:temporal` instead of `dc:coverage`). When it is difficult to decide about appropriate mapping of provider's metadata element to ESE, then `europeana:unstored` could be used.

As far as ESE represents lowest common denominator for object metadata by converting datasets to a Dublin Core profile. Thus, it supports interoperability but is supposes a loss of original metadata. Other drawbacks of ESE [15] consists in the "flat" ESE model using only simple string values and preventing linking items ingested by Europeana to other CHOs such as contextual entities (e.g., naming variations of the CHO's creator) or more specific concepts. As well, ESE aggregates in one record metadata fields applying to different entities like providers' use rights. That's why the new EDM specification was created.

4.2 Europeana Data Model

In order for solving the shortcomings explained over, in last years the Europeana team works for creation of a new, RDF-based Europeana Data Model (EDM) aiming at replacement the old the ESE schema. It aims at enabling Europeana transition (in Europeana Version 2.0) from a closed data repository to an open Web 3.0 information space of Linked Data. The last version of EDM is given in [16].

The EDM design is focused at enabling data integration and interoperability by means of semantic linking between objects, while preserving original metadata (unlike ESE) and, as well, supporting an enriched functionality (like semantic search).

For reaching these goals, EDM makes use of fundamentally new principles for ingesting, managing and publishing metadata about CHOs [17], such as:

- data integration in an open environment (because it is feasible to anticipate all submitted metadata)
- rich and extendible functionality
- broad reuse of existing (standard) models

EDM provides abilities to distinguish any provided CHO (e.g., a painting, book, video, etc.) from its digital representation(s) and, on the other side, from its metadata record. It allows as well co-existing of several possible (even contra-dictionary) descriptions of the same in Europeana, because different organizations may submit different descriptive metadata. Another feature of EDM is sequence or partitioning needed to support representation of complex items and compositions. For this purpose, Open Archive Object Reuse and Exchange Model (OAI-ORE[4]) is used as a reference model for the description and exchange of aggregations of CHOs metadata. It enables compatibility with different abstraction levels of description, based on using the latest version of DCMI Metadata Terms[5] specified as RDF model. The W3C SKOS[6] model for Knowledge Organization Systems is used together with OAI-ORE and DC as a vocabulary format that can be specialized.

The Europeana Linked Open Data Pilot[7] is the first Europeana EDM pilot system [15]. It includes three core classes:

- a provided cultural heritage object (`edm:ProvidedCHO`) such as a painting, book, movie, music record, etc.)
- one or more digital representations of this object accessible via Web including its previews (`edm:WebResource`)
- an aggregation (`ore:Aggregation`) aggregating `ore:ProvidedCHO` and one or several `edm:WebResource` via two sub-properties of `ore:aggregates` - `edm:aggregatedCHO` and `edm:hasView`, respectively (fig. 2).

Aggregations capture digital environment of given provided CHO by attaching descriptive and contextual information about different features of the resource, allowing both *object-centric* and *event-centric* approaches. While the object-centric approach is focused to the object metadata like `dc:format`, `dcterm:title`, `edm:hasMet`, `edm:currentLocation` and `edm:hasType`, the more complex event-centric approach enriches this data with contextual classes like `edm:Agent` (representing persons or organizations), `edm:Event` (with `edm:wasPresentAt` property), `edm:Place` (with `edm:happenedAt` property), `edm:TimeSpan` (with `edm:occurredAt` property) and `skos:Concept` (for SKOS entities thesauri and classification schemes).

As far as Europeana receives data from many providers, some data may represent multiple views on the same CHO. In order not to merge these different metadata records about the same object, Europeana supports provider's proxy of this CHO, modeled using the `ore:Proxy` resource. A proxy is specific to given aggregation and

[4] `http://www.openarchives.org/ore/`
[5] `http://dublincore.org/documents/dcmi-terms/`
[6] `http://www.w3.org/2004/02/skos/`
[7] `http://data.europeana.eu`

represents description of the provided CHO specific for the provider and therefore for that aggregation. The proxy is attached to the aggregation that contextualizes it using `ore:proxyIn`. Thus, Europeana may support different, even conflicting metadata about any CHO, received by different providers about that CHO. Fig. 2 presents an event-centric EDM description of a painting of the Bulgarian artist Tsanko Lavrenov titled "The Old Plovdiv". The artist is described via a Virtual International Authority File (VIAF) record.

Finally, EDM supports Europeana's aggregation (`edm:EuropeanaAggregation`) which bundles together all providers aggregations for a given CHO. Meta-level statements on the creation and publication of ORE aggregations are supported by means of OAI-ORE Resource maps.

