Technical Report

Quality evaluation of the LIDO and EDM Metadata that ATHENA, Linked heritage, AthenaPlus and MUSEU supplied Europeana with.

(tender reference number: 2016/MCA/09/002)

Institute of Communication and Computer Systems (ICCS)

20 July 2017
Athens, Greece

Table of Contents

1. Background

Quality of metadata has been an issue for Europeana from its early days and one of the main reasons that led to the shift from Europeana Semantic Elements (ESE) to Europeana Data Model (EDM).
In order to understand the problem of metadata quality we first have to go through the process the providers follow in order to publish their metadata to Europeana.

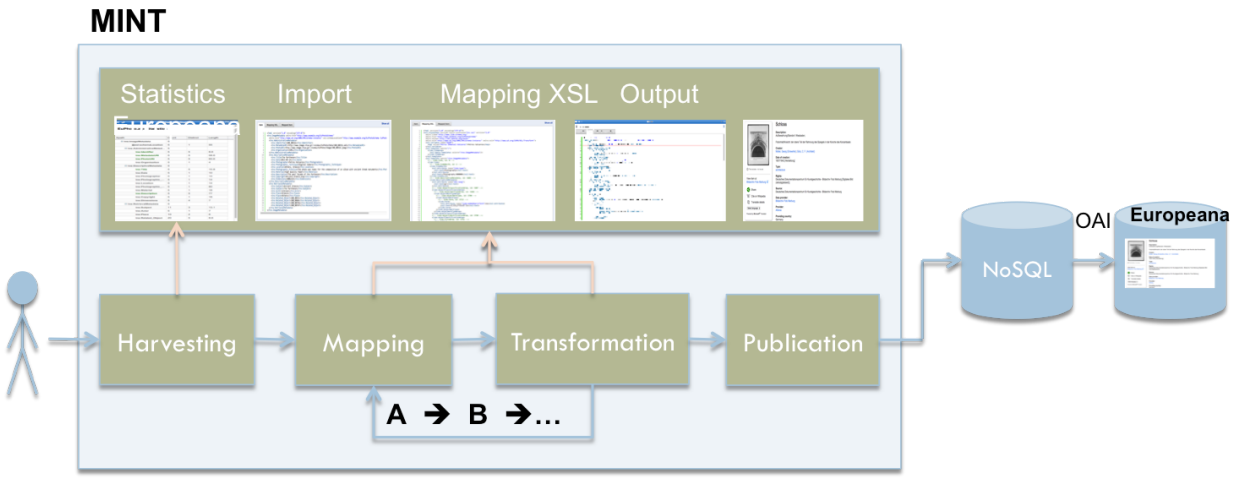

Figure 1. Europeana Publication Flow

First step is importing datasets of records in Mint.After this creation of a mapping from the imported metadata schema to the target metadata schema that is LIDO. This is done through the Mapping editor of the MINT mapping tool. During this step the provider can preview the input metadata, the crosswalk (XSL) she created through the mapping editor from the imported schema to LIDO, the metadata in LIDO and EDM and also the Europeana preview.
It is also important to mention that apart from the preview interfaces an inline XSD and schematron validation is also provided for LIDO and EDM. In that way content providers have full control of the produced metadata and how their records will look like when published on Europeana.
Once a valid mapping is created, the provider can use it to transform the metadata and then publish them – by sending them to the MongoDB NOSQL database that exposes them using the OAI protocol. The actual publication in Europeana happens at the end of every month, when the metadata is harvested and checked on this server by the Europeana Ingestion office.
It is worth mentioning that valid EDM records can be discarded during this check by Europeana due to mistaken rights or low quality metadata (broken links, non-sense description of DCHOs). Providers can produce good quality and valid metadata using these features but this is not always the case. They can also skip them by just creating a valid mapping to LIDO using only few of the project recommended elements and to publish in-expressive EDM records. The only way of ensuring that one of the main objectives of the project, that is the delivery of high quality

metadata to Europeana, is achieved is a quality evaluation on them before the Europeana harvesting process. The production of high quality metadata has been one of the Europeana main objectives towards this direction a set of bookmarks has been created into the MINT mapping tool pointing to specific LIDO elements. These in turn end up to specific EDM elements, ensuring a minimum set of elements required for the description of a Digitized Cultural heritage record.

2. Overview

In this document we describe our efforts to collect data statistics and evaluate the LIDO records that were published during these projects's publication process. Our focus is to present the progress of the 'metadata quality' of LIDO records as a metric of the appearance of elements of specific interest that fall into 4 categories.

- Digital Objects – Web Resources
- Geo-spatial information
- Vocabularies
- Events

2.1 Occurrence Rate

As a metric of metadata quality in the following chapters we will use the rate of the number of occurrences of a particular LIDO or EDM element of interest as it is described by it's XML XPath over the overall number of the records in the dataset. As an example for the Museu project published EDM dataset we counted 243874 records. While the number of 'edm:object' elements in all records was 165865. Thus we calculate the **Occurrence Rate** of 'edm:object' within the published Museu records to be :

Edm object Occurrence Rate 165865 / 243874 = 0.68

This would be mean that on an average the 85.5% of the records published contains the edm:object property. Within the tables in this document we present and compare the value of the "occurrence rate" of an element of particular interest in a dataset.

2.2 Data input

In order to compare and evaluate the use of LIDO and EDM we used as an input source all LIDO and EDM records published during Museu, AthenaPlus, LinkedHeritage and also 75% of the Photography published records.

We are using the Photography project as a reference because of the similar way LIDO and EDM were used, and mostly because a Special Metadata Task Force was created for this project for ensuring the delivery of high quality metadata to Europeana.

2.3 Digital Objects

As one would expect all records appear to carry at least one linkresource. During the creation of mapping to LIDO or EDM schema in the Mint Mapping tool a record is required to have at least one. The number of link resources is significantly increased in Musu and Photography.

2.3.1 LIDO Digital Objects

In LIDO most records tend to have more than one link resource on average.

LIDO Provider	Unique link resource occurrence rate
Photography	1.347
Museu	1.671
AthenaPlus	0.95
LinkedHeritage	1.211

Table 1. LIDO link resource occurrence rates per project

In this table we can read that in the Photography published set of metadata records on average 1.3 link resource elements exist for every record. Respectively 1,6 for Museu 0.95 for AthenaPlus.The table contents are shown graphically in the following chart.

