



→ EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy: final report

Contributors

Agnès Simon, Bibliothèque Nationale de France
Daniel Vila Suero, Universidad Politécnica de Madrid
Eero Hyvönen, Aalto University
Esther Guggenheim, National Library of Israel
Lars G. Svensson, Deutsche Nationalbibliothek
Nuno Freire, The European Library
Rainer Simon, AIT Austrian Institute of Technology
Rodolphe Bailly, Musical Instrument Museums Online
Roxanne Wyns, LIBIS
Seth van Hooland, Université Libre de Bruxelles
Shenghui Wang, Online Computer Library Center
Vladimir Alexiev, Ontotext

Editors & Task Force Chairs

Juliane Stiller, Humboldt-Universität zu Berlin
Antoine Isaac, Europeana Foundation
Vivien Petras, Humboldt-Universität zu Berlin

Executive Summary	3
1. Introduction	4
2. Enrichments in Europeana	6
3. Use Cases	7
3.1. Rijksmuseum (dataset identifier: 90402)	7
3.2. Saxon State and University Library (01004)	10
3.3. Instituto de Archaeologia Iberica Universidad de Jaen - CARARE (2020715)	13
3.4. HISPANA (2022703)	15
3.5. Hungarian Jewish Archives (09315)	17
3.6. Europeana 1914-1918	19
3.7. Bernstein Collection (09802)	20
4. Results and Recommendations	24
4.1. Metadata Quality	25
4.2. Vocabularies	28
4.3. Enrichment Process	28
4.4. Further Ideas	29
5. Conclusions	30
Acknowledgements	31
References	31
Appendix 1 – Europeana Enrichment Process	32
Appendix 2 – Broader GEMET Terms	36
Appendix 3 – Vocabularies	40

Executive Summary

The semantic and multilingual enrichment of metadata in Europeana is a core concern as it improves access to the material, defines relations among objects and enables cross-lingual retrieval of documents. The quality of these enrichments is crucial to ensure that highly curated content from providers gets represented correctly across different languages. To ensure that those enrichments unfold their whole potential and act as facilitators of access, a semantic and multilingual enrichment strategy is needed. The EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy set out to analyze datasets in Europeana and to evaluate them with regard to their enrichment potential and the enrichments that were executed. The goal was to drive a strategy for enriching metadata fields adding value for users. To achieve this, the members of the task force held a one-day workshop in Berlin where they analyzed randomly selected datasets from Europeana, their metadata fields and their enrichment potential.

This report aggregates the results and derives findings and recommendations regarding the metadata quality (source), vocabulary used (target) and the enrichment process. It was found that especially during mapping and ingestion time, metadata quality issues arise that influence the success of the enrichments. Tackling these issues with better documentation, training and the establishment of quality scores are some of the recommendations in this field. Furthermore, Europeana should encourage the delivery of specialized vocabularies with resolvable URIs which would also lead to less need for enrichments by Europeana itself. With regard to the enrichment process, clear rules for each field need to be established.

1. Introduction

Semantic enrichment aims at adding new information at the semantic level to the data about certain resources. This is a rather vague notion, which has different interpretations depending on the disciplinary context. For example, in the Linked Data context, it chiefly refers to the creation of new links between the enriched resources and others, preferably coming from an existing, reference dataset. In Information Retrieval, it means adding new terms to a query or document and therefore reaching a higher visibility of documents within the document space.

In the context of Europeana and its related projects, many approaches and tools have been proposed under this notion. Within the EuropeanaTech network, a FLOSS (free/libre, open source software) inventory¹ was made available with numerous tools for enriching metadata. Furthermore, in the same project, a market study was conducted which gives an overview of tools for semantic extractions (Olensky, 2012).²

Inspired by a recent report from the PATHS project, we can identify main (practical) semantic enrichment categories (Stevenson et al, 2013):

- Identification of Key Entities: Named Entity Extraction can be used to flag the terms that correspond to important persons, places or events in existing metadata.
- Item Similarity: the creation of links between pairs of resources of a same type in a same dataset or collection, based on similarity. They can be qualified by a measure and/or a type.
- Background Links: the creation of links between resources in a dataset and external resources, such as a controlled vocabulary (thesauri and other knowledge organization systems) or Wikipedia. In Europeana, the representations of cultural heritage data should be interlinked and contextualized with semantic network resources (Gradmann, 2010). Within the project EuropeanaConnect³, many experiments were conducted on how to leverage these links for semantic search in Europeana and a prototype was delivered (de Boer et al, 2010). It can be accessed in the ThoughtLab⁴.
- Groupings: the grouping of resources along a common origin, shared themes or general similarity measures. These groupings may be based on the approaches mentioned in the two previous points. They may result from the linking to a (hierarchical) semantic vocabulary, or based on various levels of object-metadata similarity. A study on semantic clustering based on hierarchically structuring Europeana objects was conducted giving an analysis of the categories that emerged (Wang et al, 2013).

Europeana enrichment denotes a composite process made up of two main steps:

1. Matching the metadata of Europeana objects to external semantic data resulting in links between these objects and resources from external (reference) datasets, which can be also re-used by users of the Europeana data services (e.g., API). For example, http://europeana.eu/portal/record/2020715/uid_iid_3493855_HA_4013.html, a 'recipiente cerámico' is matched with the concept of 'ceramics' in the GEMET thesaurus: <http://www.eionet.europa.eu/gemet/concept/1266>.

¹ https://docs.google.com/spreadsheets/ccc?key=0Ag_7rVJw0CpdFRJOEJxdEk4ZEMxQ01jaDgxQXFSTkE

²

http://ec.europa.eu/information_society/apps/projects/logos//2/270902/080/deliverables/001_DeliverableD74MarketStudyToolsSemExtractfinal2.pdf

³ <http://www.europeanaconnect.eu/>

⁴ <http://pro.europeana.eu/thoughtlab/linked-open-data>

2. The exploitation of these links through adding data to the index behind the Europeana.eu portal that enhances the user's experience by triggering semantic or multilingual enrichments to retrieve documents that match a given query. That means for the example given above that the record Solr⁵ indices behind Europeana.eu are supplemented with all the translated labels of the GEMET concept (e.g., 'céramique'), as well as with a link to the broader concept in GEMET (<http://www.eionet.europa.eu/gemet/concept/4260>) and all its translated labels (e.g., 'industrial product').

The semantic and multilingual enrichment of metadata in Europeana is a core concern as it improves access to the material, defines relations among objects and enables cross-lingual retrieval of documents. The quality of these enrichments is crucial to ensure that highly curated content from providers gets represented correctly across different languages. To ensure that enrichments unfold their whole potential and act as facilitators of access, a semantic and multilingual enrichment strategy is needed, which targets the following points:

- datasets and their objects,
- metadata fields, and
- controlled vocabularies.

The main work of the task force consisted of studying collections and their metadata fields as cases for developing a multilingual and semantic enrichment strategy and identifying possible pitfalls in the enrichment process.

The activities of the task force included:

- choosing multilingual datasets from Europeana which can serve as cases for an implementation of a semantic and multilingual enrichment strategy,
- gathering vocabularies for enrichment that are open and multilingual, and
- holding a workshop with all participants to deliver hands-on results by developing an enrichment strategy for each use case and overall for Europeana by
 - analyzing the metadata fields and their potential for enrichment, which included assessing their semantic meaning and possibly their ambiguity,
 - choosing controlled vocabularies that suit these fields, and
 - suggesting rules to ensure high quality enrichments.

The one-day workshop took place on November 8, 2013 and was attended by the organizers and the task force members.

From the use cases, areas of concern for enrichment were identified and strategic recommendations for enrichment were derived. Here, the focus was on actions Europeana can take to improve the enrichment quality and find measures to prevent errors in upcoming enrichments.

The report is structured as follows: section 2 describes the enrichment process in Europeana, the vocabularies used and the rules applied. Section 3 describes the findings of the use case analysis during the workshop; section 4 summarizes the findings and gives recommendations for Europeana to implement a better semantic and multilingual enrichment strategy. Section 5 concludes the report.

⁵ <https://lucene.apache.org/solr/>

2. Enrichments in Europeana

For enriching metadata, Europeana currently uses the Annocultor⁶ tool. As mentioned in the introduction, Europeana enrichments focus on *linking* Europeana objects to other resources.

The main components of the enrichment process can be described as:

- the *target of enrichments*, i.e. the set of resources to which objects in Europeana are linked;
- the *source of enrichment*, i.e. the fields in EDM data from which the links are derived - mostly by matching the string value of these fields to the labels of the contextual resources;
- a *rule* that specifies how a match is obtained between the source metadata field(s) and the (labels of the) target contextual resources.

Currently Europeana enriches objects by creating links to places from the GeoNames⁷ dataset, concepts from GEMET⁸ (GEneral Multilingual Environmental Thesaurus), agents that mostly come from DBpedia⁹ and time periods from the adhoc Semium Time vocabulary¹⁰.

More details on the enrichment, including the number of objects that are enriched for each category, can be found in Appendix 1.

⁶ <http://sourceforge.net/projects/annocultor/>

⁷ <http://www.geonames.org/>

⁸ <http://www.eionet.europa.eu/gemet/>

⁹ <http://dbpedia.org/About>

¹⁰ <http://semium.org/time/>

3. Use Cases

In this section, the analysis of six datasets in the Europeana portal is described. The purpose of the task force is to give recommendations on how to achieve better enrichments of the Europeana data, therefore we chose datasets without contextual resources from providers because they have more potential value for enrichment by Europeana itself. This has resulted in picking the less representative sets for projects like CARARE or MIMO. Other criteria were the objects described, the language of the metadata and the richness of the metadata. The sample is exploratory and purposeful and does not reflect Europeana datasets in their entirety.


Each dataset was analyzed by two workshop participants. The task was

- to analyze the usage of each field in relation with its semantics,
- to study the multilinguality aspects of the dataset, and
- to determine what kind of enrichments were undertaken, why they possibly failed and how they could be improved.

The participants only used the portal and its search functionality to derive their conclusions. The focus was on getting a view on the data as a user would see it, no other tools were used.

3.1. Rijksmuseum (dataset identifier: 90402)

This dataset has 111,657 objects in total. Figure 1 shows a typical object represented in the Europeana portal. Table 1 gives an overview over the enrichments in this dataset.



Public Domain marked

View item at Rijksmuseum

Share

Cite on Wikipedia

Translate details

Select language

Powered by Microsoft Translator

Theepot met deksel, veelkleurig beschilderd met chinoiserieën

Description: Bolvormige theepot van beschilderd porselein. De pot heeft een C-vormig oor, een tuit en een gewelfd deksel. De tuit ontspringt uit een masker in reliëf. De tuit heeft een verguld zilveren montuur met een klep. Het deksel is gevat in een verguld zilveren montuur. De pot is beschilderd met Hóroldt-cinoiserieën binnen cartouches. De theepot hoort bij een theeservies (BK-17420-A t/m BK-17420-0) en is gemerkt. Het servies staat op een stellage of onderstel (BK-17017).

