Undefined 0 (0) 1
IOS Press

Europeana Linked Open Data – data.europeana.eu

Antoine Isaac [a], Bernhard Haslhofer [b]
[a] *Europeana, The Hague, The Netherlands*
[b] *Cornell Information Science, USA*

Abstract.
Europeana is a single access point to millions of books, paintings, films, museum objects and archival records that have been digitized throughout Europe. The data.europeana.eu *Linked Open Data* pilot dataset contains open metadata on approximately 2.4 million texts, images, videos and sounds gathered by Europeana. All metadata are released under Creative Commons CC0 and therefore dedicated to the public domain. The metadata follow the Europeana Data Model and clients can access data either by dereferencing URIs, downloading data dumps, or executing SPARQL queries against the dataset. They can also follow the links to external linked data sources, such as the Swedish cultural heritage aggregator (SOCH), GeoNames, the GEMET thesaurus, or DBPedia. The latest dataset release has been published in February 2012.

Keywords:, Europeana, Linked Data, Libraries, Cultural Heritage

1. Introduction

Europeana is a single access point to millions of books, paintings, films, museum objects and archival records that have been digitized throughout Europe, gathered from hundreds of individual cultural institutions,[1] with the help of dozens of data aggregators and providers. The *Europeana Linked Open Data* pilot dataset contains open metadata on approximately 2.4 million texts, images, videos and sounds. These collections encompass more than 200 cultural institutions from 15 countries. They cover a great variety of heritage objects, such as a Slovenian version of *O Sole Mio* from the National Library of Slovenia,[2] or memories on the herring business from the Tyne and Wear Archives & Museums in Newcastle.[3]

Version 1.1 of the dataset, which is now available at `http://data.europeana.eu`, has been released in February 2012. The data is represented in the *Europeana Data Model (EDM)*, as we explain in more detail in Section 4. It is served according to the Linked Data principles: the described resources are addressable and dereferenceable by their URIs; especially, depending on its `Accept` parameter, an HTTP GET request against a `data.europeana.eu` URI leads either to an HTML page on the Europeana portal for the object it identifies or to raw, machine-processable data on this object. See `http://pro.europeana.eu/tech-details` for examples. The data is also available for bulk download at `http://data.europeana.eu/download/`, where the metadata are organized by dataset version, data provider, and RDF serialization format (RDF/XML, N-Triple). Clients can also execute structured queries against the publicly available SPARQL endpoint: `http://europeana-triplestore.isti.cnr.it/sparql`.

[1] Around 1500 institutions have contributed to Europeana including renowned names such as the British Library in London, the Rijksmuseum in Amsterdam and the Louvre in Paris but also many smaller cultural heritage organizations and libraries across Europe.
[2] `http://data.europeana.eu/item/92056/BD9D5C6C6B02248F187238E9D7CC09EAF17BEA59`
[3] `http://data.europeana.eu/item/09405f/533F9A826CB038D02C05A9814CF97E5D1B49BBEE`

2. Opening Cultural Data

data.europeana.eu is one of the results of more than one year of campaigning from Europeana to convince its community of opening up their metadata.[4] Currently it serves metadata coming from 8 data aggregators who have reacted early and positively to these efforts and agreed to publish their metadata under the Creative Commons CC0 Public Domain Dedication,[5] which means that "[Anyone] can copy, modify, distribute and perform the [data], even for commercial purposes, all without asking permission".

Including only a subset of the total Europeana collection, which encompasses more than 20M objects at the time of writing, is deliberate. In fact the first version of our dataset contained metadata for approximately 3.5M objects but the licensing was not explicit. With 2.4M objects in version 1.1 we clearly favored openness of metadata over quantity.

At the moment, data.europeana.eu serves as a prototype for unlocking metadata and rights on metadata, on a massive scale. In so-called hackathons (Hack4Europe[6]) developers can learn about this prototype and other access mechanisms to cultural data: Europeana also has an API and semantic mark-up on pages. We hope they will be used by third parties to develop innovative applications and services. This would in turn help to convince our partners to release more open data, next to other actions such as the release of an animation that bridges Linked Data technology with Open data policies[7].