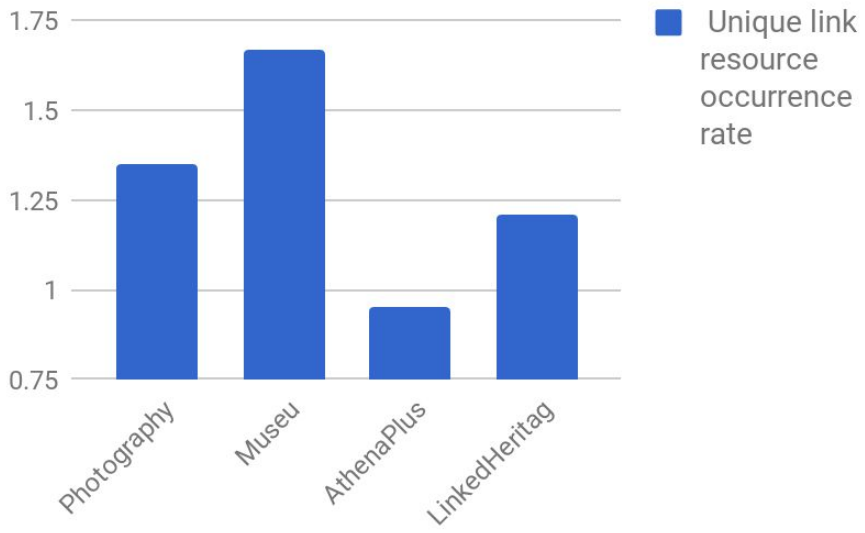

Figure 2. LIDO ink resource occurrence rates per project

2.3.2 EDM Digital Objects

EDM Provider	Unique isShownBy occurrence rate	Unique isShownAt occurrence rate	Unique edm object occurrence rate	Unique hasview occurrence rate
Photography	0.955	0.999	0.855	0.019
Museu	1.01	0.879	0.68	0.698

Table 2. EDM digital objects occurrence rates per project

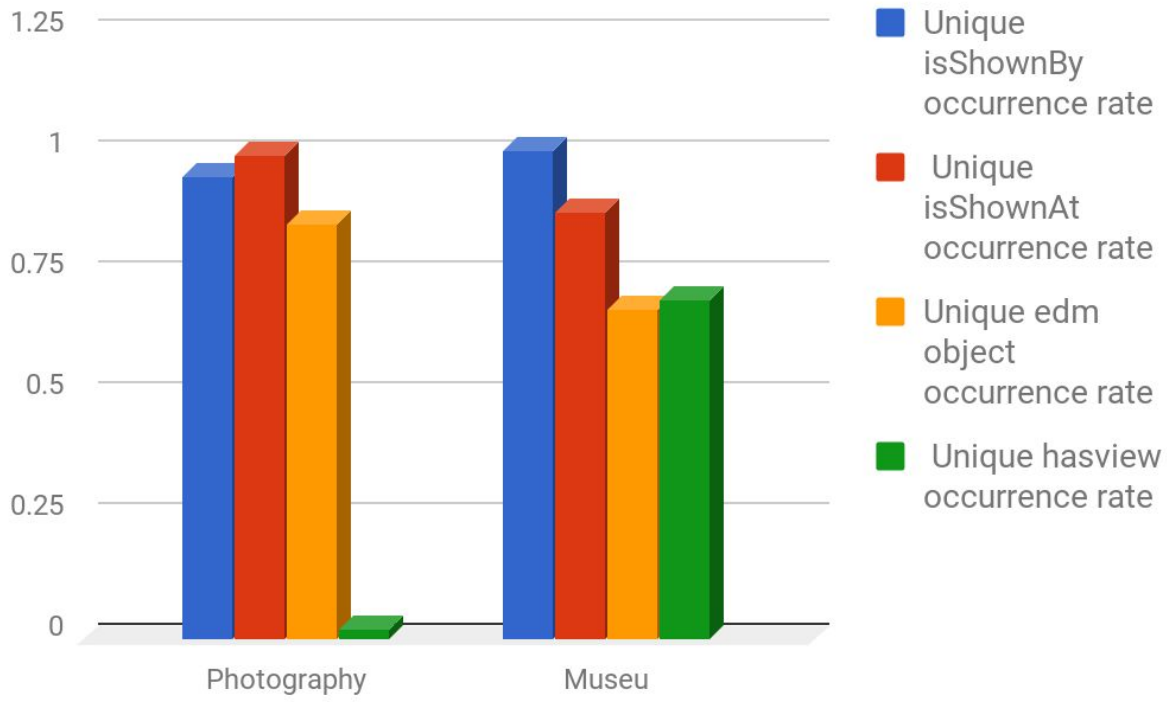

Figure 3. EDM Digital Objects occurrence Rates per project

2.4 Geo-spatial Information

2.4.1 LIDO Repository Location

The increase in the rate of occurrences of geo spatial information is remarkable for the LIDO schema. In Museu on average more spatial related elements are declared compare to the older AthenaPlus and LinkedHeritage projects.This is the result of better mapping guidelines and also of the use of a bookmarked elements list to map during the mapping creation phase in the Mint Mapping tool.

LIDO Provider	repositoryLocation occurrence rate	RepositoryLocation/placeId occurrence rate	repositoryLocation/partOfPlace occurrence rate	repositoryLocation/partOfPlace/placeId occurrence rate
Photography	0.073	0	0	0
Museu	0.951	0.629	0.951	0.793
AthenaPlus	0.005	0	0	0
LinkedHeritage	0.289	0.006	0.013	0.006

