Submitted on: 8/11/2014

Links, languages and semantics: linked data approaches in The European Library and Europeana.

Valentine Charles
Europeana Foundation and The European Library, Den Haag, The Netherlands.
valentine.charles@europeana.eu

Nuno Freire
The European Library, Den Haag, The Netherlands
nuno.freire@theeuropeanlibrary.org

Antoine Isaac
Europeana Foundation, Den Haag, The Netherlands.
antoine.isaac@europeana.eu

Abstract:

The European Library and Europeana have both an extensive experience in aggregating metadata for bibliographical records or digital resources from the cultural heritage institutions of Europe. For both of them meeting the challenges offered by multilingual and heterogeneous data is an ongoing effort. The growth of the Semantic Web and the more generalised publication of knowledge organisation systems as linked open data offer the possibility to make these services truly multilingual.

This paper shows how The European Library and Europeana exploit the semantic relations and translations offered by knowledge organisations systems in order to solve the problem of data integration at a European scale. It also demonstrates the potential of Linked Open vocabularies for enabling multilingual search and retrieval services.

Keywords: libraries, knowledge organization systems, linked open data, multilingual, vocabulary alignment.

THE EUROPEAN LIBRARY AND EUROPEANA: TWO DATA AGGREGATION SERVICES

Europeana[1] provides access to digitised cultural resources from a wide range of cultural institutions (libraries, museums, archives...) all across Europe. The European Library[2] gives access to documents from the national and research libraries of Europe, playing the role of library-domain aggregator for Europeana.

Both initiatives seek to enable users to search and access knowledge in all the languages of Europe. This is done either directly, via their respective web portals, or indirectly, via third-party applications built on top of their data services (search APIs and Linked Open Data prototypes).

Both services are also based on the aggregation and exploitation of (meta)data about the digitized objects[3] from very different contexts. To provide a seamless, efficient services on top of such aggregation, they have to solve hard data integration issues. To address these, The European Library and Europeana have developed infrastructures and workflows for ingesting, indexing, normalising and publishing data.

The Europeana Network has defined the Europeana Data Model (EDM) to be used by Europeana as its single data model [1]. This approach to data harmonization is not monolithic, though: it embraces the Semantic Web principles to integrate in an open environment the various models used in Cultural Heritage data. EDM re-uses OAI-ORE[4], Dublin Core[5] SKOS[6], RDA Elements[7] and others. It is also aligned with CIDOC-CRM and FRBRoo[11]. It is also easy to extend EDM to implement richer functions for the Europeana service or more specific projects [12]. This is in line with the vision of *linked open vocabularies*, in which vocabularies can be seamlessly networked at the semantic level and across languages[8].

The European Library currently supports the ingestion of most common bibliographic standards used in European Libraries: MARC21, UNIMARC, MODS, METS, Dublin Core.... It also converts metadata to EDM. For The European Library, this facilitates the participation of the library-domain into the web of data with the cultural heritage institutions represented in Europeana.

Yet data modelling alone does not solve all complexity and heterogeneity issues. For example, The European Library works with libraries following similar data creation practices. Yet, in spite of the standardization efforts for cataloguing practices and bibliographic data models, we still find very heterogeneous data across libraries. Cataloguing rules leave room

[1] http://www.europeana.eu

[2] http://www.theeuropeanlibrary.org

[3] The European Library, addresses a more specific sector, which allows it to provide extra services based on full-text content and collection-level metadata.

[4] http://www.openarchives.org/ore/

[5] http://purl.org/dc/terms/

[6] http://www.w3.org/2004/02/skos/core#

[7] http://rdaregistry.info

[8] http://lov.okfn.org/dataset/lov/details/vocabulary_edm.html

for interpretation. The result of cataloguing can be too subtle to be fully represented in machine-processable structured fields; librarians often resort to general note fields. Information systems are not always up-to-date with the standards or do not fully enforce cataloguing practices. There is a lot of legacy data created according to older cataloguing rules or in older systems, which have been subjected to automatic data migration processes. Therefore, the same information may be represented quite differently from library to library, and even within the same library. Heterogeneity issues are of course even more numerous in Europeana's cross-sector metadata. In fact Europeana relies on its domain aggregators, like The European Library, to try and deal with the local practices of data providers.