3. Data Anatomy

3.1. Coverage

As said, Europeana aggregates metadata about more than 20M millions books, paintings, films, museum objects, archival records and other types of cultural objects. data.europeana.eu represents the "public domain" subset of the collections that can be accessed through Europeana. It currently holds metadata about 2,381,745 digitized objects, which were aggregated from 8 aggregators representing 221 individual institutions from 15 countries across Europe. Please note that the following statistics apply to this "open" subset of the total Europeana collection. We also excluded data about 4 objects, which were added to the dataset for illustrative purposes.

Table 1
Open data contribution by country.

Country	Number of objects
Spain	1,468,460
Norway	248,987
Austria	224,147
Sweden	102,850
Belgium	68,516
Denmark	45,041
Germany	40,729
Slovenia	40,281
United Kingdom	39,243
Ireland	33,651
Luxembourg	24,890
Serbia	16,852
Czech Republic	10,849
Italy	9,088
Portugal	8,161

In Table 1, which shows the "public domain" metadata contribution by country, we can clearly see that institutions from Spain, with 1.47M objects, are currently the major data contributors.

While the 10 largest data providers (see Table 2) contribute 80% of all data (1,902,380 objects), the remaining 20% (479,365 objects) are contributed by the 211 smaller institutions or come from collections for which we do not have explicit information on individual data providers, as is currently the case for the majority of Swedish objects. Two data providers even contribute only one single object to the current dataset.

These statistics show the importance of Europeana and intermediate data aggregators that contribute to it, such as http://hispana.mcu.es or The European Film Gateway. The distribution of data aggregation efforts allows unifying the access to objects from a huge diversity of institutions, with limited effort. The resources it takes to consume data available at an aggregator is much lower than the effort of setting up a solution at each data provider's side.

3.2. Data gathering, linkage, and processing

The process of preparing the data for data.europeana.eu has been described in a separate

[4] See Europeana's new Data Exchange Agreement and actions in support for open data at http://pro.europeana.eu/support-for-open-data

[5] http://creativecommons.org/publicdomain/zero/1.0/

[6] http://pro.europeana.eu/hackathons

[7] http://vimeo.com/36752317

Table 2
The 10 largest data providers and their aggregators.

Aggregator	Data Provider	Number of objects
Hispana	Biblioteca Virtual de Prensa Histórica	956,496
Norsk Kulturråd	Fylkesarkivet i Sogn og Fjordane	248,368
The European Library	Österreichische Nationalbibliothek - Austrian National Library	223,847
Hispana	Galiciana: Biblioteca Digital de Galicia	136,473
Hispana	Repositorio Biblioteca virtual de Andalucía	100,775
Hispana	Gredos (Universidad de Salamanca, Spain)	65,567
The European Film Gateway	Det Danske Filminstitut	45,041
Hispana	Biblioteca Digital de Madrid	44,825
The European Film Gateway	Deutsches Filminstitut - DIF	40,729
The European Library	National and University Library of Slovenia	40,259

technical paper [1]. The prototype is deployed directly on top of metadata that has already been gathered by Europeana, either via OAI-PMH servers or from batch files. These metadata are formatted according to the *Europeana Semantic Elements (ESE)* XML Schema,[8] which is essentially a flat record structure that uses the Dublin Core Element Set[9] with some Europeana extensions.

For the Europeana Linked Open Data set we converted this ESE metadata into the new *Europeana Data Model (EDM)*,[10] which has been developed with a much stronger Linked Data focus. We thus defined a mapping[11] between ESE and EDM and implemented it as an executable ESE-EDM transformation library,[12] which can be applied on the legacy ESE data.

Parallel to this, we currently follow two strategies for linking `data.europeana.eu` resources with other Web resources: first, we fetch *semantic enrichment* data that is being created by Europeana, after it has ingested metadata from its data providers. This data consists of links to four types of reference resources:[13] Geonames for places (1.7M links), GEMET for general topics (863K links), the Semium time ontology for time periods (1.9M links), and DBpedia for persons (1304 links). Since the enrichments are links they perfectly fit EDM and Linked Data approach, as seen in the following section. Second, as a simple ad-hoc linking strategy, we rely on existing resource identifiers that are part of the metadata and create links to other Linked Open Data services, which hold information about objects that are also served by `data.europeana.eu`: for the moment this only concerns the Swedish cultural heritage aggregator (SOCH).