Table 3. LIDO Repository spatial elements occurrence rates per project

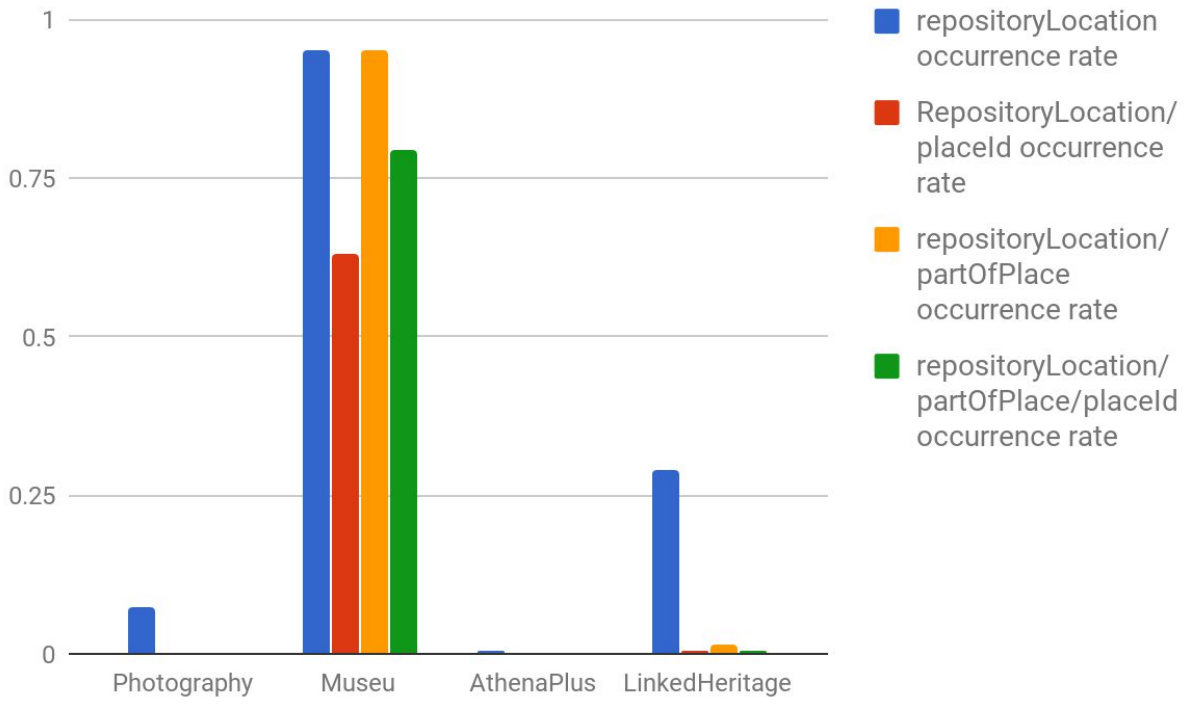

Figure 4. LIDO Repository spatial elements occurrence rates per project

2.4.2 LIDO Event Places

LIDO Provider	/event/eventPlace occurrence rate	/event/eventPlace/place occurrence rate	/event/eventPlace/place/placeID occurrence rate
Photography	1.689	0.729	0.051
Museu	0.305	0.196	0.141
AthenaPlus	0.804	0.773	0
LinkedHeritage	0.538	0.176	0.092

Table 4. LIDO Event spatial elements occurrence rates per project

For Event places the Photography project has much more occurrences of this elements. This happens mainly because of more event elements bookmarked for the particular Photography project in the Mint Mapping Tool. This list of bookmarks guides the user in a friendly and efficient way to create a more complete mapping and producing better quality metadata.

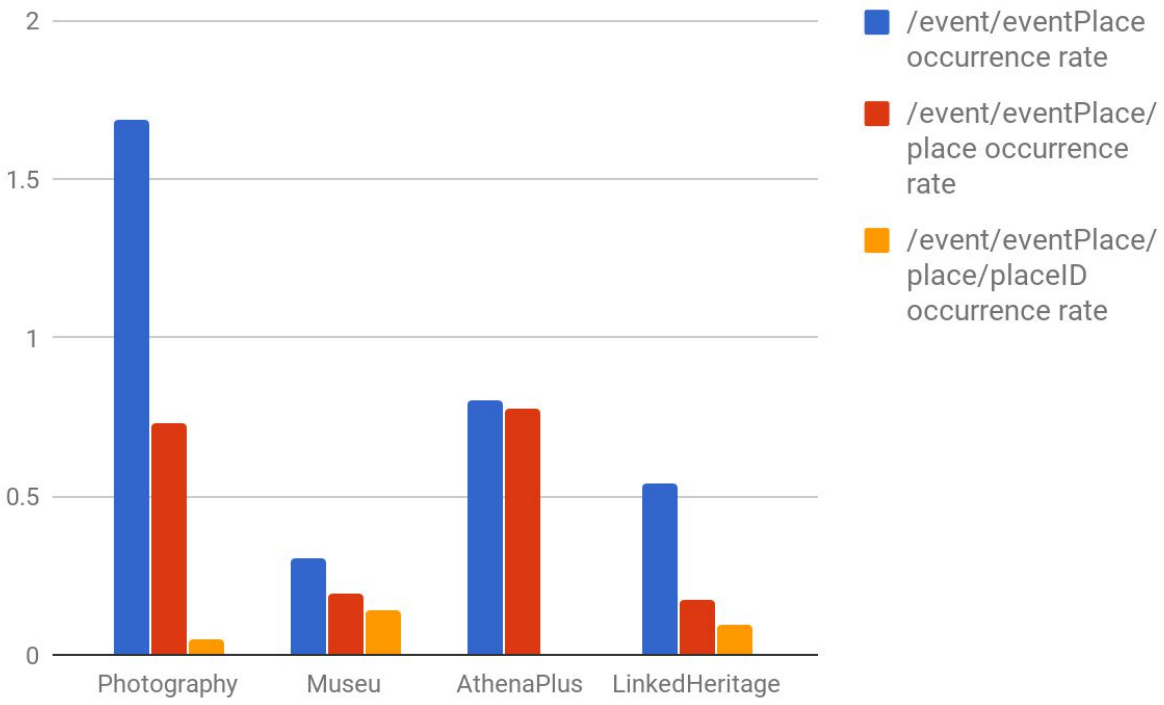

Table 5. LIDO Event spatial elements occurrence rates per project

2.4.3 LIDO Subject Places

Far more Subject spatial information has been created during the Museu project compared to AthenaPlus and LinkedHeritage projects.