On top of this, the dimension and richness of both datasets (language coverage, types of resources) calls for advanced user services, including homogeneous subject-based access, multilingual search, semantic relationships between the described resources or with the resources available in the Semantic Web.

EXPLOITATION OF MULTILINGUAL KNOWLEDGE ORGANISATION SYSTEMS IN THE EUROPEAN LIBRARY AND EUROPEANA

The integration of data in Europeana and The European Library comes with challenges; but there are unprecedented opportunities, too. As a first attempt to satisfy the above mentioned user expectations, both initiatives are exploring the generalized use of knowledge organisation systems (KOS) to get more from the objects' context [3]. This includes resources from multilingual "value vocabularies" like thesauri, authority lists, classifications, either coming from our network of providers or from third-party data sources, which we aim to gather in a so-called "semantic layer" [2].

In Europeana

Because of the diversity of the resources it collects, Europeana started to perform automatic metadata enrichment itself [9] with a selection of external value vocabularies and datasets. We defined requirements for selecting vocabularies that comply with its diversity of domains and languages:

- be technically available (through Linked Data or in dedicated repositories), properly documented, and in open access.
- be well-connected together, e.g., equivalent elements in other vocabularies are indicated
- be multilingual

Currently, Europeana continuously enriches objects by creating links to contextual resources[9]:

- 7M objects are connected to places from GeoNames[10],
- 144K objects are connected to agents from DBpedia[11]
- 9.1M objects are connected to concepts from GEMET[12] and DBpedia.

Using these reference resources, one can identify "who", "what", "when" and "where" foci in object metadata, and complete it with additional information such as translations. At the time

[9] Note that the figures also include the links to contextual resources delivered by data providers.
[10] http://geonames.org
[11] http://dbpedia.org
[12] http://www.eionet.europa.eu/gemet

of writing, enrichments are visible in our portal as links in individual metadata fields - see (1) in fig. 4 - and the "Auto-generated tags" foldout (2). Clicking on the foldout reveals labels associated with the matched resources - see (3) in fig. 5 - and their parent resources in the vocabularies (4). Multilingual semantic enrichment also enables users to retrieve more documents for a given query.

Fig 4: Links to vocabulary resources shown in object display

Fig 5: Data fetched from vocabularies after enrichment

The Europeana team is currently working on improving the display of contextual resources on the portal [13] and the automatic enrichment process, especially trying to cope with the challenges raised in a multilingual environment [13].

Finally, Europeana encourages its partners to provide contextual resources at their own level. We would like to benefit from data links made by domain-specific aggregators or individual data providers. The diversity of ingested resources forces indeed Europeana to focus its automatic enrichment efforts on large, multilingual and rather generic vocabularies. But being able to exploit specific, "local" semantic resources can already bring us a long way. Actually these resources are sometimes used by many partners and benefit from tremendous data curation efforts.

An obvious case is the Getty Art and Architecture Thesaurus (AAT)[14], which is used by dozens of museums represented in Europeana. Until now, AAT concepts only appeared as simple, undistinguished labels in object records sent to Europeana. Europeana itself had access to the AAT vocabulary data, but linking it to records required too much collection-specific work. The recent publication of the AAT as Linked Data[15] changes this. At the time of writing, we are working with a handful of data providers to include the new AAT URIs in their data. This will enable Europeana to fetch all the multilingual semantic data attached to

[13] Editor's note: the work on display will be finished in time for the Satellite meeting.
[14] http://www.getty.edu/research/tools/vocabularies/aat/
[15] http://blogs.getty.edu/iris/art-architecture-thesaurus-now-available-as-linked-open-data/

them, via the centralized open Getty service.[16] This is even easier as the AAT data uses the SKOS model that EDM expects for conceptual data.

The increased reliance of Europeana services on semantic enrichment illustrates the need for rich, open resources, giving more values to initiative like Getty's.

In The European Library

The European Library has brought a particular focus on the knowledge organisation systems it has access to, applying information extraction techniques to match to open data pivot sources.

The European Library has access to a large set of subject headings through the bibliographic records it aggregates. However a survey of the various types of subject systems used in the bibliographic data concludes that libraries using subject classifications use most of the time local, language-dependent vocabularies rather than language-independent classification systems such as Dewey Decimal Classification or the Universal Decimal Classification.

