→ EDM Case Study: Europeana Linked Open Data pilot

Europeana has released its own Linked Open Data (LOD) pilot, data.europeana.eu. This pilot is the first case of large-scale open release of Europeana data, sending signals to our community and mobilising our partners. It is also the second example of Europeana data released using EDM in combination with Linked Data[1] technology, including RDF.

We learnt a first series of useful lessons on porting legacy (ESE) metadata into EDM. Though it uses the existing ‘semantic enrichments’ done by Europeana, the LOD pilot does not create new metadata itself. The mapping exercise was thus about distributing fields from existing ESE records onto the new EDM resources (e.g. ore:Aggregation, ore:Proxy, edm:WebResource…) that correspond to each of these records. For example, the europeana:dataProvider field in ESE translates in EDM into an edm:dataProvider statement that is subjected to an aggregation resource.

Here, most issues come from providers’ not entirely following the ESE guidelines. This results in ambiguous values for certain fields, which cannot be attributed with certainty—i.e., for every object of the entire Europeana dataset—to the resources one would think of first. ESE quality issues will directly impact the quality of EDM data generated from it, if no data cleaning is carried out in the process.

The LOD pilot was also an opportunity to test and showcase how more advanced features of EDM (proxies) can be used to represent the semantic enrichment that Europeana creates.

[1] http://linkeddata.org

One key point was to allow distinguishing between original metadata from a provider and later enrichments by other actors. The LOD pilot shows how this can be done by creating several proxies that represent different perspectives on the same object, carrying different bits of metadata.

This however results in quite verbose data, especially compared to the ESE records we started with. It was an important lesson learnt from the LOD pilot: we should try to 'hide' this complexity when it is not needed, or reveal the full complexity and power of EDM in successive steps that make the full picture easier to understand for data providers and consumers alike. This has directly influenced the way we further specified how EDM should be used for data ingestion into Europeana.

Another point was the ease of representation of the enrichment data itself. Europeana semantic enrichment consists mostly of linking items to resources (places, concepts), which are already represented as Linked Data on other services, for instance geonames.org. Using RDF for representing such links (without duplicating the data available elsewhere for the linked resources) was quite straightforward. The main issue was the low granularity of the enrichment data served in the end; while the enrichment process creates semantic links from specific fields (e.g. dc:subject), that information is not recorded. As a result, we do not know whether, say, a given city is the subject (topic) of an item or its place of production. We had to use a quite abstract EDM property that merely expresses that the item is 'generally linked' to that place. This is in fact a first step towards determining what 'data quality' could mean for EDM.

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

Finally, the LOD pilot has provided a great opportunity for experimenting with RDF storage tools for Europeana metadata and evaluating their scalability. The current storage set-up does not deliver impressive performance for complex queries. Yet, it still allows for the simple service of RDF data for over two million items.

Further information on all these aspects is available at data.europeana.eu and the technical paper published in the proceedings of the 2011 Dublin Core conference.[2]

[2] Bernhard Haslhofer, Antoine Isaac. *data.europeana.eu - The Europeana Linked Open Data Pilot*. International Conference on Dublin Core and Metadata Applications (DC 2011). The Hague, 2011. Available at http://dcevents.dublincore.org/index.php/IntConf/dc-2011/paper/view/55