LIDO Provider	/subject/subjectPlace occurrence rate	/subject/subjectPlace/place occurrence rate	/subject/subjectPlace/place/placeID occurrence rate	/subject/subjectPlace/displayPlace occurrence rate
Photography	0.722	0.305	0	0.305

Museu	0.645	0.388	0.041	0.275
AthenaPlus	0.002	0.002	0	0
LinkedHeritage	0.005	0.004	0.003	0.001

Table 6. LIDO Subject spatial elements occurrence rates per project

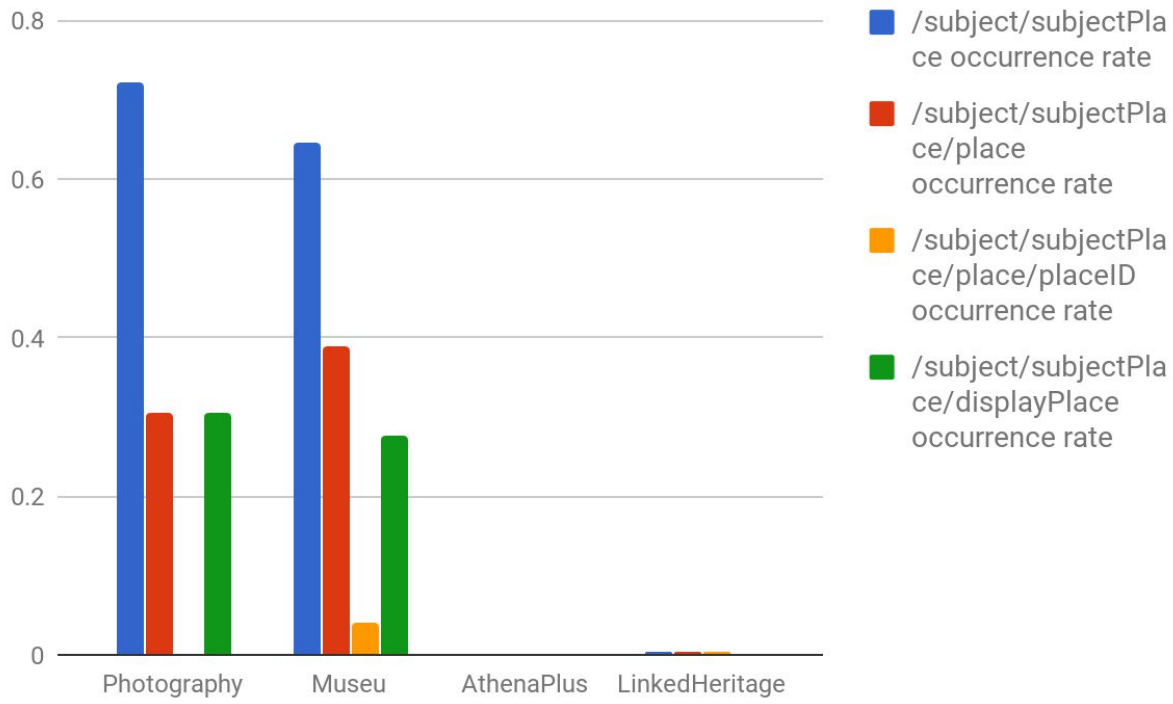

Figure 5. LIDO Subject spatial elements occurrence rates per project

Much increased are the occurrence rates of Subject spatial elements in Museu compared to LinkedHeritage and Museu provided records.Approximately a 65% of the records contain a Subject place.

EDM Spatial

EDM Provider	/RDF/ProvidedCHO/spatial occurrence rate	/RDF/ProvidedCHO/spatial/@lang occurrence rate	/RDF/ProvidedCHO/spatial/@resource occurrence rate	/RDF/ProvidedCHO/currentLocation	/RDF/Place occurrence rate

Photography	1.415	0	0	0	0
Museu	0.417	0.218	0.001	0.001	0

Table 7. EDM spatial elements occurrence rates per project

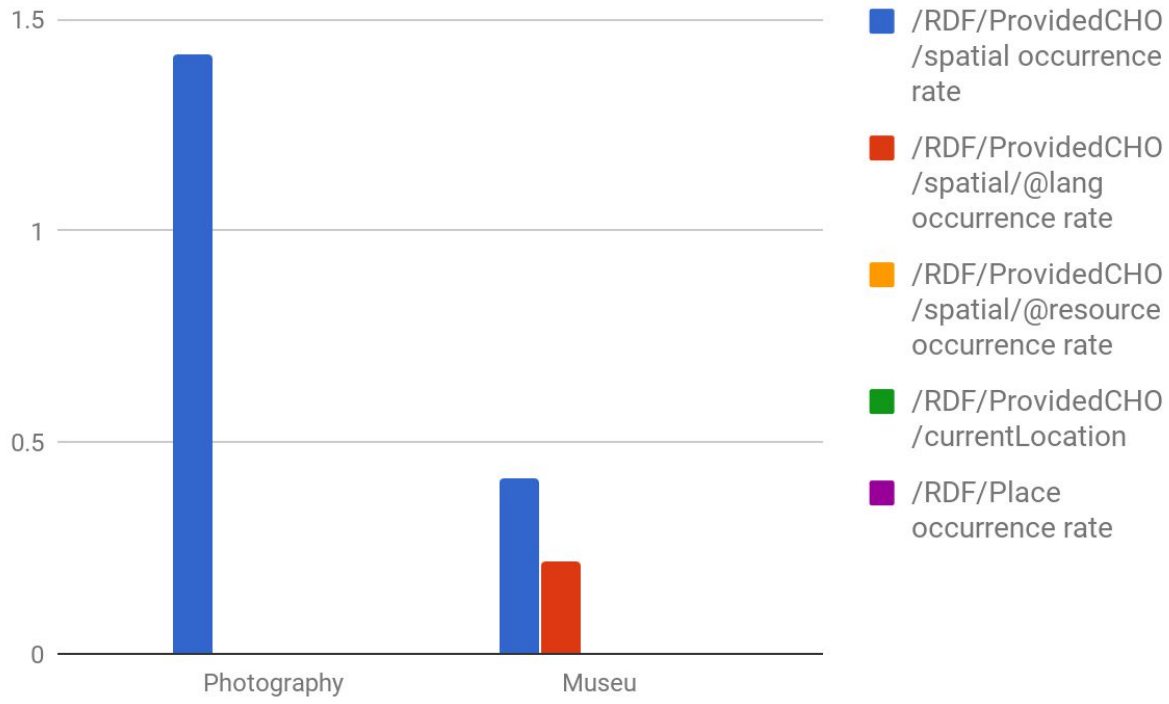

Figure 6. EDM spatial elements occurrence rates per project

Compared to the records provided for Photography, the Museu records present a smaller occurrence rate.Forty per cent 40% of the records carry a spatial element.More than 20% contain a @lang attribute for the spatial element.

2.5 Vocabularies

2.5.1 LIDO Object Work Type / Classification Vocabularies

In the following table we can see the objectWorkTypes and its respective conceptID element occurrence rates. The comparison between the two values gives us a metric of Vocabulary use by the provider for this element.