Aligning these language-dependent subject vocabularies is therefore required to provide better services. It is a challenge because of the diversity of types and languages of the knowledge organisation systems used across European libraries. An additional difficulty is introduced by the information loss resulting from the aggregation of data from various sources: subject data is often striped of its original richness (synonyms, hierarchical links).

The Europeana Library has begun this effort, focusing on the least language-dependent vocabularies. As a starting point, it bases its subject access functionality on the Common European Research Classification Scheme[17], a research-oriented vocabulary part of the Common European Research project Information Format, CERIF [8].

[16] See one of the first ingested objects at http://www.europeana.eu/portal/record/2026116/Partage_Plus_ProvidedCHO_Bildarchiv_Foto_Marburg_obj_20887058_fmd470404.html.

[17] http://www.eurocris.org/Index.php?page=CERIFreleases&t=1

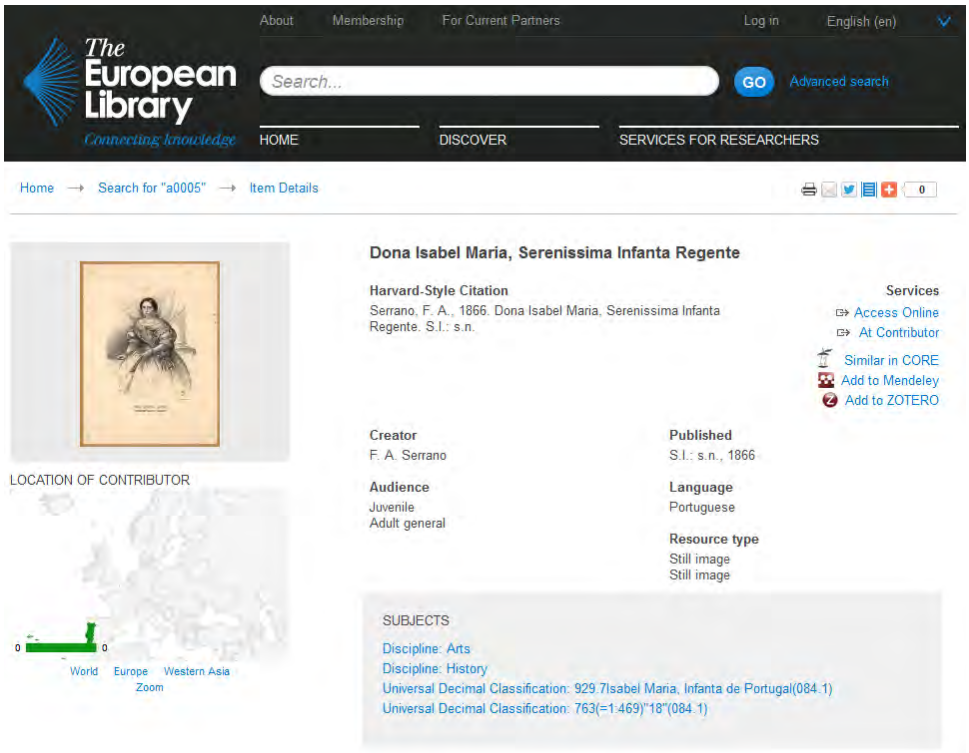

Fig 3: Alignment of bibliographical records with CERIF in The European Library portal

In addition, The European Library exploits the results of the project Multilingual Access to Subjects (MACS)[18] for subject browsing. The MACS project has produced manual and semi-automatic alignments between three major systems: the Library of Congress Subject Heading (LCSH) [19], the Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU)[20] and the Schlagwortnormdatei (SWD), covering English, French and German respectively. By the end of 2012, 120.000 cross-language links were established between the subject headings most often used in the collections of the French, British, German and Swiss national libraries.

A subject normalisation process has been applied to all bibliographic data in The European Library, including MARC data from libraries not participating in MACS, as well as Dublin Core based data, where subject data is often present but the used knowledge organisation system is not known. The outcome is a cross-language subject base, which is used by the portal to provide a comprehensive listing of the bibliographic resources that address a same topic.