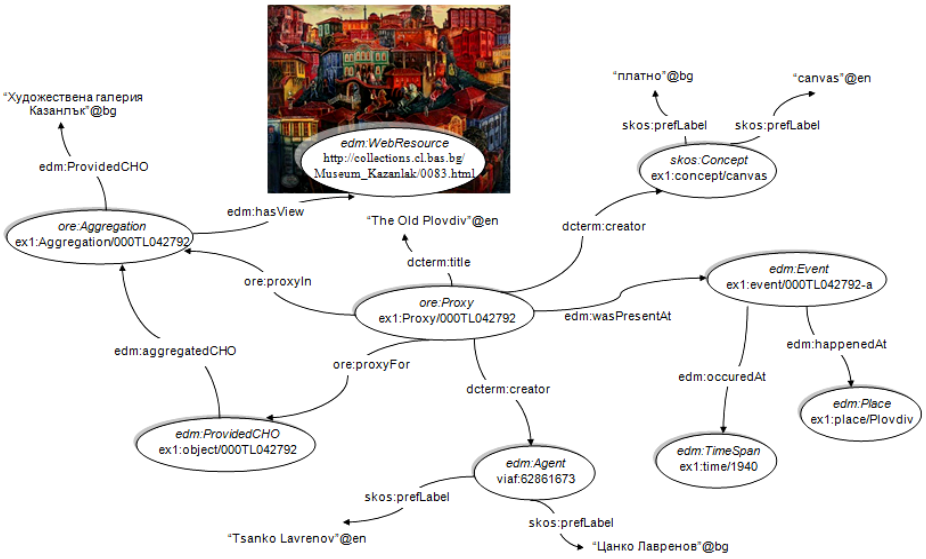

Fig. 2. Event-centric description of "The Old Plovdiv" by Tsanko Lavrenov (1940).

5 Conclusions

As explained above, ESE metadata format supposes flat lists of property-value pairs while EDM specifies how the resources as networked. Therefore, one-to-one mapping from ESE to EDM is not possible. On other hand, EDM structure leads to complex network of aggregations, proxies and other resources, which makes the RDF graphs rather complex. Tracking object data provenance without using proxies will be expected after starting applying RDF named graphs [18] when W3C will finalize its specification.

Though the complexities which EDM does incur, it allows flexible metadata modeling with valuable semantic issues. It preserves original metadata of the providers while facilitating data interoperability. As well, it does support effectively both object-centric and event-centric advanced metadata modeling, allowing property

relations between provided objects, aggregation structures for hierarchical objects, versioning relations and compatibility with descriptions and conceptual schemas [16].

The launch of data.europeana.eu has set the first trial of making the Europeana metadata set available as Linked Open Data. Next developments will emphasize even more on the EDM and on a smooth transition from ESE to EDM. Therefore, contributors to Europeana who got used with the flat metadata structure of ESE, should be prepared to EDM entrance.

References

1. Bontchev, B., Jardim-Gonçalves, R.: Ontology-based content development in collaborative environments with semantic services, Proc. of Int. Conf. Collab. Technologies, ISBN 978-972-8939-21-2, 26-28 July 2010, Freiburg, Germany, 2010, pp. 183--188, IADIS (2010)
2. Ioannou, E., Niederée, C., Velegrakis, Y.: Enabling entity-based aggregators for web 2.0 data, Proc. of 19th Conf. on WWW, ISBN: 978-1-60558-799-8, pp.1119--1120, ACM, NY, USA (2010)
3. Concordia, C., Gradmann, S., Siebinga, S.: Not just another portal, not just another digital library: A portrait of Europeana as an application program interface, IFLA Journal, Vol. 36, No. 1, pp.61--69 (2010)
4. Meghini, C., Isaac, A., Gradmann, S., Schreiber, G.: The Europeana Data Model: tackling interoperability via modelling, DL.org Autumn School, Athens (2010)
5. Georgia, A.: Guidelines for AV content providers to make their contents available to Europeana, PrestoPRIME, D6.2.2 (2010)
6. Europeana Aggregators' Handbook, Edition 1, Europeana, May (2010)
7. Döring, U.: Intermediate report on DMG-Lib OAI-PMH interface, thinkMOTION, D2.1, http://www.thinkmotion.eu/wp-content/uploads/D2.1_draft.pdf (2012)
8. Clayphan, R.: Europeana Semantic Elements Specification, Europeana v1.0, Ver. 3.4 (2011)
9. Meghini, C., Clayphan, R., Isaac, A.: Definition of the Europeana Data Model Elements, Version 5.2.3, Europeana v1.0 (2012)
10. Europeana v1.0 Metadata Mapping & Normalisation Guidelines for the Europeana Prototype, Version 1.2.1 (2010)
11. Europeana v1.0 Europeana Semantic Elements V3.4 XML Schema, Ver. 3.2, http://www.europeana.eu/schemas/ese/ESE-V3.4.xsd (2011)
12. Europeana Content Checker version 2.0 User Guide – revised for ESE v3.4, (2011)
13. Reis, D., Freire, N.: OAI-PMH implementation and tools guidelines, ECP-2006-DILI-510003, TELplus D-2.1 (2008)
14. Bergheim, S., McKenna, G., Urtegaard, G., Rowlatt, M., Rehak, R., Davies R., Clayphan, R.: EUROPEANALOCAL Technical Primer - Overview, Revised version (2008)
15. Haslhofer, B., Isaac, A.: data.europeana.eu - The Europeana Linked Open Data Pilot, Proc. of Int. Conf. on Dublin Core and Metadata Appl., ISSN: 1939-1366, pp.94--104 (2011)
16. Isaac, A., Clayphan, R.: Europeana Data Model Primer, Europeana v1.0 (2011)
17. Clayphan, R., Charles, V., Isaac, A.: Europeana Data Model Mapping Guidelines, Ver. 1.0.1, Techn. report, Europeana v1.0 (2012)
18. Gibson, T., Schuchardt, K., Stephan, E.: Application of named graphs towards custom provenance views, Proc. of TAPP'09, Art. No.5, USENIX, Berkeley, CA, USA (2009)