Creator: porseleinfabriek: Meissener Porzellan Manufaktur

Geographic coverage: Meissen; <http://sws.geonames.org/2872155/>

Date: tweede kwart 18e eeuw

Date of creation: ca. 1725 - ca. 1730

Type: eet- en drinkgerei; theepot

Format: image/jpeg; geheel: hoogte 11.9 CMcmcm; geheel: breedte 17.0 CMcmcm; geheel: diameter 10.7 CMcmcm; porselein; zilver

Subject: Iconclass code: 48A98211; Iconclass code: 48A9876

Identifier: BK-17420-A; RM0001.COLLECT.315793

Is part of: BK-17420

Language: nl

Publisher: Rijksmuseum, Amsterdam

Data provider: Rijksmuseum

Provider: Rijksmuseum

Providing country: Netherlands

Auto-generated tags

Where

Place Term: <http://sws.geonames.org/2921044/>

Place Label: [德國] (zh); [germania] (ro); [tyskland] (no); [németország] (hu); [vokietija] (lt); [jerman] (id); [bundesrepublik deutschland, deutschland] (de); [de, deutschland, federal republic of germany] (def); [saksa] (fi); [allemagne] (fr); [tyskland] (sv); [германия] (bg); [nemčija] (sl); [německo] (sk); [німецьчина] (uk); [tyskland] (da); [germania] (it); [německo] (cs); [уърловия] (el); [germania] (la); [alemania] (pt); [niemcy] (pl); [federal republic of germany, germany] (en); [германия] (ru); [saksamaa] (et); [alemania] (es); [deutschland] (nl)

Place Term: <http://sws.geonames.org/2872155/>

Place Label: [misnia] (it); [meißen, mißeñ] (cs); [meißen] (de); [meißen] (def); [meißen] (pt); [misnia] (la); [misnia] (pl); [meissen, meißen] (sv); [meißen] (fr); [meissen, meißen] (en); [meißen] (ru); [meißen] (et); [meißen] (es); [meißen] (nl)

Search also for:

Title
Theepot met deksel, veelkleurig beschilderd met chinoiserieën (2)

Who
porseleinfabriek: Meissener Porzellan Manufaktur (1006)

What
eet- en drinkgerei (848)
theepot (320)
Iconclass code: 48A98211 (123)
Iconclass code: 48A9876 (293)
image/jpeg (2815060)

Provider
Rijksmuseum (111657)
Rijksmuseum (111657)

Figure 1: Example object of the dataset Rijksmuseum.

Enrichment	Number of Objects
Agent enrichment	0
Place enrichment	67,525
Timespan enrichment	0
Concept enrichment	16,739

Table 1: Number of enrichments in dataset Rijksmuseum.

3.1.1. Findings

Per record, there were approximately 10-15 metadata fields covered. The analysis of these fields follows in more detail below. The subheadings refer to the displayed name fields that can but do not have to match the Dublin Core (DC) `field` name in the metadata provided.

Title & Description

The differentiation between title and description, which are often very similar in the museum context, was seen as problematic during the analysis. The title often reflects object names or a physical description of the object that is then repeated in the description field in a more complex fashion¹¹.

Creator & Contributor¹²

For the creator field, no enrichment worked as the field contains not only person names but also values describing different actor roles, e.g. "uurwerkmaker: Montjoye, Louis; schilder: Dodin, Charles". The specific roles and the associated person names are separated by a semicolon, person name and role are distinguished with a colon. No enrichment of actors was executed as the distinction between the role and the name is not possible.

Solving this issue by asking more granular data to the provider is tricky: during the mapping to EDM, it is impossible to express the roles of persons involved. In the `dc:creator` field, the roles could be removed but then this information would be lost. A specialization of `dc:creator` could be introduced by an EDM extension or a "creation event" with different persons attached with their role (which also requires an EDM extension)¹³.

It may be possible to employ some basic regular expression-based techniques to extract artists' names, e.g. using the presence of colons as in "schilder: Dodin, Charles"¹⁴. But in the `dc:contributor` field, for example, there is often too much natural language in front of the actual name of the contributor to extract the person name for enrichment, e.g. "Schenking van de ervan van de heer A. Isaac"¹⁵.

¹¹ One example is "Snuifdoos van goud, rechthoekig, versierd met bloemen- en vruchtenslingers in émail [...] in rocaille-omlijsting" http://europeana.eu/portal/record/90402/collectie_BK_17138.html.

¹² So far, the Rijksmuseum was not part of the two ingestion cycles which enriched objects with Dbpedia artists.

¹³ The task force on EDM mappings, refinements and extensions used the following definition for an EDM extension: "An extension to EDM is required when existing EDM classes and properties cannot represent the semantics of providers' data with sufficient details." More info: <http://pro.europeana.eu/web/network/europeana-tech/-/wiki/Main/Task+force+on+EDM+mappings+refinements+and+extensions>

¹⁴ http://europeana.eu/portal/record/90402/collectie_BK_16672.html

¹⁵ http://europeana.eu/portal/record/90402/collectie_BK_1955_395.html

Geographic Coverage & Date

In the geographic coverage field, the granularity is often not indicated, e.g., in Dutch “Venetie”¹⁶ can indicate the city of Venice or the province of Venice. In the given example, “Venetie” was mapped to the province of Venice in GeoNames.

The fields for the date and the date of creation often contain the same information in different formats, e.g. `dc:date`: “derde kwart 17e eeuw; vierde kwart 17e eeuw” and `dcterms:created`: “ca. 1675 - ca. 1675”.¹⁷ No enrichments for timespan were executed in this dataset.

Subject

74,426 objects have a subject field whose values are Iconclass¹⁸ codes (notations) such as “43A4316”¹⁹ without a URI that would allow to fetch machine-readable metadata for the concept (here, <http://iconclass.org/43A4316>, with machine-readable data at <http://iconclass.org/43A4316.rdf>).

The original metadata shows the same without the keywords attached to the codes (notations). At the Rijksmuseum, these might be indexed and might be searchable, but they are lost for Europeana and are not understandable for the user. Europeana could match the codes of Iconclass to the appropriate values (keywords) or the provider gives the full URIs²⁰. As it is now, the subject field does not offer the possibility for Europeana to retrieve the multilingual keywords that could populate its indexes for the good of the user.

Format & Type

The field `dc:type` often mixes physical description such as “bottle” with art forms such as “sculpture”. The format field combines the display of dimensions and formats of the digital representation and object e.g. “*image/jpeg; geheel: lengte 25.5 CMmcm; geheel: breedte 3 CMmcm; geheel: hoogte 4.3 CMmcm; papier; zijde; papier*”²¹. The values of the field in the metadata are coming from `dc:format`, `dcterms:extent` and `dcterms:medium`. The measuring units are unnecessarily repeated here in `dcterms:extent`. The metadata would support enrichments here, e.g. for the `dcterms:medium` field.

3.1.2. Recommendations

The general recommendation would be to improve the metadata overall²². One suggestion is to ask data providers to not map different values to the same field and keep originally different fields separated. When referring to identifiers of authorities, strings should be avoided and URIs should be used instead which can then be de-referenced by Europeana.

Additionally, if there are two actor roles, one could be to deliver one actor role with the actor name in different fields. EDM can handle actor roles which can result in a separation of actor name and person name.

¹⁶ http://europeana.eu/portal/record/90402/collectie_RP_P_H_H_1184.html

¹⁷ http://europeana.eu/portal/record/90402/collectie_SK_A_2098.html

¹⁸ <http://iconclass.org>

¹⁹ http://europeana.eu/portal/record/90402/collectie_RP_P_OB_11_076.html

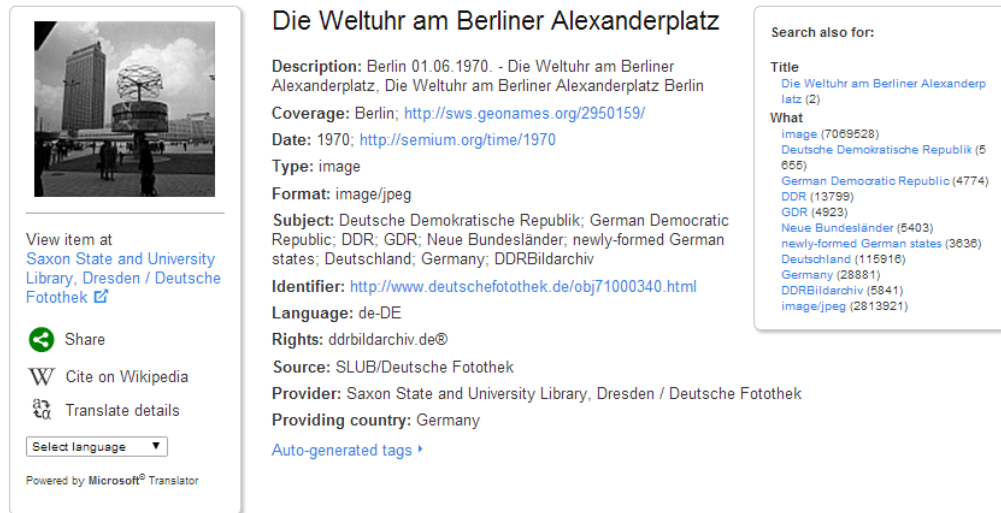
²⁰ The standard procedures for these enrichments is the provision of the full URIs by the provider, which are then de-referenced mapping the Iconclass metadata to the field `edm:concept`.

²¹ http://europeana.eu/portal/record/90402/collectie_BK_1960_56_B.html

²² The rich metadata of the provider could have been accommodated better with a mapping to EDM than to ESE (with auto-conversion to EDM).

3.2. Saxon State and University Library (01004)

This dataset has a total of 1,104,117 objects with metadata mainly in German. 1,097,320 have a `dc:description`, 889,961 have a `dc:coverage`, 741,727 have a `dc:creator` field. All of them have title, dates, format, identifier, subject fields etc. Figure 2 shows a typical object represented in the Europeana portal. Table 2 gives an overview over the enrichments in this dataset.



Die Weltuhr am Berliner Alexanderplatz

Description: Berlin 01.06.1970. - Die Weltuhr am Berliner Alexanderplatz, Die Weltuhr am Berliner Alexanderplatz Berlin

Coverage: Berlin; <http://sws.geonames.org/2950159/>

Date: 1970; <http://semium.org/time/1970>

Type: image

Format: image/jpeg

Subject: Deutsche Demokratische Republik; German Democratic Republic; DDR; GDR; Neue Bundesländer; newly-formed German states; Deutschland; Germany; DDRBildarchiv

Identifier: <http://www.deutschefotothek.de/obj71000340.html>

Language: de-DE

Rights: ddrbildarchiv.de©

Source: SLUB/Deutsche Fotothek

Provider: Saxon State and University Library, Dresden / Deutsche Fotothek

Providing country: Germany

[Auto-generated tags ▸](#)

Search also for:

Title
Die Weltuhr am Berliner Alexanderplatz (2)

What
image (7089528)
Deutsche Demokratische Republik (565)
German Democratic Republic (4774)
DDR (13799)
GDR (4923)
Neue Bundesländer (5403)
newly-formed German states (3636)
Deutschland (115916)
Germany (28881)
DRRBildarchiv (5841)
image/jpeg (2813821)

View item at [Saxon State and University Library, Dresden / Deutsche Fotothek](#)

[Share](#)

[Cite on Wikipedia](#)

[Translate details](#)

Select language ▾

Powered by Microsoft® Translator

Figure 2: Example object from the collection Saxon State Library.

Enrichment	Number of objects
Agent enrichment	1,228
Place enrichment	698,978
Timespan enrichment	761,604
Concept enrichment	852,635

Table 2: Number of enrichments in the dataset Saxon State Library.


3.2.1. Findings

The category system of the provider is not completely represented in Europeana. In general, the original documents at the provider are richer than the metadata shown in Europeana. Often, material and categorization are partly missing²³. The original has more textual description, more dates and a Google Maps localization. During the analysis, many objects were found with dead links or non-working identifiers²⁴.