Another example from The European Library is the consolidation of author information to link objects to Virtual International Authority File (VIAF)[21] [4]. The European Library matches the data about the contributors of a work present in the bibliographic records to the VIAF entries. Our first matching process relies on direct comparison of (numerical) authority record identifiers. Its success depends on (i) the involvement of the library in VIAF; (ii) the presence of authority record identifiers in the data. The second matching process relies on the

[18] http://www.nb.admin.ch/nb_professionnel/projektarbeit/00729/00733/index.html?lang=en
[19] http://id.loc.gov/authorities/subjects.html
[20] http://rameau.bnf.fr/
[21] http://viaf.org/

other data available in the original bibliographic records and the VIAF records: structured person names, variants of names in different languages, birth and death dates. In data originating, or concerning, countries where other alphabets than Latin are used (namely Greek and Cyrillic) the matching of names is applied in conjunction with standardized alphabet transliteration algorithms.

Based on this work, further enrichment of contributors' data has been performed against the International Standard Name Identifier (ISNI), which identifies uniquely and authoritatively public entities in various domains [7]. The European Library submitted data on work contributors to the ISNI International Agency. A total of 1,539,060 works were assigned ISNI identifiers, representing an improvement of the contributor data for 5,613,872 bibliographic records. The ISNI identifiers are currently being integrated in The European Library.

INTEGRATING LIBRARY-DOMAIN LINKED DATA IN THE WHOLE CULTURAL HERITAGE DOMAIN

The European Library is now working on integrating these results into its services and start disseminating them to other research infrastructures or Europeana. One of the major challenges being addressed, is their publication under a data model suitable for Linked Data [10], as well as an adequate licensing framework.

We have started the design of a new data model, which we will use to publish data about bibliographic records, digital objects and collections as Linked Open Data. This model will accommodate the different levels of semantic detail in the data aggregated by The European Library, from complex MARC data to simple Dublin Core data. The model similarly to EDM will be based on several vocabularies such as RDA, Dublin Core and SKOS.

Our Linked Open Data service also aims to make openly available the results of the data alignment and enrichment processing described earlier, including for instance:

- Links for FRBR Group 2 entities (person and corporate body), between the aggregated bibliographic data and VIAF.
- Links between concepts (FRBR Group 3) across the aggregated bibliographic data sources by exploiting the MACS results.
- Links between the aggregated bibliographic data and Geonames places (as FRBR Group 3 entities).

As an example of the issues ahead of us, let us focus once more on the results of the MACS project. They would be a great asset for many communities, since (multilingual) vocabulary building and alignment is a key aspect in many sectors[22]. The subject headings systems aligned in MACS (LCSH, RAMEAU and SWD) are available as linked open data. But the integration of the MACS links in each of the linked datasets is neither a concerted nor a continuous process, which can result in inconsistencies. Furthermore, their current representation in these linked datasets as simple SKOS links next to all other vocabulary-specific statements can hamper fine-tuned, provenance-sensitive data usage, e.g., by applications that would treat differently the links established by the vocabulary owners

[22] One of the authors has received requests from researchers for using the MACS alignments as a gold standard for testing automatic alignment tools, years after a first experiment had been set up for the Ontology Alignment Evaluation Initiative, http://oaei.ontologymatching.org/2009/library/.

themselves and the ones contributed by third parties. Giving MACS data its own space and a finer-grained representation, inspired for example from the Expressive and Declarative Ontology Alignment Language[23] or the PROV ontology[24] would enable:

- better dissemination and re-use of MACS by third parties (Europeana, research infrastructures)
- easier integration with new alignments between the existing MACS vocabularies or with new subjects headings systems
- better visibility of library subject headings system on the Web.

Making MACS available as linked open data will be achieved using the European Library platform already in place. Initially, it will also be made available as a downloadable dataset and in a second stage, a SPARQL endpoint will be set up for online querying.

The publication of data enrichment by The European Library will complement the ongoing activities of national libraries with linked open data and will help connect the library-domain data with the data from other sectors. In Europeana, for example, the enrichments from The European Library will co-exist with the ones mentioned in the previous section, including the links to the AAT linked open data. Later on, both Europeana and The European Library could integrate more cross-vocabulary alignments (similar to the mapping networks around AGROVOC[25] or the STW Thesaurus for Economics[26]) to further populate the "semantic layer" outlined earlier in this paper.