At the moment we manually execute the ESE-EDM transformation and fetch the enrichment data whenever we release a new dataset version and ingest the resulting RDF data into a separate triple store. This is clearly a temporary solution, only suitable for a pilot. In the long term, all human- and machine-readable Europeana interfaces, including the Linked Data one, should be directly fed from one single data repository.

4. EDM data modeling patterns

For publishing metadata at `data.europeana.eu`, we "upgrade" ESE data to the Europeana Data Model (EDM), which has been developed by the Europeana community and is a more flexible and precise model. It offers the opportunity to attach every statement to the specific resource it applies to and also reflects some basic form of data provenance. The main EDM requirements include:

- distinguish between a "provided item" (painting, book) and digital representations
- distinguish between an item and the metadata record describing it
- allow ingesting multiple records for a same item, containing potentially contradictory statements about it

EDM allows to represent different perspectives on a given cultural object. It also enables to represent complex, especially hierarchically structured objects as in

[8] `http://pro.europeana.eu/technical-requirements`
[9] `http://dublincore.org`
[10] `http://pro.europeana.eu/edm-documentation`
[11] `http://europeanalabs.eu/wiki/EDMPrototypingTask15`
[12] `https://github.com/behas/ese2edm`
[13] Accessible respectively at `http://www.geonames.org`, `http://www.eionet.europa.eu/gemet/`, `http://semium.org` and `http://dbpedia.org`

the archive or library domains. Finally, it allows us to represent contextual information, in the form of entities (places, agents, time periods) explicitly represented in the data and connected to a cultural object.

In the following we explain in more detail the basic structure of EDM networked resources, which is shown in Figure 1, together with the properties we expect to be applied to their instances. Further information, including dereferencable example resources are available at `http://pro.europeana.eu/web/guest/tech-details`.

4.1. Item (Provided Cultural Heritage Object)

Item resources (typed as Provided Cultural Heritage Object (CHO)) represent objects (painting, book, etc.) for which institutions provide representations to be accessed through Europeana. Provided CHO URIs are the main entry points in `data.europeana.eu`. A Provided CHO is the hub of the network of relevant resources. When applicable (see Section 3.2), the URIs for these objects link, via `owl:sameAs` statements, to other linked data resources about the same object. In our pilot, no descriptive metadata (creator, subject, etc.) is directly attached to object URIs. It is instead attached to the proxies that represent a view of the object, from a specific institution's perspective (either a Europeana provider or Europeana itself, see below). Depending on the feedback received during this pilot, we may change this and duplicate all the descriptive metadata at the level of the item URI. Such an option is costly in terms of data verbosity, but it would enable easier access to metadata, for data consumers less concerned about provenance.

4.2. Provider's proxy

Proxies originate from the OAI-ORE model [2] and are used as subjects of descriptive statements (creator, subject, date of creation, etc.) for the item, which are contributed by a Europeana provider. They enable the separation of different views for a same resource, in the context of different aggregations. This allows us to distinguish the original metadata for the object from the metadata that is created by Europeana. Descriptive properties that apply to these proxies, as we can generate them from ESE metadata (see Section 3.2) mostly come from Dublin Core. Proxies are connected to the item they represent a facet of, using the `ore:proxyFor` property. They are attached to the aggregation that contextualizes them, using the `ore:proxyIn` relationship. This design was chosen because of the lack of support for named graphs (aka "quadruples") in the RDF standard. OAI-ORE introduced Proxies in order to support referencing resources in the context of a specific graph. Eventually, named graphs may be natively supported by RDF, which could supersede the Proxy construct.

4.3. Provider's aggregation

These resources provide data related to a Europeana provider's gathering of digitized representations and descriptive metadata for an item. They are related to digital resources about the item, be they files directly representing it (`edm:object` and `edm:isShownBy`) or web pages showing the object in context (`edm:isShownAt`). They may also provide controlled rights information applying to these resources (`edm:rights`). Finally, provenance data is given in statements using `edm:provider` (the direct provider to Europeana in the data aggregation chain) or `edm:dataProvider` (the cultural institution that curates the object). The aggregation is connected to the item using the `edm:aggregatedCHO` property.