There were also several cases where multiple objects were generated from a single one. In some cases the metadata is exactly the same, in others it might differ slightly. Many duplicate objects were found, which do not resolve at the original provider website (each object has a different ID and one of the objects gets resolved), see figures 3 and 4 for an example. These duplicates might therefore differ in the way they got enriched.

²³ compare the above images with the original record at <http://www.deutschefotothek.de/obj82061202.html>

²⁴ <http://europeana.eu/portal/record/01004/C9C54E9F7709B4C75DB103DD5B1E6E37F9A066DE.html>



View item at
[Saxon State and University Library, Dresden / Deutsche Fotothek](#)

Share

Cite on Wikipedia

Translate details

Select language

Powered by Microsoft® Translator

Hohe Brücke

Description: Hohe Brücke Pirna-Graupa
Creator: Dreßler, Klaus (Fotograf)
Coverage: Pirna-Graupa
Date: 1701; <http://semium.org/time/1701>
Type: image
Format: image/jpeg
Subject: Foto; Fotos; <http://www.eionet.europa.eu/gemet/concept/6205>
Identifier: <http://www.deutschefotothek.de/obj80061201.html>
Language: de-DE
Source: SLUB/Deutsche Fotothek
Provider: Saxon State and University Library, Dresden / Deutsche Fotothek
Providing country: Germany
[Auto-generated tags](#)


Search

Title
Hohe

Who
Dreßl

What
imag
Foto
Fotos
http://
conce
imag

Figure 3: First duplicate record with no enrichment for geographic coverage.²⁵



View item at
[Saxon State and University Library, Dresden / Deutsche Fotothek](#)

Share

Cite on Wikipedia

Translate details

Select language

Powered by Microsoft® Translator

Hohe Brücke

Description: Hohe Brücke Pirna
Creator: Dreßler, Klaus (Fotograf)
Coverage: Pirna; <http://sws.geonames.org/2853572/>
Date: 1701; <http://semium.org/time/1701>
Type: image
Format: image/jpeg
Subject: Brücken; Fotos
Identifier: <http://www.deutschefotothek.de/obj80061202.html>
Language: de-DE
Source: SLUB/Deutsche Fotothek
Provider: Saxon State and University Library, Dresden / Deutsche Fotothek
Providing country: Germany
[Auto-generated tags](#)

Search

Figure 4: Second duplicate record with an enrichment of geographic coverage. (Compare metadata fields description and identifier, where differences occur. The first record does not resolve at the provider website.)²⁶

²⁵ <http://europeana.eu/portal/record/01004/E63473164B7D77219B8E9E7BF44E9E19D8AEE740.html> The original object can be found here: <http://www.deutschefotothek.de/obj82061202.html>

²⁶ <http://europeana.eu/portal/record/01004/73A931E651DA2FBA5805DAB844A8EEEBBD464327.html>

Creator

An ambiguity is introduced by the creator of picture being recorded as the creator of the object, in the example of figure 3 the creator is the photographer and not the architect of the bridge. This problem might be solved in a new mapping to EDM where the metadata of the object could be separated from the metadata of the picture of the object.

Geographic Coverage & Dates

Figure 3 and Figure 4 show two duplicated objects that were enriched differently. In the second picture, the enrichment worked for “Pirna” in the `dc:coverage` field; in the first object the coverage is “Pirna-Graupa”, which is a district of Pirna. It might be worth to break up such geo-location names and enrich them word by word although in some languages this might be problematic.

Dates are sometimes wrongly enriched. The dates are correct at the provider, e.g. “1000 ante” but at Europeana the date shows: “1000” which was enriched with “1000 AD”²⁷. In some cases the date was not enriched because the date “8000 AD” does not exist (but “8000 BC” would have been correct)²⁸. The metadata of the provider shows for this record “8000ante/5000ante” in the date field.

Another issue is that the provider presents time periods in the manner: “1000/1250”. In Europeana, these dates were mapped to the first year occurring in the string, in this case “1000”. So, time periods become exact years due to incorrect mappings.²⁹ Another example, at the provider, the date field shows: “1601/1700” - meaning the 17th century. Europeana presents this period as 1601, so a specific year is displayed and enriched instead of a time span³⁰.

Subject

The provider delivers a `dc:subject` field with the value “Foto” or/and “Fotos” for 970,891 objects. 791,853 objects of the dataset were enriched in this field. 179,038 of the objects were not enriched because the terms in the `dc:subject` field were not found in the vocabulary, e.g. a value in the field is “Kunst und Kultur; Volkskunst; Ballett; Oper; Kontaktbogen; Fotos; Pressearchiv Höhne/Pohl”³¹ and none of the terms was matched. In contrast, 495,045 of the objects were enriched with the GEMET concept: “photograph”.

Often, the subject fields are rather filled with types than topical keywords, e.g. “Foto”, “Fotos”, “Malerei”, “Bild”, “Sonstiges”. The sophisticated category system from the provider is not completely represented. Some subjects were not contributed in the metadata. Table 3 shows four of the most often occurring concept enrichments from GEMET within this dataset and the associated broader terms. The enriched concept terms are of very generic nature.

Concept enrichment	Broader term	Number
photograph	documentation	506,506
architecture	human science	65,781
art	cultural heritage	16,892

²⁷ <http://europeana.eu/portal/record/01004/D30DB3F0DF6B80B7A664D3B911CEBF804DC902BA.html>

²⁸ <http://europeana.eu/portal/record/01004/13AE7D2367AD1EDDB88EA48CF6697BFF72EF95A1.html>

²⁹ <http://europeana.eu/portal/record/01004/032EB48DC28A8EA71E92670F86465D54B3621806.html>

³⁰ <http://europeana.eu/portal/record/01004/E023CE8C81114CAD272D73348D6293C6BF464C73.html>

³¹ <http://europeana.eu/portal/record/01004/7E31184EBD8FFA37CFEBC06847EBCDEB78149E5B.html>

ceramics	industrial product	213
----------	--------------------	-----

Table 3: concept enrichments from GEMET and their broader terms.

3.2.2. Recommendations

Outgoing links should be checked ideally by the provider and then by Europeana at ingestion time to make sure that they are resolvable. As the location is mapped by the original provider, it should be used for enrichment. For the enrichment of `dc:subject`, it would have been beneficial to stem the keywords before matching as often plural words are used where no match is found, e.g. “Foto” vs. “Fotos”. The original categories could really provide a huge pool for enrichment, although it would have been on the provider’s side to deliver these.

The mapping of the date fields is another issue in this dataset. The original objects are more complete. For the subject enrichment, it would be advisable to not only use the first words for enrichment. For date fields, the main problem is with date ranges and the cut off of “BC” during mapping. This could be solved by better mapping, as ranges are enabled in EDM with the class `edm:TimeSpan`. Another problem, much harder to solve, will be to handle approximate dates (“early 18th century”).

3.3. Instituto de Archaeologia Iberica Universidad de Jaen - CARARE (2020715)

This is a dataset which was provided to Europeana within the project CARARE, a best practice network for archaeological content. It contains 4,563 objects. Figure 5 shows a typical object represented in the Europeana portal. Table 4 gives an overview over the enrichments in this dataset.



View item at [Instituto de Arqueología Ibérica-Universidad de Jaén](#)

Share

Cite on Wikipedia

Translate details

Select language

Powered by Microsoft Translator

Imagen en color, CC5_617_1.
Recipiente cerámico procedente de la necrópolis ibérica de Castellones de Ceal (Hinojares, Jaén)

Description: Recipiente cerámico realizado a torno, de cocción oxidante. Tiene el borde exvasado, con el labio vuelto. El cuerpo es de perfil globular, y el pie está indicado. Al exterior presenta decoración pintada de color rojo formando bandas horizontales y paralelas que enmarcan motivos de semicírculos concéntricos alternando con ondulados verticales y paralelos. Diámetro de la boca: 8,2 cm. Altura: 12 cm. Bibliografía: CHAPA, T. PEREIRA, J. MADRIGAL, A. MAYORAL, V. (1997) La Necrópolis ibérica de Castellones de Ceal (Hinojares, Jaén) Consejería de Cultura Junta de Andalucía - Universidad de Jaén. Sevilla.

Geographic coverage: iid:3494160/SP.1; Castellones de Ceal; Castellones de Ceal

Date: S. IV a.C.

Time period: Ibérico

Type: image

Format: image/jpeg

Subject: cerámica; <http://www.eionet.europa.eu/gemet/concept/1266>

Identifier: 1091

Publisher: Colección CATA: <http://cata.cica.es/cata.html> Bibliografía: CHAPA, T. PEREIRA, J. MADRIGAL, A. MAYORAL, V. (1997) La Necrópolis ibérica de Castellones de Ceal (Hinojares, Jaén) Consejería de Cultura Junta de Andalucía - Universidad de Jaén. Sevilla.

Data provider: Instituto de Arqueología Ibérica-Universidad de Jaén

Provider: CARARE

Providing country: Spain

[Auto-generated tags](#)

Search also for:

Title
[Imagen en color, CC5_617_1. Recipiente cerámico procedente de la necrópolis ibérica de Castellones de Ceal \(Hinojares, Jaén\) \(1\)](#)

What
[image \(7069528\)](#)
[cerámica \(53089\)](#)
<http://www.eionet.europa.eu/gemet/concept/1266> (8885)
[image/jpeg \(2813921\)](#)

Provider
[Instituto de Arqueología Ibérica-Universidad de Jaén \(4866\)](#)
[CARARE \(2008809\)](#)

Figure 5: Example object from the collection Instituto de Archaeologia Iberica.

Enrichments	Number of Objects
Agent enrichment	0
Place enrichment	4
Timespan enrichment	0
Concept enrichment	4563

Table 4: Number of enrichments in the dataset Instituto de Archaeologia Iberica.

3.3.1. Findings

Different views of the same object are shown in Europeana as separate items (e.g. text, 3D PDF, image, video) where the type of the object is (most of the time) included in the title. As the metadata is identical, it is hard to see the difference between these items for users. Furthermore, the connection between the different items is lost in Europeana³². We were unlucky to have selected a CARARE dataset that seems to not implement the good practices for mapping representation set by the project.³³

Geographic Coverage & Dates

No geographic enrichment was found. The place names have two levels: a generic location and a very specific archaeological site. Only the former may be enriched. CARARE uses indeed identifiers (e.g., "iid:3493855/SP.1") in the `dcterms:spatial` fields, which come from their own list and are most of the time not enriched in Europeana. An explanation might be the fact that this dataset has been converted from EDM to the simple ESE (the format that predated EDM) for a first ingestion prior to EDM implementation on the Europeana side and then back from ESE to EDM, but via a default conversion procedure. The structure of the place data that initially came in the records has been lost in this round-tripping. A new mapping and publication could solve this issue as there would be real places associated to the object. However, the enrichment by Europeana could still fail, as it is based on human-readable labels in the fields attached to the object, and not on labels attached to EDM place resources associated with that object. The enrichment rules should be redesigned to fit such case, which becomes then the one of *aligning* different place lists.