LIDO Provider	/objectWorkTypeWrap occurrence rate	objectWorkTypeWrap/objectWorkType/conceptID occurrence rate	objectClassificationWrap/classificationWrap/classification occurrence rate	objectClassificationWrap/classificationWrap/classification/conceptID occurrence rate
Photography	1.039	0.004	1.039	0.005
Museu	1	0.306	3.161	2.092
AthenaPlus	1	0	0	0
LinkedHeritage	1.002	0.028	1.367	0.229

Table 8. LIDO objectWorkType/Classification elements occurrence rates per project

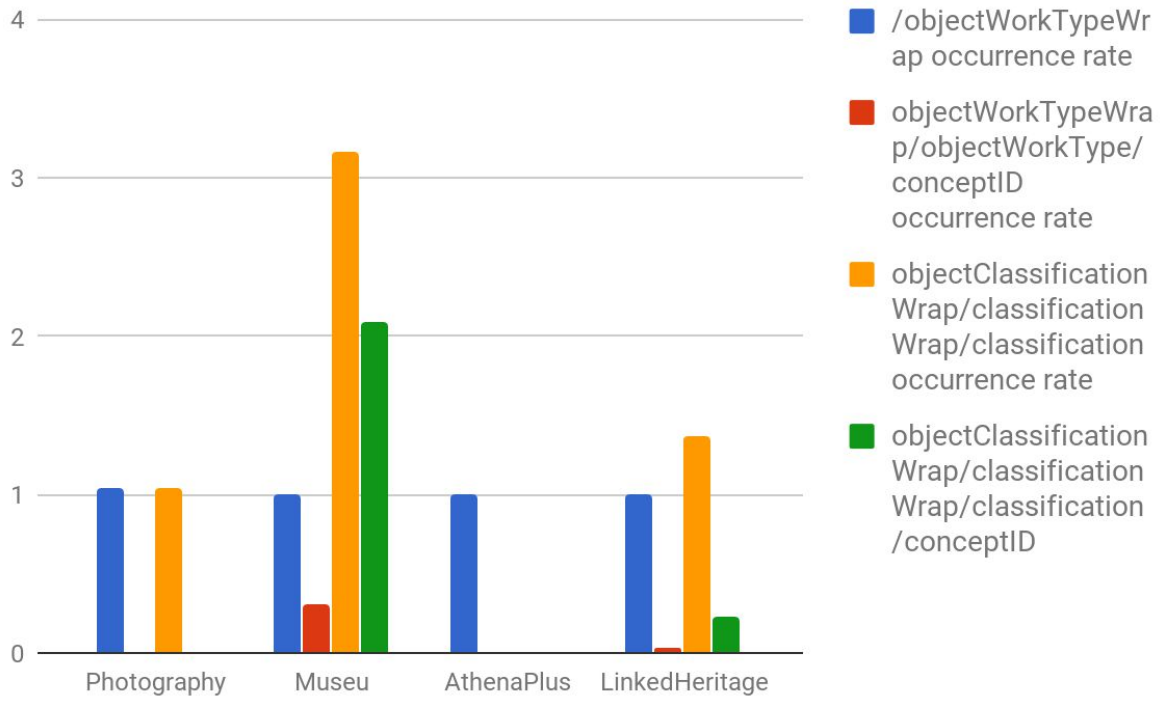

Figure 7. LIDO objectWorkType/Classification elements occurrence rates per project

The use of vocabulary in the object's Classification in Museu is much increased when compared to other projects datasets .

2.5.2 LIDO Event Vocabularies

Here we present the rate of occurrences of LIDO Event elements through the use of Vocabularies.Event Elements where a conceptId sub-elemen exists indicate the use of a Vocabulary.Thus for the Museu project more than 40% of thne provided records are linked to a Vocabulary value.

LIDO Provider	/event/eventType occurrence rate	/eventType/conceptID occurrence rate	/event/eventActor/actorInRole/roleActor per actor	/eventActor/actorInRole/roleActor/conceptID	event/eventMaterialsTech/materialsTech/termMaterialsTech occurrence rate	eventMaterialsTech/materialsTech/termMaterialsTech/conceptID occurrence rate
Photography	1.006	1.006	0	0	1.111	0.077
Museu	1.096	0.401	0.59	0	0.89	0.003
AthenaPlus	1.798	0.877	0	0	0	0
LinkedHeritage	1.373	0.13	0.462	0	1.195	0.34