One concrete step to realize this data interconnection vision has been made for the Europeana 1914-1918 initiative[27]. The European Library provided metadata from 11 institutions across Europe on the First World War theme. To enable multilingual subject-based access, several Europeana-coined terms for that theme have been selected, based on a subset of Library of Congress subject headings, and augmented with translations in the providers' languages. The structured representation of these as SKOS concepts for ingestion as EDM data enables us to keep all the data added in the context of the project, while providing explicit links to the original concepts published at the Library of Congress. Future work involve the further integration of this SKOS data in the portal and its promotion as a reference for any data providers sending metadata about First World War to Europeana.

CONCLUSION

The task of creating linked data is demanding in terms of human and computational resources, and requires a large range of expertise in information science and semantic technology. We argue that library data aggregators can provide an organisational environment where conducting linked data activities becomes less demanding for libraries. Such organisations are indeed in position to leverage existing information and communication technologies as part of their operations; their expertise in both library data and the semantic web can bring many benefits as it is coupled with direct access to centralised data. This aspect is being addressed in cooperation between The European Library and Research Libraries UK[28]. Domain-specific aggregators are also in a position within the domain network to facilitate synergies between

[23] http://alignapi.gforge.inria.fr/edoal.html
[24] http://www.w3.org/TR/prov-o/
[25] http://aims.fao.org/node/16917/
[26] http://zbw.eu/stw/versions/latest/mapping/about.en.html
[27] http://www.europeana1914-1918.eu/en
[28] http://www.rluk.ac.uk

existing library initiatives, including the larger-scale aggregation networks such as Europeana, which are then in position to connect them to other cultural heritage domains.

References

[1] Europeana (2012). Definition of the Europeana Data Model elements. Retrieved from http://pro.europeana.eu/edm-documentation (March 07, 2014)

[2] EuropeanaConnect (2011). Europeana Semantic Data Layer. Retrieved from http://www.europeanaconnect.eu/results-and-resources.php?page=1 (March 7, 2014)

[3] Gradmann, S. (2010). Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana. Retrieved from http://pro.europeana.eu/c/document_library/get_file?uuid=cb417911-1ee0-473b-8840-bd7c6e9c93ae&groupId=10602 (March 07, 2014)

[4] Bennett, R.; Hengel-Dittrich, C.; O'Neill, E.; Tillett, B. (2006): VIAF Virtual International Authority File): Linking Die Deutsche Bibliothek and Library of Congress Name Authority Files. World Library and Information Congress: 72nd IFLA General Conference and Council.

[5] Freire, N., Scipione, G., Muhr, M., Juffinger A. (2013): Supporting Rights Clearance for Digitisation Projects with the ARROW Service. LIBER Quarterly, 22(4), pp. 265-284.

[6] Freire, N., Borbinha, J., Martins, B. (2008): Consolidation of References to Persons in Bibliographic Databases. Bibliographic Databases. ICADL 2008 – The 11th International Conference on Asian Digital Libraries.

[7] International Organization for Standardization (2012): Information and documentation. International standard name identifier, ISO 27729

[8]CERIF Introduction. Retrieved from http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1 (March 07, 2014)

[9] Final report of the EuropeanaTech Task Force on Multilingual and Semantic Enrichment Strategy, *forthcoming* http://pro.europeana.eu/web/network/europeana-tech/-/wiki/Main/Task+force+multilingual+semantic+enrichment

[10] Isaac, A., Waites, W., Young, W. & Zeng, M. (2011). Library Linked Data Incubator Group. Datasets, Value Vocabularies, and Metadata Element Sets. http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/

[11] Final Report of the EuropeanaTech EDM – FRBRoo Application Profile Task Force http://pro.europeana.eu/web/network/europeana-tech/-/wiki/Main/Task+Force+EDM+FRBRoo

[12] Final Report of the EuropeanaTech Task force on EDM mappings, refinements and extensions http://pro.europeana.eu/web/network/europeana-tech/-/wiki/Main/Task+force+on+EDM+mappings+refinements+and+extensions

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