Note that the situation here is different from what happens in other CARARE collections, where CARARE solved the problem upfront by providing human-readable labels next to the identifiers that the Europeana portal cannot yet display properly in its object pages. These collections have often been enriched³⁴. Note also that there are potential matches between CARARE's places and Pelagios' gazetteer.³⁵

There are no temporal enrichments. The original value of the date field coming from a field "cronologia" is split in Europeana in the `dc:date` and the `dcterms:temporal` (field "time period" in display) field³⁶. The time period is always "Iberico", whereas the dates follow specific conventions, e.g. date: "S. IV a. C." which seems to be very domain-specific. In the original source data, date and time period are wrapped into one statement e.g. "Periodo Iberico. S. IV a. C." To enrich these fields, existing vocabularies could be extended or specific vocabularies

³² http://europeana.eu/portal/record/2020715/uid_iid_3493855_HA_4013.html

http://europeana.eu/portal/record/2020715/uid_iid_3493855_DR_1287_1.html

http://europeana.eu/portal/record/2020715/uid_iid_3493855_DR_3880_2.htm

³³ <http://pro.europeana.eu/carare-edm>

³⁴ http://europeana.eu/portal/search.html?query=edm_place:*geonames*&qf=PROVIDER:CARARE

³⁵ <http://pelagios-project.blogspot.co.uk/2013/10/a-web-of-gazetteers.html>

³⁶ http://europeana.eu/portal/record/2020715/uid_iid_3493924_DR_5014_3.html

which can deal with this type of dates would be needed, ideally delivered by the provider with the domain knowledge.

Subject & Type

All objects (4563) in the dataset have a `dc:subject` field. All of them have the keyword “cerámica” which was also enriched with the GEMET thesaurus and worked well in this case. The broader term also enriches these objects with the term “Industrial product” which might be slightly off-topic.

For many objects, the `dc:type` field is enriched. The ambiguity of the term “application” which is the value in `dc:type` field for 1143 objects³⁷ leads to an enrichment with the term “enforcement” as application in French means enforcement. The broader term carries this mistake further as it is “administrative procedure”. The problem also comes from the language (the language of metadata in Spanish, but the label in GEMET that was matched is French). Similarly, the broader term for the type “video” does not add any value, as it is “documentation”. Here again, the match was correctly executed but in the post-enrichment the choice of using the broader term was made. The benefits of using broader terms should be evaluated and for this dataset it would probably make sense to abandon.

3.3.2. Recommendations

The Pleiades time vocabulary³⁸ could be used for enrichments of time periods but the URIs for that need to be delivered by the provider as Europeana has not the resources to match suitable vocabularies to datasets. The enrichments with the broader terms of GEMET do not help in this dataset as they are too general. Again, a more specific or restricted vocabulary would be needed here.

In this dataset, the mapping hurts. In principle, the `dc:format` value ('text') that is currently attached to the object (proxy) should be attached instead to the `WebResource`, which represents the digital representation and should carry such information: it is the representation that is textual, not the object itself. The trick would be then not to enrich the attributes of the `WebResource`, or enrich them with a very specific (technical/media) vocabulary. Note again that this stems from the fact that this dataset has been converted from the simple ESE (the format that predated EDM) into EDM via a default conversion procedure. A new mapping and publication could solve this issue.

Providers should state the language of their metadata or add a language tag to specific fields.³⁹ This would help Europeana to match the right language of the vocabulary entries to the fields.

3.4. HISPANA (2022703)

This is one of the datasets of one of the main Spanish providers, HISPANA. It contains 133,696 objects and the language of the metadata is Spanish only. Contrary to other HISPANA datasets, it doesn't include URIs in the `dc:creator` and `dc:subject` fields. Figure 6 shows a typical object represented in the Europeana portal. Table 5 gives an overview over the enrichments in this dataset.

³⁷

http://europeana.eu/portal/search.html?query=+europeana_collectionName%3A2020715*+AND+cc_skos_pref_Label%3Aapplication

³⁸ <http://pleiades.stoa.org/vocabularies/time-periods>

³⁹ <http://pro.europeana.eu/data-multilinguality>



© Free Access - Rights Reserved

View item at
[CER.ES: Red Digital de Colecciones de museos de España](#)

Share

Cite on Wikipedia

Translate details

Select language

Powered by Microsoft® Translator

Autómata

Description: Pato con patas y pico naranja, cabeza verde, cuerpo rojo, alas grises, rojas, marrones claritas, verdes y colita verde. Mecanismo de cuerda

Creator: Georg Köhler (Nuremberg) (act. 1932-1960[ca])

Geographic coverage: Alemania; <http://sws.geonames.org/2921044/>

Time period: 1946-1954

Date of creation: 1946-1954

Type: Autómata

Format: Altura = 6,50 cm; Anchura = 9 cm; Profundidad = 4 cm; Hojalata

Subject: Autómata; Juguetes; Hojalata; Pato; Georg Köhler (Nuremberg)

Identifier: oai:euromuseos.mcu.es:euromuseos/MT-CE081792

Is part of: Museo del Traje. Centro de Investigación del Patrimonio Etnológico

Language: spa

Rights: Ministerio de Educación, Cultura y Deporte Museo del Traje. Centro de Investigación del Patrimonio Etnológico Fotografía: Museo del Traje. Centro de Investigación del Patrimonio Etnológico

Publisher: Ministerio de Educación, Cultura y Deporte

Source: Museo del Traje. Centro de Investigación del Patrimonio Etnológico

Data provider: CER.ES: Red Digital de Colecciones de museos de España

Provider: Hispana

Providing country: Spain

Search also for:

Title
Autómata (69)

Who
Georg Köhler (Nuremberg) (act. 1932-1960[ca]) (5)

What
Autómata (69)
Autómata (69)
Juguetes (3963)
Hojalata (941)
Pato (2011)
Georg Köhler (Nuremberg) (6)

Provider
CER.ES: Red Digital de Colecciones de museos de España (133696)
Hispana (2104434)

Figure 6: Example object from the collection Hispana 2022703.

Enrichments	Number of Objects
Agent enrichment	14
Place enrichment	72,124
Timespan enrichment	80,777
Concept enrichment	92,368

Table 5: Number of enrichments in the dataset.

3.4.1. Findings

Title & Description

Named entity recognition could reveal more specific entities not covered in the subject information. For example: subject mentioned cars, while the title mentioned specific brands or models.⁴⁰

Subject

There were not many matches found in GEMET for the subject headings although many keywords are provided. The National Library of Spain Subject Headings (BNESH) might provide more matches than GEMET. For example one object has the terms “Autómata; Juguetes; Hojalata; Pato; Georg Köhler (Nuremberg)” which produces no matches in GEMET but finds three matches using BNESH (Juguetes, Hojalata, Pato). By following several links, subject data could be linked to several languages.

⁴⁰ <http://europeana.eu/portal/record/2022703/D828D667D67037F7739AF01998961BAA6FF59906.html>

Format

In the format field, the concepts of height and length were written in Spanish. An enrichment process could process these values, structure the metadata, and display it in the language of the user. This might be not so valuable for search but would counteract confusion of users.

Temporal coverage

In this field, historical time periods are mentioned. Some time periods were not linked, because of difficulties in date format, particularly for B.C./A.D.⁴¹ Temporal enrichment could be done with Wikipedia (some matches were found) which would also offer different languages.

3.4.2. Recommendations

A broader exploitation of the title and description fields would have been helpful for this dataset. Linking keywords to subject headings requires lemmatization. Furthermore, this might be aggravated by the different indexing practices in museums and libraries. Mapping subjects from libraries to data from museums might prove to be problematic. Perhaps a way to proceed with direct matching is to do a first step that identifies general topics (e.g. cars) and then based on the result we use a more specific vocabulary (e.g. for brand of cars).

Furthermore, Wordnets (e.g. Babelnet) could be used for non-formal terms such as "coche"⁴² (a less formal word for "car" in Spanish). As it may introduce too much noise, mapping should be only done when there is no formal match in the controlled vocabularies.

3.5. Hungarian Jewish Archives (09315)

This collection of the Hungarian Jewish Archives has 2002 documents with metadata in Hungarian and English. Figure 7 shows a typical object represented in the Europeana portal. Table 6 gives an overview over the enrichments in this dataset.



© Free Access - Rights Reserved

View item at
[Hungarian Jewish Archives](#)

Judaica Postcard

Title: Judaica képeslap

Description: Publisher serial number: N.T.G.L6;
Repository/Location: Hungarian Jewish Archives Budapest

Contributor: N.T.G. [publishing]

Date: [1929]

Type: grafikus

Format: height: 10 cm; width: 14 cm; Format: landscape; Format: black and white

Subject: Lilien Jeruzsálem utca; Lilien Jerusalem Street

Identifier: local HJA K505 [Metadata]

Rights: Hungarian Jewish Archives

Source: Hungarian Jewish Archives

Data provider: Hungarian Jewish Archives

Provider: Judaica Europeana

Providing country: Hungary

Figure 7: Example object from the collection Hungarian Jewish Archives.

⁴¹ <http://europeana.eu/portal/record/2022703/444B7A62DE5EB1A83D66D5D99EA2E96A7DE81070.html>

⁴² <http://babelnet.org/search.jsp?word=Coche&lang=ES>

Enrichments	Number of Objects
Agent enrichment	0
Place enrichment	0
Timespan enrichment	0
Concept enrichment	0

Table 6: Number of enrichments in the dataset Hungarian Jewish Archives.

3.5.1. Findings

Title & Descriptions

All objects have 'Judaica postcard' in the title. This title does not occur in the original metadata. It might have been created to fulfil the requirement of providing a title for each document. The problem is that this title is not unique therefore providing no added value when these objects show up in the search results. The title might also be a description of the whole collections without referring to individual objects.

For place names, several different fields are used and these are often found in the `dc:description` field. Splitting values (using the semi-colon) at the mapping stage would have helped to enrich place name.

Coverage & Dates

The value in `dc:date` is '1929' for all objects. However, the original metadata for these objects has no date. It looks as if the provider has created a mapping that includes by default this value for all objects. Perhaps this date is relevant at the level of the entire collection or for another aspect of the collection process, but it is unclear why it is present at the level of the individual objects and what date it refers to.

Format & Type

The whole dataset has 30 different unique types that were retrieved with a SPARQL query. Often the types are in two languages, but the translations are not consistent.

Subject and geographic coverage

Geographic coverage is hidden in subject, often the subject field has also other values such as dates, etc. No enrichment worked on these objects.

Multilinguality

The original metadata is in English and Hungarian. The provided metadata in Europeana is often in both languages across all fields, but often only one value is displayed. There are cases where both languages are displayed in the same field. As this collection is multilingual, it would have benefitted from the use of the language tag.

3.5.2. Recommendations

Most of the issues found in this analysis seem to originate from the mappings. Especially the field values which could not be found in the original metadata are of concern here. One recommendation would be to specify the mapping guidelines for values that cannot be found in metadata but are required by Europeana. Having required fields might increase the probability of creating fields just for Europeana. The task force on EDM mappings, refinements and



extensions⁴³ also identified this creation of Europeana-specific metadata as a process which might introduce metadata quality issues, e.g. for the `edm:rights` metadata element.

Further, a lot of the contextual information was mapped to the subject field and therefore could not be enriched. Especially the chance to use enrichments for geographic coverage is missed here.

The opportunity for multilingual metadata is missed here as no language tags are provided and the fields are mixing the languages. There is documentation on the multilingual metadata and Europeana encourages the use of the language tags for the fields, but practically it is not implemented.

3.6. Europeana 1914-1918

The Europeana 1914-1918 project aggregates the user-generated stories around the First World War. The original metadata can be found at <http://www.europeana1914-1918.eu/en>. In total, the dataset has 64,255 documents. The description of the documents is very long containing the whole story and it is very hard to extract the relevant information from it. Table 7 gives an overview over the enrichments in this dataset.

Fields	Number of Objects
Agent enrichment	0
Place enrichment	791
Timespan enrichment	3669
Concept enrichment	14,660

Table 7: Number of enrichments in the dataset.

3.6.1. Findings

Multilinguality

All documents have the multilingual facet but the language for each item is not known. The providing country is for all documents “Europe”, which is not very granular: these objects are user-generated and their origin was not recorded.