Table 9. LIDO Event elements occurrence rates per project

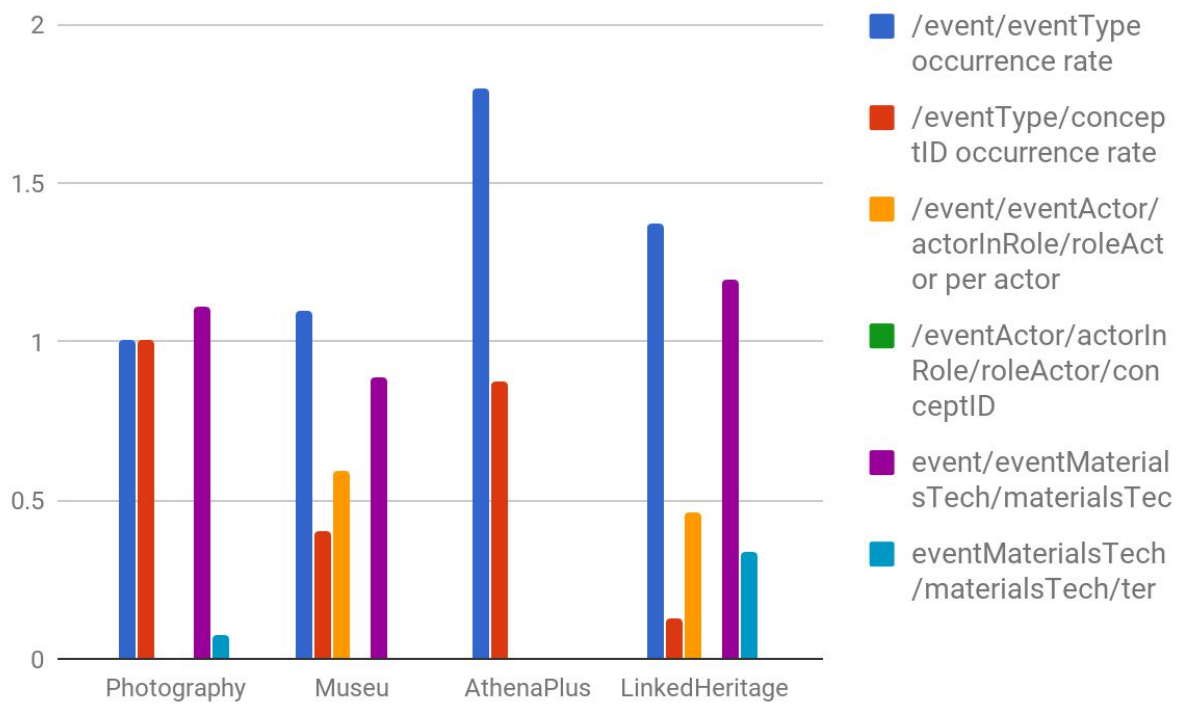

Figure 8. LIDO Event elements occurrence rates per project

2.5.3 LIDO Subject Vocabularies

The usage of Vocabularies in LIDO subjects is notably increased in the Museu dataset, compared to the LinkedHeritage and AthenaPlus. For Photography the use of Subject vocabulary is increased when compared to other projects because more vocabularies are used for this project.

LIDO Provider	subjectWrap/subjectSet/subject/subjectConcept occurrence rate	subjectWrap/subjectSet/subject/subjectConcept/conceptID occurrence rate
Photography	2.188	6.562
Museu	0.555	0.732
AthenaPlus	0.773	0
LinkedHeritage	0.197	0.11

Table 10. LIDO Subject elements occurrence rates per project

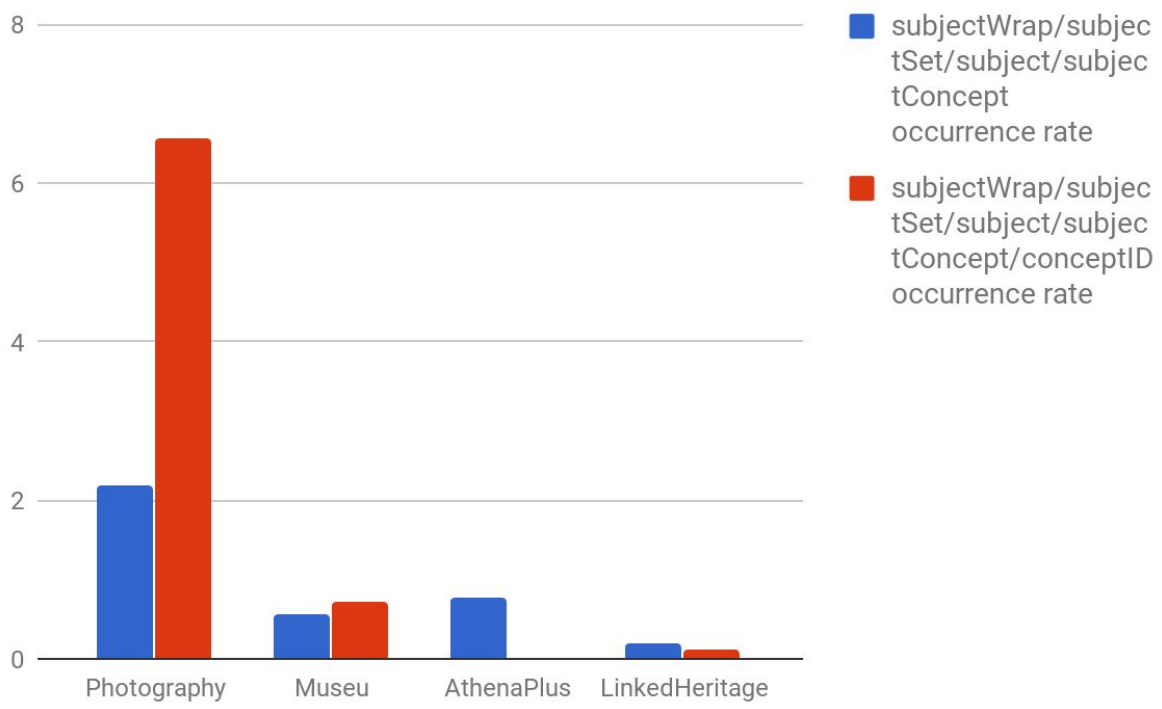

Figure 9. LIDO Subject elements occurrence rates per project

2.5.4 EDM Vocabularies

The Photography dataset appears have much more increased vocabulary usage for the Subject elements both in LIDO and EDM, because of the extra vocabularies used for this project. In the following table the @resource elements resource indicate the use of a Vocabulary value in this element.

EDM Provider	/RDF/ProvidedCHO/format occurrence rate	/RDF/ProvidedCHO/format/@resource occurrence rate	/RDF/ProvidedCHO/subject occurrence rate	/RDF/ProvidedCHO/subject/@resource occurrence rate	/RDF/ProvidedCHO/medium occurrence rate	/RDF/ProvidedCHO/medium/@resource occurrence rate	/RDF/Concept occurrence rate	/RDF/Concept/broader/@resource occurrence rate
Photography	1.775	0.893	9.735	1.681	0.43	0.048	3.08	2.778
Museu	0.613	0.03	0.786	0.005	0.349	0.035	1.124	0.004

