A study of multilingual semantic data integration

Presentation

Douglas Tudhope
Hypermedia Research Group,
University of South Wales, UK
douglas.tudhope@southwales.ac.uk

Ceri Binding
Hypermedia Research Group,
University of South Wales, UK
ceri.binding@southwales.ac.uk

Keywords: knowledge organization systems; linked open data; query expansion; natural language processing; Getty Art & Architecture Thesaurus; CIDOC Conceptual Reference Model; archaeology;

Abstract

The availability of the various forms of open data today offers great opportunity for meta level research that draws on combinations of data previously considered only in isolation. There are also great challenges to be overcome; datasets may have different data models, may employ different terminology or languages, project data may only be represented by the final textual report. However, metadata and controlled vocabularies have the potential to help address many of these issues.

Previous work by the authors has explored semantic integration of English language archaeological datasets and reports (Binding et al., 2015; Tudhope et al., 2011). This presentation reflects on experience from a semantic integration exercise involving archaeological datasets and reports in different languages. Different forms of Knowledge Organization Systems (KOS) were key to the exercise. The Getty Art and Architecture Thesaurus (AAT) was used as the underlying value vocabulary and the CIDOC CRM ontology as the metadata element set (Isaac et al. 2011) for the semantic integration. Linked data expressions of the vocabularies formed part of an integration dataset (RDF) extracted from the source data, together with subject metadata automatically generated from the reports via Natural Language Processing (NLP) techniques.

The data was selected following a broad theme of wooden material, objects and samples dated via dendrochronological analysis. The investigation was conducted as an advanced data integration case study for the ARIADNE FP7 archaeological infrastructure project (ARIADNE 2017), with the datasets and reports provided by Dutch, English and Swedish ARIADNE project partners.

The presentation will outline the data cleansing, NLP and integration methods and present illustrative scenarios from the web application Demonstrator (2017). A template based tool was used for data conversion of extracts from the archaeological datasets and also the data resulting from NLP information extraction from the archaeological reports (STELETO 2016). Following the approach used in the ARIADNE Portal (2017), terms from different languages were intellectually mapped to concept identifiers from the Linked Open Data implementation of the Getty AAT (2018), in order to support cross search (via the AAT) over subject metadata in different languages. The user is shielded from some of the complexity of the metadata framework and the underlying SPARQL implementation by an interactive query builder. The search system exploits the AAT's hierarchical relationships and specialised associative relationships to provide a query expansion capability using SPARQL 1.1 property paths.

The case study shows that it is possible to semantically integrate information extracted from datasets and grey literature reports in different languages and provide KOS-based search. The presentation reflects on lessons learned, including the need to allow resources for extensive data cleansing. Although more work on the NLP extraction methods is needed for an operational capability, the study was able to generate CRM/AAT based RDF from English, Dutch and

https://doi.org/10.23106/dcmi.952139116

Swedish texts in the same format as that derived from the datasets, thus allowing cross search. A pattern based mapping methodology helped ensure the validity and consistency of the ontology mappings and the lower level implementation details. The Demonstrator also illustrates the possibility of domain application oriented user interfaces for searching RDF datastores. Automatically generated metadata from natural language does not have the same reliability as metadata automatically derived from datasets (after data cleansing); future work should express the provenance of the subject metadata extracted and also the method by which it was extracted. Details of the case study methods and results can be found in Binding et al. (2018).

Acknowledgements

This work was supported by the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193 (the ARIADNE project).

References

AAT. (2018). Getty Art & Architecture Thesaurus as Linked Open Data, Getty Vocabulary Program, Retrieved May 5, 2018, from http://vocab.getty.edu/

ARIADNE. (2017). ARIADNE Project. Retrieved May 5, 2018, from http://www.ariadne-infrastructure.eu

ARIADNE Portal (2017). Retrieved May 5, 2018, from http://portal.ariadne-infrastructure.eu/

Binding Ceri, Michael Charno, Stuart Jeffrey, Keith May and Douglas Tudhope. (2015). Template Based Semantic Integration: From Legacy Archaeological Datasets to Linked Data. International Journal on Semantic Web and Information Systems, 11(1), 1-29.

Binding Ceri, Douglas Tudhope and Andreas Vlachidis. (2018). A study of semantic integration across archaeological data and reports in different languages, Journal of Information Science, Retrieved Aug 22, 2018, from https://doi.org/10.1177/0165551518789874. (an open access 'author accepted version' is available at https://pure.southwales.ac.uk/files/2683350/Archaeology_integration_JISauthorversion2.docx).

Demonstrator. (2017). Demonstrator for dendrochronological data integration case study. Retrieved May 5, 2018, from http://ariadne-lod.isti.cnr.it/description.html

Isaac Antoine, William Waites, Jeff Young J and Marcia Zeng. Eds. (2011). Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets. W3C Incubator Group Report, Retrieved May 5, 2018, from http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/

STELETO. (2016). STELETO open source code, Retrieved May 5, 2018, from https://github.com/cbinding/steleto/

Tudhope Douglas, Keith May, Ceri Binding, Andreas Vlachidis. (2011). Connecting archaeological data and grey literature via semantic cross search. Internet Archaeology, 30, Retrieved May 5, 2018, from https://doi.org/10.11141/ia.30.5

https://doi.org/10.23106/dcmi.952139116