Dates

The date field is not very precise. 6173 objects have a `dc:date` field with dates ranging from 1811-2012. It is often not clear whether the date refers to the photograph taken or the specific event described. This leads to a confusion of dates of occurrences on the user side, e.g. photo taken in 2012 vs. a battle in 1917.

Contributor and Creator

There are confusing roles of agents: who was there, who wrote the text or created document and who made the digitization. For the users, this might not be clear and it made the analysis of fields difficult. The field names as such are hard to understand and often ambiguous. Furthermore, the non-standardized way of formatting the values of the field yield to separation of surname and forename, e.g. Patrick Flocard was spread over two lines leading to two

⁴³ <http://pro.europeana.eu/web/network/europeana-tech/-/wiki/Main/Task+force+on+EDM+mappings+refinements+and+extensions>

separate strings (Flocard; Patrick) where Patrick was used a facet (instead of Patrick Flocart)⁴⁴. This name does not occur in the original provider's metadata.

Subject and Coverage

This field is often enriched with keyword "world war one" (11,831 objects). In cases it was not enriched, "world war one" was not the first keyword in the field or the keyword was missing completely. The original site also has categories that are mapped to the `dc:coverage` field. In general, the `dc:coverage` field covers dates or geographic locations, here it is contextual spatial and temporal terms such as "western front" and "balkan". No enrichment was executed for these terms.

Type

In most cases, for `dc:type` the keyword "item" is used which does not add a lot of value. This keyword cannot be found in the original metadata and was added during the mapping, as a way to distinguish between stories and items in a story. Confusion between items in images vs. items related to documentation might occur for users.

Coverage and Spatial

Only 791 enrichments, which were mostly based on the `dcterms:spatial` fields, could be found. Again these fields are ambiguous and it is often not clear to which geographic location they refer in which context. For example, a post-card from a hospital in Freiburg can have at least two places: the sender's address and where the card was sent. Furthermore, some of the field were not enriched because the geographic coordinates were given instead of the string representation, e.g., "Stryj" of Ukraine vs "49.2561742,23.84655459999999"⁴⁵.

3.6.2. Recommendations


In general, it was recommended to provide the cultural organizations with annotation tools and vocabulary services. This was considered more useful than trying to align the metadata after the ingestion.

3.7. Bernstein Collection (09802)

This collection has been analysed prior to the workshop, as a prototype case. The dataset consists of 91,741 objects representing watermarks. It shows different watermarks, which often seem to be very similar as they have the same motifs but they are mostly unique. The data was provided in ESE and then auto-converted to EDM: precise re-mapping to the much richer EDM is often too resource-intensive for providers. Figure 8 shows a typical object represented in the Europeana portal. Table 8 gives an overview over the enrichments in this dataset.

⁴⁴ http://europeana.eu/portal/record/2020601/attachments_62982_5710_62982_original_62982_JPG.html

⁴⁵ http://www.europeana.eu/portal/record/2020601/attachments_36069_2211_36069_original_36069_jpg.html



View item at
Bernstein project: <http://www.memoryofpaper.eu>

Share

Cite on Wikipedia

Watermark: Oiseau - Figure entière - Sans motif accessoire - Pieds composé d'une ligne seul - Cygne

Description: Distance between chain lines: 39 mm, Depository: StA K, Shelf mark: O.B.A.

Coverage: Tuchsels

Date: 1414; <http://semium.org/time/1414>

Type: image

Format: Height: 62 mm, Width: 35 mm

Subject: paper; Papier; papier; carta; papel; бумага; watermark; Wasserzeichen; filigrane; filigrana; филигрань; <http://www.eionet.europa.eu/gemet/concept/6023>

Identifier: PO:42100

Rights: Landesarchiv Baden-Württemberg, Hauptstaatsarchiv Stuttgart, Deutschland

Publisher: Hauptstaatsarchiv Stuttgart, Deutschland

Provider: Bernstein project: <http://www.memoryofpaper.eu>

Providing country: Europe

Auto-generated tags ▾

What ▾

When ▾

Search also for:

Title
Watermark: Oiseau - Figure entière - Sans motif accessoire - Pieds composé d'une ligne seul - Cygne (7)

What
image (7639795)
paper (170127)
Papier (208217)
papier (208217)
carta (483738)
papel (150581)
бумага (140582)
watermark (119979)
Wasserzeichen (119969)
filigrane (120056)
filigrana (119983)
Филигрань (119961)
<http://www.eionet.europa.eu/gemet/concept/6023> (2956)
Height: 62 mm, Width: 35 mm (52)

Figure 8: Example record of Bernstein collection.

Fields	Number of Objects
Agent enrichment	0
Place enrichment	68,217
Timespan enrichment	91,741
Concept enrichment	91,741

Table 8: Number of enrichments in the dataset.

3.7.1. Findings

Title

Two objects have titles missing. The description of what the watermark depicts is in the title. The title represents the classification of the provider (figure 9), which is in hierarchical order. The flat representation in Europeana does not reflect the richness of the original classification. The provider offers this classification in German, English and French, but the metadata in Europeana is only in one of these languages.

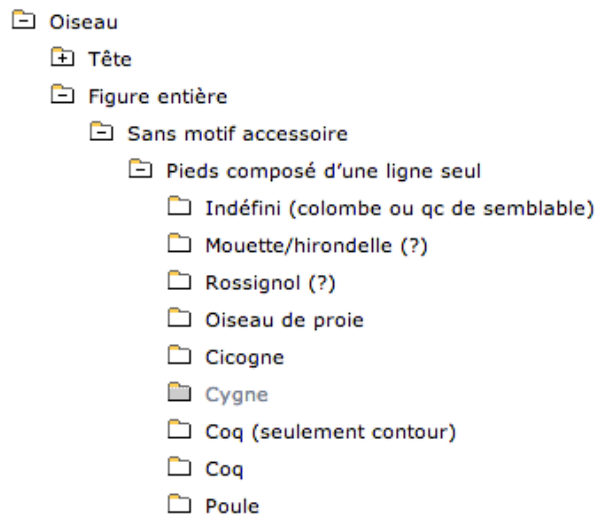


Figure 9: Classification of the provider which is reflected in the title in the objects

The classification of watermarks seems to be universal, the highest class of watermark motifs seems to be the same across different providers. Keeping this information and enriching it would have been important. Most of the objects seem to be in German, but some are also in French. It is not obvious what determines the language of the document in Europeana in this case.

Creator

22,344 (24.36%) objects have a `dc:creator` field.

Coverage & Dates

83,975 objects (91.54%) have a `dc:coverage` field which was enriched when the location was found in GeoNames. In the cases when it was not enriched it, the location was small, old or an abbey or similar.

Almost all objects have a date field, for two objects the date field is missing, all dates have been enriched.

Type & Format

The value in the type field is always “image”, although “watermark” would have been the better choice, here. Most objects have a format field, 1.58% of the objects have no format field.

Subject

All of them have a subject field and have the same subject keywords. The separation of terms in the subject field worked well. The different language versions are not indicated in the subject field and have no language tag which could be used for display purposes later.

Providing country

All objects have this field with the value “Europe”.

3.7.2. Recommendations

The quality of enrichments could benefit from taking “watermark” and the first hierarchy in the classification of the provider: `watermark:oiseau` as a title. The rest of the terms of the classification describing the watermark could be put in the description field in all three languages. In the subject field, it would have been beneficial to use the term watermark and the first term in the hierarchy. This could then be enriched, e.g. enrich “watermark” and enrich “oiseau”.



Enrichments

All the creators are not well known, so a very specific vocabulary is needed here. All documents were enriched with a time span. 23,524 (23.46%) documents have no geographic enrichment although 15,758 (17.18% of the whole dataset) of these objects do have a `dc:coverage` field. Here, the geographic locations refer to an abbey or a small place that were not found in Geonames. 7766 (8.47%) were not enriched because they do not have a `dc:coverage` field.

All document were enriched with the concepts label “Paper” and the broader term: “Industrial product”. The better enrichment term here would have been “Watermark” which reflects the scope of the collection better. For the enrichment term “Paper”, the vocabulary that the project “Partage Plus” uses, would have been the better choice:

<http://partage.vocnet.org/html/vocItem.php?uri=http://partage.vocnet.org/part00575>

Multilinguality

The dataset is multilingual, the provider has the objects in three different languages. The title in Europeana is in one of the three languages but mainly in German whereas the description and format fields have values in English. This mix of languages within one record and within the dataset is confusing.

4. Results and Recommendations

This section structures the findings of the analysis of the datasets and highlights the problems with the enrichments from different angles.

As Olensky et al. (2012) pointed out, there are three levels that influence the quality of enrichments:

- the metadata level (the source of enrichment),
- the vocabulary level (the target of enrichment), and
- the workflow level (the process to create enrichments).

Figure 10 shows an overview graphic from this paper showing the sub-categories on each level.



Figure 10: Impact levels on enrichment (simplified model based on Olensky et al., 2012).

Similar to these findings, this task force also identified metadata quality as one of the main points for poor enrichments. Analyzing the datasets, we found that enrichment flaws can be caused during on of the three stages that the metadata undergoes until it is displayed in Europeana:

1. Creation of the metadata by the provider.
2. Mapping to EDM.
3. Ingestion into Europeana.

During the process of enrichment itself, two choices are key to success or failure:

- the vocabulary used for enrichment, and
- the rules established for using the right terms for enrichment (on the target and the source side).

The goal of our task force is to determine measures Europeana can take to ensure high quality of its multilingual and semantic enrichments. The focus has been put feasible solutions that are executable by Europeana.

4.1. Metadata Quality

4.1.1. Tackling quality issues in the original metadata

The analysis showed that many enrichment flaws can be explained by shortcomings in the original metadata. The task force did not find many examples of purely wrong metadata. But in those cases, enrichments made the problem much worse as they exposed it to the users in many languages. Detecting these issues automatically is tricky; often they can be discovered only by manual review. To facilitate this, Europeana could have a **feedback form** that allows users of its web portal **to flag incorrect metadata and enrichments**.

Enrichments have often been missed because of syntactic aspects such as single strings that introduce different subjects using different separators (e.g. either comma or semicolon). In case where formal formatting rules may apply, such as dates, Europeana could check if these formal rules are applied for the field considered for enrichment. These thorough **validation rules** should be provided by Europeana. Only then an enrichment would happen based on this particular field.

Providers could also seek to **upgrade the original data**, by using tools like OpenRefine⁴⁶, which could be used to clean data, remove inconsistencies and standardize values in certain fields. Another option could be to tackle the issue at the mapping stage, with Europeana **documenting and promoting best practices to re-format fields** when providers submit their data.

An issue that results in many low-value enrichments is that the metadata assigned to individual objects can sometimes lack precision. For example, when the type assigned to an object is very vague or all objects of a collection have the same `dc:coverage` value, an enrichment with a very broad concept may deliver only little value. Europeana could **enrich objects only when their metadata reaches a certain quality score**, checking for example the diversity of titles and descriptions within the same dataset.⁴⁷

Finally, the links to the objects on the original sites were often unresolvable. While not a problem for the enrichment process itself, it really hinders its evaluation and an **automatic link checking / resolving** process should be instituted as part of the ingestion process. This should be done by Europeana

Especially, (persistent, linked data) identifiers were mentioned throughout the workshop as requirements for high quality metadata and enrichments. For example, providers should provide URIs for conceptual resources available in SKOS, like classes from the Iconclass classification (e.g. <http://iconclass.org/25F34>, rather than using mere strings for these resources ("owls") or even worse, codes without further information ("25F34")).