Table 11. EDM Vocabulary elements occurrence rates per project

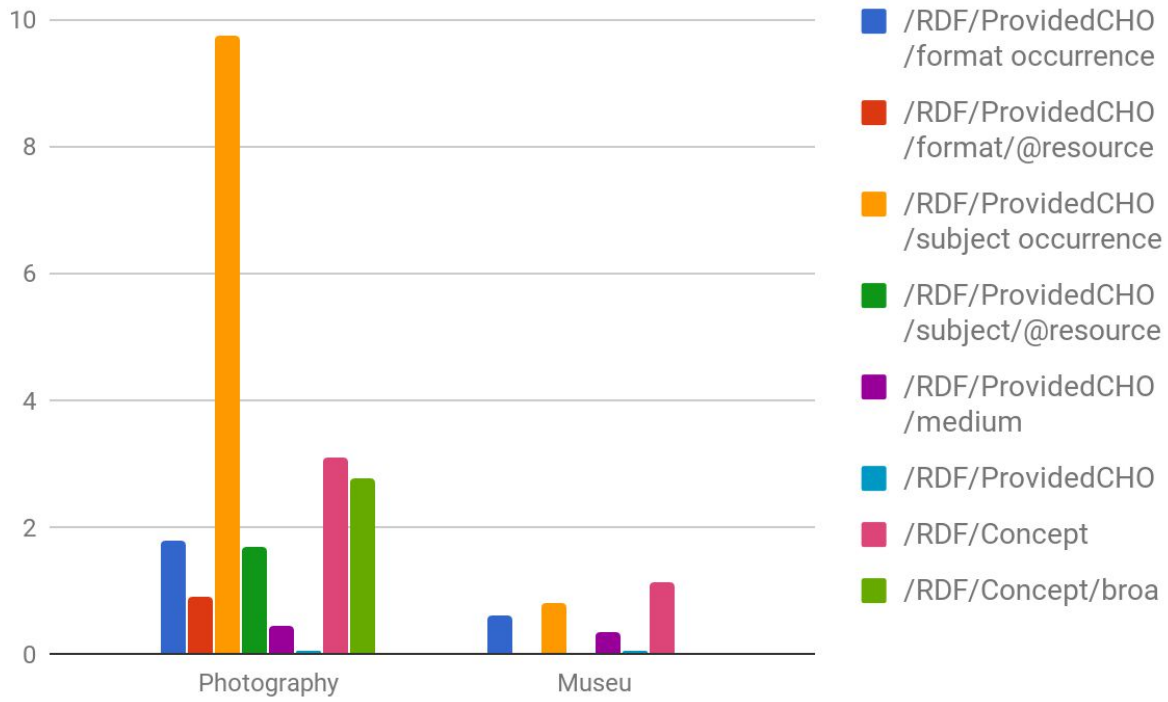

Figure 10. EDM Vocabulary elements occurrence rates per project

2.6 Events

The usage of LIDO Events is highly encouraged during the metadata mapping stage.
This creates a highly structured set of information that contains spatial, time, involved persons Information.

2.6.1 LIDO Events

LIDO Provider	/eventSet/event occurrence rate	eventSet/event/ eventType/con ceptID occurrence rate	eventSet/event/ eventType/term occurrence rate
Photography	1.006	1.006	0.227

Museu	1.096	0.401	1.081
AthenaPlus	1.798	0.877	0.851
LinkedHeritage	1.373	0.13	1.681

Table 12. LIDO Event elements occurrence rates per project

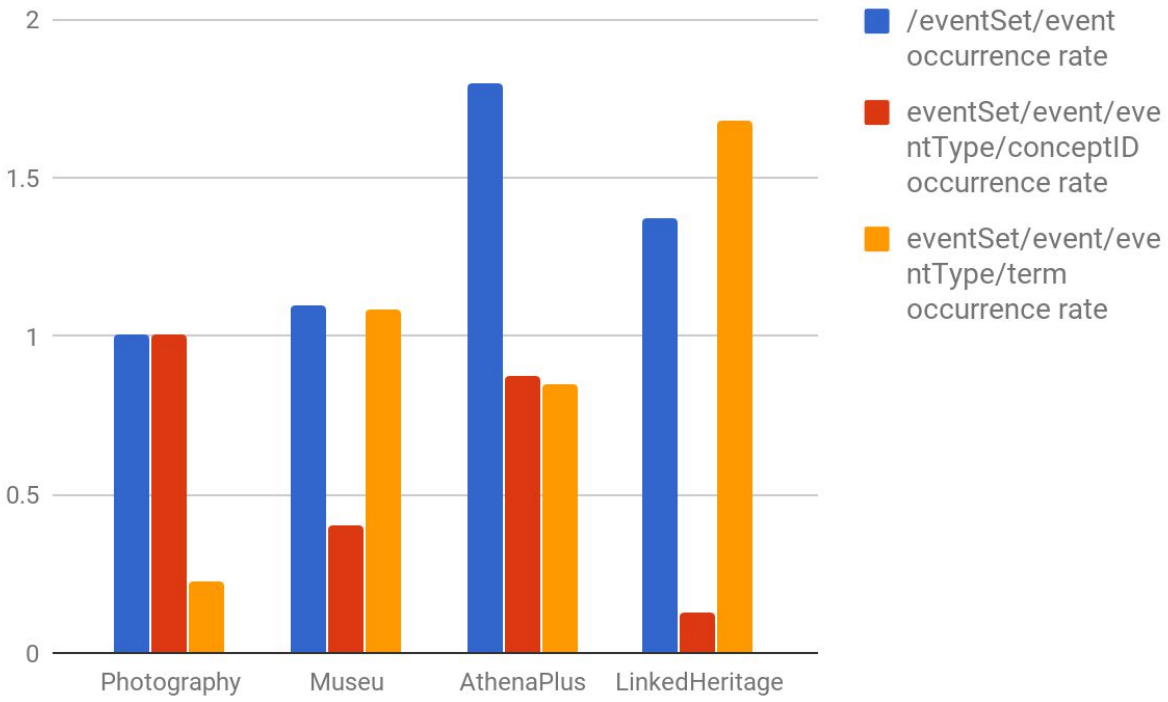

Figure 11. LIDO Event elements occurrence rates per project

2.6.2 LIDO Event Details

LIDO Provider	eventSet/event/eventDate/date occurrence rate	/event/eventMaterialsTech/materialsTech/termMaterialsTech occurrence rate	/eventSet/event/eventPlace/place occurrence rate	eventSet/event/eventMethod occurrence rate	event/eventActor/actorInRole/actor/nameActorSet/appellationValue per actor
Photography	0.909	1.111	0.729	0.938	1.018

Museu	0.93	1.35	0.076	0.042	0.529
AthenaPlus	0.697	1.998	1.81	0	0.107
LinkedHeritage	0.274	1.195	0.176	0	0.644