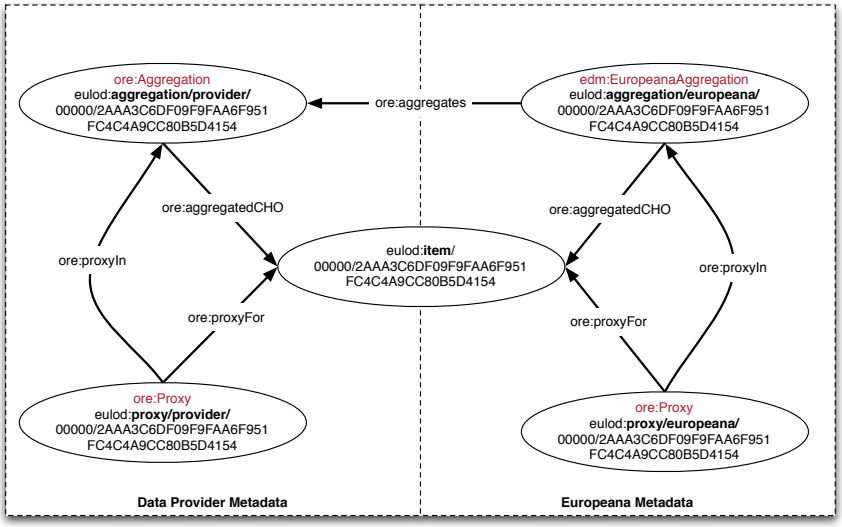
4.4. Europeana's proxy

Europeana proxies are the second type of proxies served at `data.europeana.eu`. They provide access to the metadata created by Europeana for a given item, distinct from the original metadata from the provider. Here one can find `edm:year` statements, indicating a normalized date associated with the object. Proxies also have statements that link them to places, concepts, persons and periods from external datasets, as mentioned in section 3.2. Finally, a proxy is connected to the item it represents a view of, using the `ore:proxyFor` property, as well as to the aggregation that contextualizes it, using `ore:proxyIn`.

4.5. Europeana's aggregation

A Europeana aggregation bundles together the result of all data creation and aggregation efforts for a given item. It aggregates the provider's aggregation (using `ore:aggregates`), which in turn will connect to the provider's proxy. Next to the provider aggregation, one can find the digitized resources `europeana.eu` serves for the item, i.e., an object page (`edm:landingPage`) and a thumbnail (using a combination of `edm:hasView` and

Fig. 1. Basic structure of EDM networked resources.



`foaf:thumbnail`). The Europeana proxy is also connected to this aggregation, as mentioned above.

4.6. Resource map

OAI-ORE Resource maps are constructs for indicating meta-level statements about the creation and publication of ORE data (ORE aggregations and their aggregated resources). We are exploring their use as a contextualization mechanism for the Europeana aggregation. Maps are connected to an item they are about using `foaf:primaryTopic`, and to its corresponding Europeana aggregation using `ore:describes`. They sum up the provenance of metadata using `dc:creator` and `dc:contributor` statements. Crucially, they also indicate, in a machine-readable way, that the `data.europeana.eu` RDF dataset is provided under the CC0 open license.

4.7. Vocabulary usage and interoperability

EDM is well connected to other established ontologies, most notably the Dublin Core metadata elements, SKOS and OAI-ORE. We have tried to directly re-use elements from these vocabularies whenever this was possible. When not, the newly introduced elements are semantically aligned to these ontologies, either using simple RDFS class and property specialization or OWL axioms. Such alignments allow for example to connect EDM to CIDOC-CRM, an important vocabulary for the museum domain.[14]

5. Known Shortcomings and discussion

Europeana is often confronted with the critique that its "data quality" could be enhanced. Especially, the "internal connectivity" of the dataset is currently very low. We have Provided CHO - aggregation - proxy relationships that come with the EDM model, but no "semantic" links between the items, or the proxies that represent them.