⁴⁶ <http://openrefine.org/>

⁴⁷ It should be noted that uniqueness should be weighted differently for different fields. Here, statistics can be useful to evaluate quality of datasets provided.

To accompany this, Europeana should encourage the *use of explicit and persistent (linked data) identifiers as metadata values* and transforming the corresponding controlled vocabularies into SKOS and similar interoperability formats⁴⁸.

It would also be interesting if some *reference repositories* are provided for people *to find vocabularies and other contextual resources* in the cultural heritage domain. Europeana would be in the best position to point data providers to existing reference repositories.

As this task force focuses on what Europeana can do rather than finding measures for data providers, we do not expand further here. We hope that some of the ideas will be re-used as more formal recommendations as part of a new task force on metadata quality, which has been launched shortly prior to the publication date of this report.

4.1.2. Mapping to EDM

Most of the enrichment flaws originate in the mappings which often introduce ambiguities. Our analysis indicates several spots within the mapping process, where measures could be taken to minimize enrichment errors. For Europeana this translates into three main lines of work:

Documentation: Europeana provides many documents on how to map to EDM but they might not be known or are not used as they are lengthy and tend to be technical. Here a strategy would be needed that better supports providers in finding the right information and answers to recurring mapping problems. From the analysis, it appeared that documentation should try and better answer the following *specific questions*:

- How to populate required EDM fields (e.g. `dc:title`) in the most meaningful way, when they do not exist in the original metadata for individual objects?
- How should the Dublin Core fields of an original 'one-size-fits-all' record be distributed among various EDM classes (e.g. `ProvidedCHO`, `WebResource`)?
- When/how should two separate fields in the original metadata be mapped into one value in EDM?
- When/how should multi-valued, single fields be separated into distinct fields in EDM?
- How should date intervals be mapped to proper time spans and not individual dates?
- How should persistent identifiers, e.g. identifying vocabulary terms, be handled during mappings?
- How to provide explicit and persistent links/URLs as metadata values (cf. previous section)?
- Is it appropriate for mappings to assign one value for a field (e.g. `subject`) over an entire dataset?
- How to adapt the degree of granularity of the metadata through the mapping (i.e. `dcterms:created` over `dc:date`)?

In general, the work of providers and the crucial task of aggregators that send the message to them, could be facilitated by *documentation that explains which fields are getting enriched and how*, and make it clear that enrichment could be better if they provide more appropriate metadata. It would also help if Europeana could present more openly the *difference between the metadata* that is *stored* in its production database, what is *displayed* in the Europeana.eu portal and the values that are *indexed* in the current search engine behind the portal.

Supporting tools for providers: the members of the task force noted that tools should be given for providers *for testing their mappings* in a realistic setting.⁴⁹ These tools should not

⁴⁸ Awareness to these issues has however already risen considerably, as testified for example by the Linked Heritage vocabulary guidelines (Leroi et al, 2013).

⁴⁹ In the line of a Content Checker that Europeana has provided to its providers until 2012. See <http://pro.europeana.eu/documents/900548/ae5e78e8-ce78-424d-b360-5c01eddb3564>

only focus on the display of the resulting EDM metadata for Europeana.eu users. A **preview** of the mapped metadata in the portal would strengthen the awareness on display issues and might lead to overfitting the metadata to display and make it much less optimal for data exchange (API, Linked Data) and the search index. A tool that **combines the display** of the data in the portal and **highlights the values which are enriched and searched** would be beneficial. The current presentation at europeana.eu, where enrichments are presented together with the original metadata and providers' own contextual resources, is quite confusing.

Direct communication and training: Europeana's data ingestion team organizes or participates in a number of meetings with data providers on a national or project basis in order to help them understand how to contribute metadata to Europeana. Some task force members, based on their previous experience⁵⁰, suggested that such **clinics** were a great opportunity of establishing direct contact with the professionals that curate and submit the metadata. Experience shows it is possible to teach the principles and practice of metadata cleaning and enrichment to a group of 30-40 people within a four hour session.

4.1.3. Checking metadata at ingestion time

Some recommendations regarding the quality of metadata provided to Europeana should be **enforced at the ingestion stage**. Some enrichments errors which are based on low quality metadata can be indeed avoided during ingestion time. Fields that do not respect agreed best practices should be **flagged to the providers**, e.g., dates that are in the future or that are not provided in preferred formats. Duplicates should be filtered out. Here, the use of **quality scores** can be discussed. Determining the measures for such a quality score was not part of this task force.

The following table summarizes potential measures Europeana can take to positively influence the quality of the enrichments.

Issue	Findings
Metadata quality	<ul style="list-style-type: none"> close collaboration with providers and institutions would improve metadata quality⁵¹ encourage the use of persistent, linked data URIs for vocabularies establish rules for field formatting feedback for flagging wrong metadata
Mapping to EDM	<ul style="list-style-type: none"> more specific and targeted documentation highlighting common issues supporting tools for mapping that combines display of data, indexed fields and enriched fields metadata clinics for aggregators
Ingestion	<ul style="list-style-type: none"> quality score at ingestion time to identify low quality metadata validation reports for providers to show them metadata quality issues metadata quality score threshold for executing enrichments, e.g. to ensure that fields are formatted right

⁵⁰ <http://freeyourmetadata.org/>

⁵¹ The recently launched task force on metadata quality is set out to find solutions for improving the quality.

4.2. Vocabularies

It was shown that ***the vocabulary should fit the context of the record to be enriched***. There is no vocabulary which can enrich all collections in Europeana. Often, a ***contextual vocabulary coming from the provider's context is the best solution*** but during mapping it often gets lost or is mapped to fields in a way that makes retrieving the right link very difficult, if not impossible. Additionally, the vocabulary should come with labels in the same languages as the one of the metadata to enrich. A list of multilingual vocabularies aggregated by the task force members can be found in appendix 3.

Furthermore, most of the ***datasets are embedded in rich classifications***, which in the best case are multilingual, that are often flattened in Europeana. Documentation and training should be offered on how to ***map these rich structures so they can be exploited for search and display in Europeana***. DBpedia, in particular, provides rich, multilingual descriptions of many types of cultural objects, which would be of interest: see for example the genres accessible http://dbpedia.org/resource/Category:Film_genres in the movie domain.

A mapping to such rich structures may be “indirect”: let’s consider the DBpedia resource <http://dbpedia.org/resource/Incunable>. This resource is declared semantically equivalent to <http://de.dbpedia.org/resource/Inkunabel>, which has been aligned with the resource <http://d-nb.info/gnd/4027041-5> from the German National Library. Data enrichment modules able to exploit such chains of relations could bring in the multilingual data, starting from an existing “local” German indexing. This should be considered while selecting vocabularies for enrichment, either by providers or by Europeana. Also, providers as well as Europeana should envision ***carrying out their own efforts for vocabulary mappings***, as was explored in the EuropeanaConnect project (de Boer et al., 2011).

Another issue arose in several datasets for the enrichments with the ***broader concepts*** of GEMET. They are ***often not precise*** (e.g., 'culture') and in some cases even totally off-topic (e.g. reflecting an environmental perspective on the word) as shown in Appendix 2. Europeana might benefit from skipping some of these concepts, in both enrichment steps it performs, i.e. (1) linking objects to concepts, and (2) fetching data on broader concepts to feed into the search index.

Finally, it was suggested that Europeana and its partners could grow their own ***reference resources for certain metadata fields with limited amount of values, such as format, language, country***. First scan the entire Europeana dataset to collect all different values, and then match these values to existing reference vocabularies (semi-)automatically or even manually. Gradually, synonyms or misspelled variants could be manually picked out and put into the reference resource (or in a mapping table used in the enrichment process) for a correct enrichment.

Vocabulary	<ul style="list-style-type: none"> • encourage the delivery of vocabulary fitting the collection’s context by the data provider • exploit classifications of providers • explore alignment of vocabularies and the exploitation thereof • skip the broader terms in GEMET and do not use them for enrichments
------------	---

4.3. Enrichment Process

Enrichment rules aim at establishing formal criteria for enriching certain fields. Here, several recommendations occurred from the analysis. First, it does not seem to be clear what kind of matching rules were applied for certain fields and the biggest problems arose from strings

separated by comma or semicolon. **Documentation is needed, which presents the enrichment rules for every field and enables to spot enrichment errors more easily.**⁵²

Furthermore, due to a wrong heuristic applied to older datasets, Europeana often found only one term to enrich with GEMET concepts even when the metadata provides more terms suitable for matching. Often this keyword is not one representing the objects in the best possible way. For example, the Saxon State Library provides a curated keyword list in the dc:subject field. The first word is often “Kunst”, German for art, followed by more specific keywords that were not matched and might have been a better enrichment choice⁵³. **In the best case, the automatic enrichment should try to match all keywords in a field, even if one match has already been found.**⁵⁴

Finally, matches shouldn't happen between metadata field values in one language and labels of a semantic resource in another language. Many multilingual ambiguities are introduced in the mappings due to this missing rule. This will be difficult to implement, as most of the metadata currently in Europeana doesn't state the language it comes from. It may however be possible to use the language of the country of origin of the collection as a proxy.

In some cases, it would be helpful to break geo-location names into small units before matching to GeoNames.

Enrichment process	<ul style="list-style-type: none"> • establish enrichment rules for every field, e.g. pursuing basic splitting of values and document them well • enrich all keywords within a field and do not stop enrichment after the first match in a field • match the language of the metadata field (often, the language of the country of origin is sufficient) with the language of vocabulary
--------------------	---

4.4. Further Ideas

Another point of discussion was to leverage the users' input to crowd-source and validate links, filter ambiguous meaning and relations. **Overall, users could be more involved in improving metadata quality and enrichment quality.**

⁵² This task force actually triggered a first attempt at this from Europeana, cf. Appendix 1

⁵³

http://europeana.eu/portal/search.html?query=europeana_collectionName%3A01004*+cc_skos_prefLabel%3Akunst

⁵⁴ This is now applied on every ingestion from January 2014 onwards.

5. Conclusions

This task force was initiated to establish a multilingual and semantic enrichment strategy for Europeana. For that, several datasets were analyzed regarding their metadata's potential for enrichment. Naturally, other issues were reported that were noticed by the workshop participants. Often these remarks concern the quality of the metadata and not the enrichment process per se. In many cases it has an influence on the quality of the enrichments but there cannot be much improvement regarding enrichments if the basis for it is of low quality. In the worst case, enrichments increase the problem and spread it across several languages.

This task force identified three main areas of concern for enrichment quality, namely metadata quality, vocabularies and the enrichment process. Most of the recommendations target the metadata and suggest solutions to empower providers to deliver high quality content. One aspect of this is to raise awareness of enrichments, their potential to increase visibility of objects paired with their risk when based on poor metadata. Non-technical documentation and training events for aggregators, such as metadata clinics, are one of the proposals from this task force.

Recommendations that targeted enrichment sources, targets and the process were rather light compared to the ones on original data and mapping. Given the time constraints of a one-day workshop, many pitfalls were identified that – if addressed correctly – can contribute to a higher level of metadata quality for all of Europeana. Improved enrichments are one of the many benefits of this.

Another outcome formulated by the members of this task force is the conclusion that there isn't one solution fitting all datasets. The very heterogeneous data in Europeana make it hard to generalize recommendations that can be applied to all objects. This is especially true for the vocabulary chosen for enriching the fields. For almost all analyzed datasets, there would have been a better and more specialized vocabulary fitting the context of the particular collection. Embedding more of the resources coming directly from data providers is the way to go here. In cases where providers were already doing that, there were often no URIs delivered with it. Again, Europeana should encourage best practices for delivering specialized vocabulary to make its data richer. Additionally, the one-size-fits-all approach regarding vocabularies such as GEMET can lead to a decrease in search performance as broad and generic terms are applied to the majority of objects.