Table 13. LIDO Event details elements occurrence rates per project

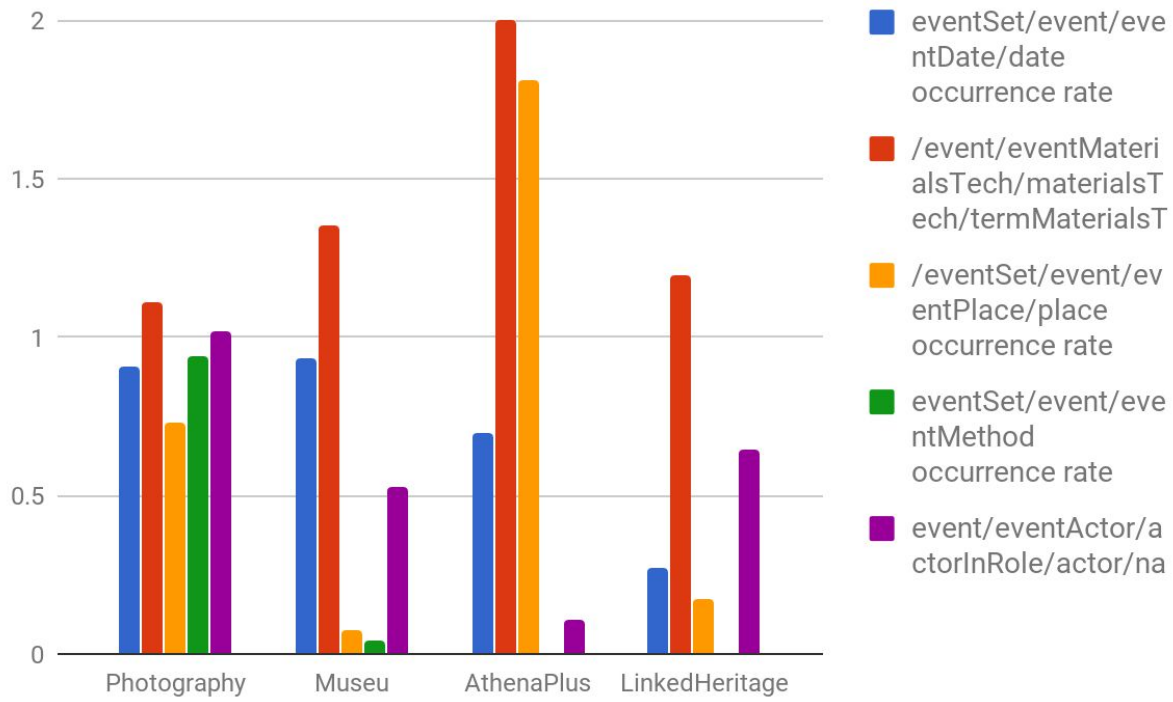

Figure 12. LIDO Event details elements occurrence rates per project

2.6.3 EDM Events

After the transformation of a LIDO record to EDM, LIDO Events, are transformed to various EDM elements like date contributor, format. The Photography dataset appears to have much more occurrences of these elements because of the increased number of vocabularies and mapping guidelines for photographies.

Xpath / Provider EDM	Photography	Museu

/RDF/ProvidedCHO/date occurrence rate	0.024	0.606
/RDF/ProvidedCHO/contributor occurrence rate	0.018	0.568
/RDF/ProvidedCHO/format occurrence rate	1.775	0.003
/RDF/ProvidedCHO/publisher occurrence rate	0.014	0.034
/RDF/ProvidedCHO/created occurrence rate	0.898	0.273
/RDF/ProvidedCHO/medium occurrence rate	0.43	0.349
/RDF/ProvidedCHO/spatial occurrence rate	1.434	0.426
/RDF/Agent occurrence rate	0	0.031
edm place occurrence rate	0	0.001
/RDF/ProvidedCHO/creator occurrence rate	1.017	0.266

Table 14. *LIDO Event elements occurrence rates per project*

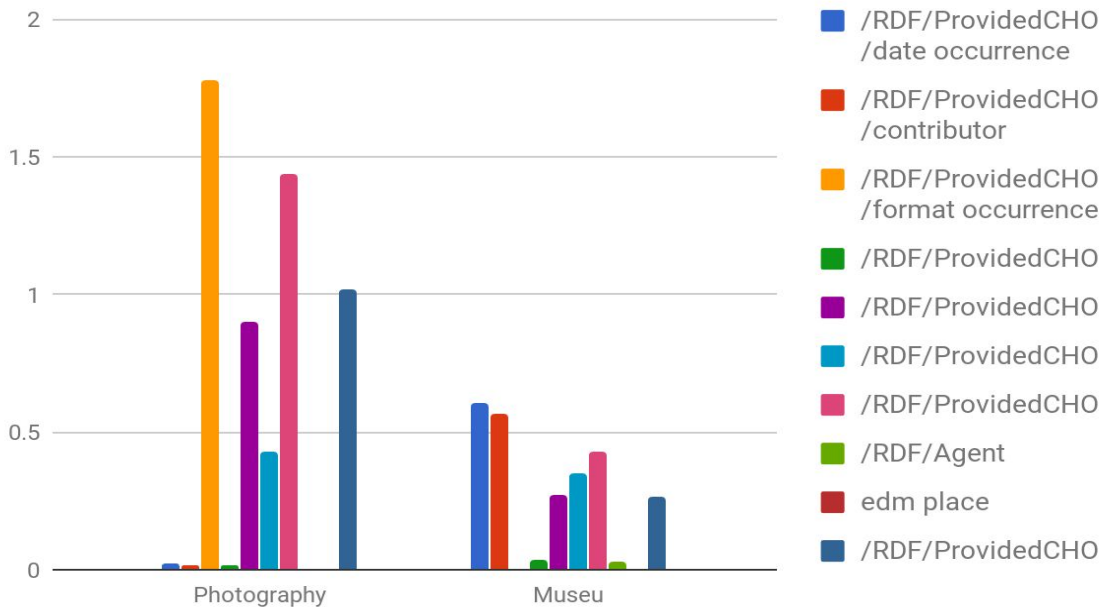

Figure 13. EDM Event-generated details elements occurrence rates per project

3. Contribution and Suggestions

Throughout this document we presented an evaluation of LIDO metadata. The quality of metadata was improved in time as Museu dataset records, appear to have more occurrences of specific elements of interest. The importance of useful mapping guidelines for content Providers should be noted. Also the encouragement of good mapping practices like the the use of a list of bookmarked elements connected with vocabularies in the Mint Mapping Editor, leads clearly to more expressive metadata of higher quality.

The metadata ‘quality’ metric’ is a field of future research that could lead to the creation of a ‘metadata quality’’ ranking system. We would like to propose the design of this ranking system to Europeana, as this could provide a way to automatically rank every EDM and LIDO record. This could be done on the fly for every record stored in the OAI. A procedure like this could make much easier the process of Europeana records filtering based on ‘metadata’ quality.

A ranking system would take into account the significance of the different metadata elements in the record, and also the values that they contain. Special elements could carry an increased weight depending on their level of importance and or their quality of their values as a metric of number of different words, or linked resources and the use of vocabularies.

Such a system has been proposed in the past by Kapidakis in his ‘Comparing Metadata Quality in the Europeana Context’ paper.This paper could be the basis for further research on a matada quality ranking system.