This is partly because the ESE metadata format, which is based on simple text fields, conceals the richness of the original metadata: many providers use contextual resources, which could be fed into Europeana and provide internal links. This includes, amongst others, concepts from shared domain thesauri, or place resources, which are already used in the description for different objects in a collection or even across collections. This contextual information is lost when the metadata is transferred to Europeana in ESE. We hope to obtain such valuable information from providers, when they can submit metadata in EDM. Europeana is currently working on it and we have case studies[15] that demonstrate how this can be done and what are the benefits. In the Amsterdam Museum Linked Open

[14] `http://www.cidoc-crm.org`
[15] `http://pro.europeana.eu/edm-case-studies`

Data prototype[16], for instance, richer original metadata has been converted to EDM and published as Linked Open Data, together with its companion thesaurus and authority file.

For achieving "external connectivity", we currently rely on Europeana's enrichment process (see Section 3.2), which generates semantic links from specific fields in the ESE data (e.g., Dublin Core's `dc:subject`), but that information is not recorded. As a result, we do not know whether, say, a given city is the subject of an item or its place of production. For our RDF data we had to use an EDM property that merely expresses that the item is "generally linked" to that place. Because it has to deal with very heterogeneous collections, Europeana is bound, for the moment, to using simple data enrichment techniques, which we know will bring errors. Still, we can do better at handling the provenance of enrichments to obtain a better data grain.

Another issue is the transition to the network model of EDM, which lead to quite verbose data. We may want to "hide" this complexity when it is not needed or reveal the full complexity and power of EDM in successive steps, which should make the full picture easier to understand for data providers and consumers alike. This important lesson learnt has directly influenced how EDM should be used for data ingestion into Europeana, i.e., with only a limited part of the pattern used for our pilot. But it is still open, whether and how `data.europeana.eu` should handle the complexity differently, as a data *publication* service.

Finally, we needed to start addressing design issues that the existing EDM specification had not touched at all. The first one is the minting of HTTP Uniform Resource Identifiers (URIs) for all EDM resources in a Linked Data environment. We realized that many patterns were possible, each corresponding to slightly different priorities in terms of representing the underlying model or enabling certain HTTP-based services. The second issue is the representation of provenance for the metadata served on `data.europeana.eu`, including such things as attribution or licenses. All the provenance information available at Europeana could be represented. The way it has been represented, though, may be revisited in the light of ongoing discussions in the community.

[16] `http://semanticweb.cs.vu.nl/lod/am` — a paper has been submitted to this Semantic Web Journal special call

6. Summary and Future Work

With `data.europeana.eu` we created a Linked Data prototype for Europeana, which is a single access point to millions of cultural digital objects that have been digitized throughout Europe. At the moment, it serves metadata of 2.4M objects under the Creative Commons CCO public domain dedication. The data originate from aggregators and providers who have reacted early and positively to Europeana's new Data Exchange Agreements. One future work goal is to convince more data providers to accept these agreements and to increase the number of objects included in Europeana's Linked Open Data service.

The exposed metadata are represented in the *Europeana Data Model (EDM)*, which has been developed by the Europeana community and allows to represent different perspectives and basic provenance information on a given cultural object. We expect that future `data.europeana.eu` dataset releases reflect the lessons we have learned with respect to the model's complexity and identification of digital objects. We will also investigate how to align the EDM with other efforts dealing with provenance on the Web, such as PROV Model developed by the W3C Provenance Working Group[17].

Increasing Europeana's internal and external connectivity by means of links between Web resources is another major goal. This can be achieved by convincing data providers to deliver their original rich metadata instead of flat ESE records and by applying named entity linkage techniques in the data ingestion phase.

References

[1] B. Haslhofer and A. Isaac, *data.europeana.eu - The Europeana Linked Open Data Pilot*, International Conference on Dublin Core and Metadata Applications, 2011, The Hague, NL.

[2] C. Lagoze and H. van de Sompel (eds.), *http://www.openarchives.org/ore/1.0/primer.html*, Available at: `http://www.openarchives.org/ore/1.0/primer.html`, Accessed: 2012-05-20.

[17] `http://www.w3.org/TR/prov-primer/`