In general, it can be observed that the product of the enrichments process is influenced by many more factors than the source, target and process decisions made. Being aware of these influences is the first step in improving enrichments overall having their effect on user experience and retrieval performance in mind.

Acknowledgements

The authors would like to thank all participants of the task force for the enthusiasm in analyzing collections, their feedback on this report and their contributions during the course of this task force. The members of the task force are Agnès Simon, Daniel Vila Suero, Eero Hyvönen, Esther Guggenheim, Lars G. Svensson, Nuno Freire, Rainer Simon, Rodolphe Bailly, Roxanne Wyns, Seth van Hooland, Shenghui Wang, Vladimir Alexiev. We would also like to thank Valentine Charles, David Haskiya, Maarten Brinkerink and Kate Fernie for providing valuable feedback.

References

Victor de Boer, Antoine Isaac, Guus Schreiber, Jacco van Ossenbruggen, Jan Wielemaker: Multilingual mapping of schemes and vocabularies. EuropeanaConnect Deliverable D2.3.1, 2011. Available at

<http://pro.europeana.eu/documents/866481/838447a1-d64d-40c5-b840-20b91e2217e1>

Stefan Gradmann: Information in Context: on the Importance of Semantic Contextualisation in Europeana. Europeana White Paper, 2010. Available at

http://pro.europeana.eu/c/document_library/get_file?uuid=cb417911-1ee0-473b-8840-ed7c6e9c93ae&groupId=10602

Marie-Véronique Leroi, Johann Holland, Stéphane Cagnot: Linked Heritage WP3 and ATHENA WP4 "Terminology and multilingualism" (ed.): Your Terminology as a part of the Semantic Web. Recommendations for Design and Management, 2013. Available at:

<http://www.linkedheritage.eu/getFile.php?id=244>

Marlies Olensky: Market study on technical options for semantic feature extraction. Europeana version 2.0. Deliverable D7.4. Available at:

http://ec.europa.eu/information_society/apps/projects/logos//2/270902/080/deliverables/001_DeliverableD74MarketStudyToolsSemExtracfinal2.pdf

Marlies Olensky, Juliane Stiller, Evelyn Dröge: Poisonous India or the Importance of a Semantic and Multilingual Enrichment Strategy. Metadata and Semantics Research. 6th Research Conference, MTSR 2012, Cádiz, Spain, November 28-30, 2012. Proceedings. 2012, pp 252-263.

Mark Stevenson and Arantxa Otegi with Eneko Agirre, Nikos Aletras, Paul Clough, Samuel Fernando and Aitor Soroa: Semantic Enrichment of Cultural Heritage content in PATHS. Paths Project, 2013. Available at: <http://www.paths-project.eu/eng/Resources/Semantic-Enrichment-of-Cultural-Heritage-content-in-PATHS>

Shenghui Wang, Antoine Isaac, Valentine Charles, Rob Koopman, Anthi Agoropoulou, Titia van der Werf: Hierarchical structuring of Cultural Heritage objects within large aggregations. In: Proceedings of the International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. pp 247-259.

Appendix 1 – Europeana Enrichment Process

Introduction

Europeana's current enrichment process is based on the Annocultor tool, <http://sourceforge.net/projects/annocultor/>

NB: For some general explanation, see <http://semium.org>, and a cached version of some of this site's page at <http://europeanalabs.eu/wiki/EDMPrototypingTask21Annocultor>

For Java developers: the main class to start exploring Europeana's version of Annocultor is https://github.com/europeana/tools/blob/master/annocultor_solr4/src/main/java/eu/annocultor/converters/solr/BuiltinSolrDocumentTagger.java

Currently Europeana enriches objects by creating links to contextual resources - places, concepts, agents and time periods. For each category, the following sections indicate:

- the *target* of enrichment, i.e. the set of resources to which objects are linked,
- the *source*, i.e. the fields in EDM data from which the links are derived - mostly by matching the string value of these fields to the labels of the contextual resources;
- some details about the *rules* that specify how the match between the field and the (labels of the) contextual resource is made;
- an attempt to quantify the *results* of the enrichment - especially the number of objects that are enriched.

NB: At the time of writing, the display of objects at Europeana.eu render enrichments in a confusing way. URIs present in the original metadata and URIs resulting from enrichment are both shown in the fields representing the original values (e.g. "Geographic coverage") and the "Auto-generated tags" section. See

<http://europeana.eu/portal/record/09003/818CA77941A28830F41B3DCD8FA25EB951E50971.html>, a correctly enriched object from the CARARE project.

http://europeana.eu/portal/record/09102/_ULEI_M0000798.html, an object from the MIMO project where the edm:Place was given in the original metadata

Places

Target resources: a subset of Geonames⁵⁵:

1. countries

https://github.com/europeana/tools/tree/master/annocultor_solr4/converters/vocabularies/places/countries

2. a selection of places from European countries from selected Geonames classes (allowed prefixes are likely to be "A", "P.PPL", "S.CSTL", "S.ANS", "S.MNMT", "S.LIBR", "S.HSTS", "S.OPRA", "S.AMTH", "S.TMPL", "T.ISL", cf definitions at <http://www.geonames.org/statistics/total.html>). See

https://github.com/europeana/tools/tree/master/annocultor_solr4/converters/vocabularies/places/EU/

Source fields: dcterms:spatial, dc:coverage

⁵⁵ <http://geonames.org>

Rules for enrichment: exact matching of the field value and labels of the place; no splitting of the field value (string) occurs prior to the matching; labels in the target vocabulary are lower-cased before matching (the original label is saved for later use, e.g., for display).

Result: 01/2014: 4.6M objects have Geonames places (11/2013: 4.6M, 12/2013: 4.4M), http://europeana.eu/portal/search.html?query=edm_place:*geonames*

NB: this query is more precise than a query that was used before, which gave all objects with Places described in their records

([http://europeana.eu/portal/search.html?query=edm_place%3A](http://europeana.eu/portal/search.html?query=edm_place%3A*), 01/2014: 5.8M, 12/2013: 5.8M, 11/2013: 5.6M⁵⁶).*

But it still only gives an upper bound for the total amount of enrichments, since it also includes objects linked to Geonames places that might have been submitted by providers in the original data. See for example this object from the MIMO dataset:

http://europeana.eu/portal/record/09102/_ULEI_M0000798.html?format=labels

Getting more precise numbers is not possible now. Indeed, Europeana's enrichments and the providers' ones are attached to EDM proxies using the same set of properties (e.g., `dcterms:spatial`), and within Solr we can't distinguish whether a link to a place is attached to Europeana's proxy or the provider's proxy. For this, we need more precise tracking, or an RDF tool using the SPARQL query language.

Concepts (topics)

Target resources:

1. 5,208 GEMET⁵⁷ concepts

https://github.com/europeana/tools/tree/master/annocultor_solr4/converters/vocabularies/concepts/gemet

NB: 1 concept was removed: Drawing, a medical concept that was almost always used to enrich objects having little to do with medicine. Some labels from other concepts have also been removed to prevent numerous harmful matches, such as linking any print to the physical “pressure” concept because of its German “Druck” alternative label, as identified in (Olensky et al, 2012).

2. A handful of WWI battles, the two categories “World War I” and the following resources from the DBpedia⁵⁸ categories:

Art	Romanesque_art
Architecture	Romanticism
Art_Deco	Rococo
Art_Nouveau	Pastoral
Baroque	Portrait
Cubism	Street_art
Contemporary_art	Surrealism

⁵⁶ This query also includes Places that are 'orphan', i.e., that are not linked to the ProvidedCHO they should have been, e.g., by using the `dcterms:spatial` field. See for example http://europeana.eu/portal/record/2020724/HA_http___www_kulturarv_dk_fundogfortidsminder_site_166_38.html, as seen in the details at http://europeana.eu/portal/record/2020724/HA_http___www_kulturarv_dk_fundogfortidsminder_site_166_38.html?format=label

⁵⁷ GEneral Multilingual Environmental Thesaurus, <http://www.eionet.europa.eu/gemet/>

⁵⁸ <http://dbpedia.org>

Dada	Symbolism
Digital_art	Music
Expressionism	Theatre
Fine-art_photography	Painting
Folk_art	Sculpture
Futurism	Drawing
Impressionism	Poster
Neoclassicism	Photograph
Pre-Raphaelite_Brotherhood	Furniture
Kitsch	Costume
Still_life	Fashion
Landscape	Jewellery
Minimalism	Porcelain
Modernism	Tapestry
Renaissance	Woodcut
Realism_(arts)	

See

https://github.com/europeana/tools/tree/master/annocultor_solr4/converters/vocabularies/concepts/wikipedia

Source fields: dc:subject, dc:type

Rules for enrichment: exact matching

Result: 01/2014: 9.1M objects are linked to GEMET concepts (12/2013: 8.8M, 11/2013: 8.7M)

http://www.europeana.eu/portal/search.html?query=skos_concept%3A*gemet*

588K objects are linked to DBpedia (12/2013: 324K, 11/2013: 230K)

http://www.europeana.eu/portal/search.html?query=skos_concept%3A*dbpedia*

NB: as for the query used to measure place enrichment, these queries indicate an upper bound for actual enrichment. They may return objects that have been linked to GEMET or DBpedia by providers, not by Europeana. This is very unlikely to happen for GEMET, but may happen for DBpedia

Agents (persons)

Target resources:

1. 14K artists (painters) from DBpedia.org

https://github.com/europeana/tools/blob/master/annocultor_solr4/converters/vocabularies/people/dbpedia.selected.artists.rdf
 see selection criteria at
https://github.com/europeana/tools/blob/master/annocultor_solr4/src/main/java/eu/annocultor/converters/people/FetcherOfPeopleFromDbpediaSparqlEndpoint.java

2. 2 artists from ULAN

Source fields: dc:creator, dc:contributor

Rules for enrichment: the matching heuristics take into account variants in (first/last) name ordering, presence of birth date/death dates in parentheses or roles in brackets, and upper-case/lower-case variants (though capitalized version is the version kept for the output). The implementation is at https://github.com/europeana/uim-europeana/blob/master/workflow_plugins/europeana-uim-plugin-enrichment/src/main/java/eu/europeana/uim/enrichment/normalizer/AgentNormalizer.java

Result: 01/2014: 44K objects are linked to DBpedia (12/2013: 34K, 11/2013: 12K),
http://www.europeana.eu/portal/search.html?query=edm_agent%3A*dbpedia*

18 objects are linked to ULAN (12/2013: 18, 11/2013: 16)

http://www.europeana.eu/portal/search.html?query=edm_agent%3A*ulan*+AND+NOT+edm_agent%3A*dbpedia*

NB: again, this query indicates an upper bound for actual enrichment. It may return objects that have been linked to DBpedia or ULAN concepts by providers, not by Europeana.

NB: providers are already submitting a great deal of Agents, e.g. Hispana provided the object:
http://europeana.eu/portal/record/2022701/lod_oai_archivoimagen_jccm_es_41509_ent1.html

The query

http://www.europeana.eu/portal/search.html?query=edm_agent%3A*

returns both our enrichments and providers' agents: 236K objects in 01/2014 (12/2013: 219K, 11/2013: 140K).

