Provenance of metadata enrichments and translations in EDM *External*

EDM Extension

Version	Beta
Author	Hugo Manguinhas
Reviewers	Antoine Isaac, Valentine Charles
Date	Nov 2021
Collaboration	Beta version was developed under the Europeana Generic Services Europeana XX Century of Change project[1]. Funded under the Connecting Europe Facility Programme under by European Health and Digital Executive Agency (HaDEA)

European Commission

European Health and Digital Executive Agency (HaDEA)

Connecting Europe Facility

[1] https://pro.europeana.eu/project/europeana-xx

1. Introduction and background

In the scope of the Europeana XX GS project, several activities involving automated (and/or user-assisted) software for translation and enrichment have been planned to improve the quality of the metadata that is provided by the data providers. In the past, similar activities being applied in the aggregation chain have resulted in an altered version of the metadata being provided to Europeana without any differentiation between what was part of the original metadata record and changes to it.

Considering that many of the enrichment activities that are applied to the metadata involve automated processes that even though may offer a certain level of accuracy, there is still a margin of error that may result in unwanted metadata. On the other hand, other enrichment activities that pass through a quality assurance done by users leave a much lesser margin for issues with metadata to slip through. Retaining the provenance of the information is therefore key to distinguish the information and make a better judgment over its quality when presenting to the end-user.

This document describes an extension to EDM that supports the representation of provenance information to be indicated in the data package that is delivered to Europeana. This particular extension is limited to the EDM external representation in EDM.xsd (RDF/XML format). This is proposed as a pilot so we can evaluate the effectiveness of the approach.

2. Extension to EDM External

We chose to put forward an option that would make life easier for providers to indicate the provenance information which involves the provision of only 2 new XML attributes to the RDF/XML representation. Due to this design choice, this option is limited to EDM external data represented according to the EDM.xsd. This representation will result in a more formal but complex version when exposed in EDM Internal through the Europeana APIs as described in the companion document.

Indicating provenance

Whenever a new metadata statement is added to the metadata record, this metadata statement must be marked using a new XML attribute “edm:wasGeneratedBy”[2] indicating whether it was generated by a person or a software agent (e.g. enrichment or translation tool). The value “Person”[3] must be used to indicate that the metadata statement was provided by a person, while the value “SoftwareAgent”[4] must be used to indicate that the metadata statement was generated by a software tool.

In the case that the metadata statement was generated by a software agent but was later reviewed and approved by a person, it must be considered as if the metadata statement was created by a person and therefore indicated with value “Person”. In case this is performed only to a sample of items, only the items in the sample can be marked as being generated by a person.

Indicating the confidence level of the new metadata statement

A confidence level can be indicated at the level of each metadata statement by using a special XML attribute “edm:confidenceLevel”[5] and indicating a [0,1] floating number as value. This value should be indicated whenever the software tool is capable of outputting the value and it must reflect only that metadata statement and not an overall measurement of confidence for the whole record. In the case where the metadata statement was created by a person, the confidence level can be omitted.

3. Some examples

3.1. Provision of metadata statements resulting from semantic enrichment

The example below presents a metadata record where the “dcterms:spatial” metadata statement was semantically enriched resulting in a new metadata statement referring to the Geonames entity corresponding to “London”. The new metadata statement is marked as being generated by a software tool (ie. enrichment service) and indicating a 0.9 level of confidence.

```
<edm:ProvidedCHO rdf:about="...">
  ...
  <dcterms:spatial>London</dcterms:spatial>
  <dcterms:spatial edm:wasGeneratedBy="SoftwareAgent" edm:confidenceLevel="0.9"
rdf:resource="http://geonames.org/2643743"/>
  ...
</edm:ProvidedCHO>
```

3.2. Provision of metadata statements resulting from machine translation

The example below presents a metadata record where the dc:title metadata statement was translated to French resulting in a new metadata statement with the French translation of the title and indicating the language within xml:lang. The new metadata statement is marked as being generated by a software tool (ie. translation service) and indicating a 0.8 level of confidence.

---

[2] This attribute was inspired by the PROV ontology but was migrated to EDM given that the range differs.

[3] Corresponds to “foaf:Person” mirroring what is specified in W3C Web Annotations Data Model and EDM profile for Annotations

[4] Corresponds to “prov:SoftwareAgent”.

[5] At the time of writing, the value is defined as a percentage of certainty but it may be reviewed to represent a degree (ie. high, medium, low) of certainty instead. No specific ontology was found as inspiration however ORCA could be a possible candidate.

```
<edm:ProvidedCHO rdf:about="...">
  ...
  <dc:title>Wimbledon School Of Art</dc:title>
  <dc:title xml:lang="fr" edm:wasGeneratedBy="SoftwareAgent"
edm:confidenceLevel="0.8">École d'art de Wimbledon</dc:title>
  ...
</edm:ProvidedCHO>
```

Additionally as presented in the next example, if the translation service was able to identify the language of the original dc:title metadata statement (considering that it was able to translate to English), a new metadata statement should also be generated with the same value but indicating now the language within xml:lang. This 2nd metadata statement should also be marked as being generated by a software tool and when possible indicate the confidence level reflecting only the language detection aspect (and not the confidence level of the translation).

```
<edm:ProvidedCHO rdf:about="...">
  ...
  <dc:title>Wimbledon School Of Art</dc:title>
  <dc:title xml:lang="en" edm:wasGeneratedBy="SoftwareAgent"
edm:confidenceLevel="0.9">Wimbledon School Of Art</dc:title>
  <dc:title xml:lang="fr" edm:wasGeneratedBy="SoftwareAgent"
edm:confidenceLevel="0.8">École d'art de Wimbledon</dc:title>
  ...
</edm:ProvidedCHO>
```

3.3. Provision of metadata statements created (or approved) by persons

The example below presents a metadata record where the “dc:creator” metadata statement was semantically enriched by an enrichment tool but was reviewed and approved by a person. The new metadata statement is marked as being generated by a person but not indicating a level of confidence.

```
<edm:ProvidedCHO rdf:about="...">
  ...
  <dc:creator>Emanuel Ungaro (Designer)</dc:creator>
  <dc:creator edm:wasGeneratedBy="Person"
rdf:resource="http://wikidata.org/entity/Q529203"/>
  ...
</edm:ProvidedCHO>
```

4. Open questions

The proposal presented in this document covers enrichments that are meant to add information to the metadata. It is unclear at this point whether there will be use cases for correcting metadata which will warrant a review of the proposal.

Additionally, the proposal does not retain information about which task/software (e.g. automatic translation, language detection, semantic enrichment) resulted in the metadata enrichment as there has not yet been identified a need for this information.