Time periods

Target resources: Semium Time⁵⁹, a vocabulary of time periods generated partly automatically (for “objective” time divisions like the 3rd quarter of 15th century), partly manually (for historical period like “Roman empire”)

https://github.com/europeana/tools/tree/master/annocultor_solr4/converters/vocabularies/time

Source fields: dc:date, dc:coverage, dcterms:temporal, edm:year (itself derived from the previous fields)

Rules for enrichment: some English words (qualifiers to dates, e.g. “made”, “printed”...) are removed from fields prior to enrichment. I.e., the enrichment uses 'copies' of these fields, which do not contain these words anymore); see source code at https://github.com/europeana/tools/blob/master/annocultor_solr4/src/main/java/eu/annocultor/c/converters/solr/BuiltInSolrDocumentTagger.java

Matching is then done by exact matches. Some splitting happens indirectly, because edm:year is sometimes generated after splitting one of the other fields.

Result: 01/2014: 12.6M objects are linked to Semium (12/2013: 12.8M, 11/2013: 13.3M), see http://www.europeana.eu/portal/search.html?query=edm_timespan%3A*semium*

NB: again, this query indicates an upper bound for actual enrichment. It may return objects linked to Semium by providers, not by Europeana. It is extremely unlikely though.

⁵⁹ <http://semium.org/time.html>

Appendix 2 – Broader GEMET Terms

The following table shows the most used broader GEMET terms in Europeana and the number of records enriched with the particular term. Please note that objects can have several broader terms.

Broader GEMET term	Number of objects with this term
documentation	2919522
human science	2201295
industrial process	812563
acoustics	369640
geographical projection	354776
industrial product	242169
document type	223079
policy	133170
physical property	132289
community facility	113764
cultural heritage	111722
mass media	103928
built environment	90015
mountainous area	87442
industry	70335
health care	68676
society	68633
cultural facility	65533
population structure	57479
recreation	55071
science	46639
agriculture	38498
world	35649
traffic infrastructure	33257
earth science	29806

mill	29006
industrial site	28463
land	27365
life science	24983
institutional structure	24657
miscellaneous product	23862
plant component	22837
military activities	20450
road	20450
leisure activity	19060
road network	18954
culture (society)	18810
education	18761
motor vehicle	17493
ecological parameter	17204
economy	17149
vehicle	16864
technical regulation	16753
public information	16698
legislation	16314
transition element	15488
rock	14887
equipment	14073
green space	13829
traffic	13822
physical process	13739
social behaviour	13158
medicine (practice)	12977
labour	12614
human disease	11793
legal form of organisations	11675
water distribution system	11236

chemical industry	11053
type of waste	10940
judicial system	10896
forest product	10663
didactics	10338
tourism	10298
craft	10034
built-up area	9984
plant (biology)	9757
socioeconomic aspect of human settlements	9707
communications	9609
chemical	9556
trade (services)	9526
animal textile fibre	9307
organism	9281
equine	8996
environmental planning	8468
health facility	8176
environment	8116
railway network	7652
building land	7649
judicial body	7333
forest	7319
age	7068
landform	6988
asia	6907
river	6887
modelling	6842
health	6805
vertebrate	6780
wetland	6773
social group	6661



public institution of administrative nature	6649
land setup	6553
law (corpus of rules)	6549
law branch	6505
extraction	6192
parameter	6065
public service	6051
farm	5954
terrestrial area	5773
flowering plant	5757
sediment	5681

Appendix 3 – Vocabularies

The following table lists the potential vocabularies for enrichment contributed and aggregated by the members of the task force. Within several projects in the past, multilingual vocabularies suitable for semantic enrichments were collected. Two of the biggest collections are:

- [Multilingual Mapping of Controlled Vocabularies](#)
- [Inventory of value vocabularies for EuropeanaConnect WP1/2](#)

Vocabulary	URL	Language	Licence	Comment
UDC summary edition	http://www.udcc.org/udcsummary/php/index.php	51 languages	summary edition available as CC0	
Getty Thesauri: AAT, TGN, ULAN	http://vocab.getty.edu	en, es, nl, zh and several others	ODC BY	TGN and ULAN scheduled for release in 2014
VIAF	http://viaf.org/	international	ODC BY	VIAF subsumes ULAN. ULAN has more relations, but there are more languages in VIAF and the scope is different
EuroVoc	http://eurovoc.europa.eu/	24 languages	http://eurovoc.europa.eu/drupal/?q=de/legalnotice	
British Museum		en		People, Places, Subjects, Object types, Materials. Not aligned to anything. Available in SKOS RDF
Pleiades	http://pleiades.stoa.org	en + latin/greek names	CC BY	Ancient place names
Pleiades Time Periods	http://pleiades.stoa.org/vocabularies/time-periods	en	CC BY	Time periods of antiquity
DBpedia	http://wiki.dbpedia.org/About	many	from version 3.4 on is licensed under the terms of the Creative Commons Attribution-ShareAlike 3.0	Excellent source for people, sub-city places (eg stadium, museum). Challenge is how to select classes to be used for indexing With a tool that takes context analysis into account, it can be used for a broad range of terms

French national library : authors, places (geographic names) subjects	http://data.bnf.fr/semanticweb-en	French and other languages, also transliterate	Open Licence (=CC-BY)	200 000 authors at this date, 1,5 million, in 2-3 years 170 000 topics 110 000 places
Thesaurus PICO	http://purl.org/pico/thesaurus_4.3.0.skos.xml	it, en	CC BY	Contact: ICCU, Sara Di Giorgio Available in SKOS RDF. On persons, organisations, classifications, periods, actor roles, object names...
Thesaurus of "The Israel Museum, Jeruzalem	http://www.judaica-europeana.eu/Search_Europeana_Collections_in_Hebrew.html	he, en		Contact Dr. Allison Kupietzky SKOSified Was used for the Europeana Judaica project
Joconde thesauri	http://www.culture.gouv.fr/documentation/joconde/fr/pres.htm	fr		Terminologies of the Joconde database, managed by MCC, French ministry of Culture and Communication Not SKOSified? Contact: Jeannette IVAIN Different vocabularies on periods, styles, entities, materials and techniques...
AM-MovE thesaurus	http://www.museuminzicht.be/public/musea_werk/thesaurus/zoeken/index.cfm	nl	Free of charge for non-profit organisations	MoVe thesaurus, often used by museum sector in Flanders. Same topics as Getty AAT. In the future they will try to align the thesaurus more with the AAT. Available as XML.
Library of congress subject headings, name authority file, classifications...	http://id.loc.gov/	en	public domain	Offered through Linked Data service and downloadable in different formats: SKOS RDF, JSON, etc. Also see OCLC FAST. LCSH problem: many subjects are pre-coordinated (e.g. " Italian poetry--16th century "), which makes them unsuitable for enrichment unless the collection used exactly LCSH

RMAH thesauri on: object names, materials and techniques, geography, musical instrument classification,	www.rmah.be ; www.carmentis.be	fr, nl, en	freely available on request	Thesauri focussing on the museum collections ranging from prehistoric times until Art Deco. Completely available in French, Dutch and English. Some concepts have scope notes. Based on AAT, British Museum and other terminology reference thesauri. Main problem: only exportable in CSV
Europeana photography vocabulary	http://www.europeana-photography.eu/	Translated in the 12 languages of the project partners	CC0?	Thematic thesaurus on photographic keywords, materials and techniques, classifications...developed within the framework of the Europeana photography project. SKOSified Contact: n.vansteen@kmg-mrah.be
Europeana Fashion vocabulary	http://www.europeanafashion.eu/	DE, SE, RS, FR, PT, GR, NL, SE, ES, AT, IT, EN	CC0	Thematic thesaurus on Fashion keywords, materials and techniques, classifications...developed within the framework of the Europeana photography project. SKOSified Contact: n.vansteen@kmg-mrah.be
KOKO Ontology Cloud	http://kansalliskirjasto.onki.fi/ - Includes ca. 15 linked ontologies of general concepts	FI, SE, EN	CC BY	Will be maintained by the National Library in 2014-; fairly comprehensive collection of general concepts (ca. 40 000) in SKOS hierarchies eero.hyvonen@aalto.fi , matias.frosterus@kansalliskirjasto.fi
Finnish History Ontology HISTO (Events)	http://www.idf.fi/dataset/history/index.html	FI	CC BY	Major historical events (ca 1100) in Finnish history. We are combining it e.g. with the national biography of 6500 short biographies. eero.hyvonen@aalto.fi
Ontology of Historical Persons TOIMO	http://www.idf.fi/dataset/agents/index.html	FI	CC BY	Over 150 000 agents harvested from various sources including ULAN. Not yet of good quality because of e.g. duplicates. eero.hyvonen@aalto.fi

Ontology of Historical Places in Finland SAPO	http://onki.fi/en/browser/overview/sapo	FI	CC BY	Finnish county history since late 1800 and some earlier places from Sweden-Finland. Planned to be extended. eero.hyvonen@aalto.fi
Various additional ontologies of the ONKI.fi service.	http://onki.fi/en/browser/	FI, SE, EN,	E.g., lots of biological name lists are available, such as birds (10 000) or mammals (6 000) of the world. eero.hyvonen@aalto.fi
MIMO vocabulary	http://www.mimo-db.eu/InstrumentsKeywords/2204	EN, FR, NL, IT, DE, SE	CC BY	rbailly@cite-musique.fr
National Library of Spain: Persons, Organizations, Subject Heading, Titles	http://datos.bne.es and http://datahub.io/dataset/datos-bne-es	ES and many other non-tagged languages	CC0	More than four million authorities (including persons, orgs and works) many of them linked to VIAF and DBpedia. Rich subject headings linked to LCSH. Linked Data, many formats available. Contact: dvila@fi.upm.es
Lista de Encabezamientos de Materia (Subject Headings) of the Spanish Public Library Network	http://datahub.io/dataset/lista-encabezamientos-materia	ES	CC0	Linked to LCSH and RAMEAU. Linked Data.
GND (Integrated Authority File)	General information http://www.dnb.de/EN/gnd RDF-Dumps http://datendienst.d-nb.de/cgi-bin/mabit.pl?userID=opendata&pass=opendata&cmd=login	mainly de, some other non-tagged languages	CC0	Updates in January, May and September. Links to VIAF, LCSH, RAMEAU, STW and others coming up. Contains Persons, corporate bodies, subject headings, works, events and more.
STW (Standardthesaurus Wirtschaft; Thesaurus for Economics)	http://zbw.eu/stw/	de, en	CC BY-NC-SA	Linked to GND and DBpedia
TheSoz (Thesaurus Sozialwissenschaften; Thesaurus for the Social Sciences)	http://datahub.io/dataset/gesis-thesoz	de	CC BY-NC-ND	Links to DBpedia

DDC (Dewey Decimal Classification)	http://dewey.info/	several (depends on edition)	CC BY-NC-ND (depends on edition and language)	
ISNI (ISO certified global standard)	http://www.isni.org	Names with linguistic and transliteration variances	copyright	6.4 million individuals 400,000 organisations The core metadata is freely available for viewing via the web site. Access is also available via a search API. Each assigned ISNI is accessible by a persistent URI.
AgroVoc	http://aims.fao.org/standards/agrovoc/ , LOD at http://aims.fao.org/standards/agrovoc/linked-open-data	Arabic, Chinese, Czech, English, French, German, Hindi, Hungarian, Italian, Japanese, Korean, Lao, Persian, Polish, Portuguese, Russian, Slovak, Spanish, Thai, Turkish (depending on state of translation)	CC BY-NC-SA (for English, French, Russian and Spanish), for other languages copyright rest with the responsible institution	32,000 concepts covering all areas of interest to FAO, including food, nutrition, agriculture, fisheries, forestry, environment etc.