Reflections on KOS based data integration

Douglas Tudhope and Ceri Binding

Hypermedia Research Group,
University of South Wales, UK

douglas.tudhope@southwales.ac.uk, ceri.binding@southwales.ac.uk

Keywords: knowledge organization systems, semantic data integration, natural language processing, Getty Art & Architecture Thesaurus, CIDOC Conceptual Reference Model.

1 Introduction

This paper briefly reviews two contrasting case studies by the authors on semantic data integration within the archaeology field and reflects on some of the key issues encountered, relevant to the NKOS Workshop theme on strategies for alignment of metadata to KOS linked data. Both projects involved diverse datasets with different schema and which employed different terminology. Both projects combined datasets with information extracted from archaeological reports via natural language processing (NLP). In both cases, the semantic framework that afforded data integration was a combination of metadata element sets organised by an ontology with a relevant value vocabulary (eg thesauri and term lists). Ontologies and value vocabularies have been seen as complementary resources for this purpose [1, 2].

2 Data Integration Projects

The STAR project case study [3] combined different archaeological thesauri (eg the Thesaurus of Monument Types and the Archaeological Objects Thesaurus) and various glossaries on the vocabulary side, subsequently most available as linked data [4]. The ontology for the study was the CRM-EH extension [5] of the CIDOC CRM [6], which specializes the CRM with additional elements for archaeological purposes. English language archaeological datasets focusing mainly on the Roman period were selected, together with an extract from the Archaeology Data Service grey literature library of reports. The Demonstrator [detailed scenarios are given in 7] explored the integration of KOS in cross search over RDF data representations. Binding et al. [8] discuss the pattern-based data extraction methods, the production of linked data and provide examples.

The more recent case study (part of the ARIADNE FP7 archaeological infrastructure project [9]) combined the Getty Art and Architecture Thesaurus (AAT) with a set of high level classes from the CIDOC CRM. Again both KOS are available as linked open data [10, 11]. 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 ARIADNE, with the datasets and reports provided by Dutch, English and Swedish ARIADNE partners. A detailed description of the study, including a review of relevant literature and scenarios from the Demonstrator [12] is provided in Binding et al. [13].

In both projects, rule-based NLP pipelines were developed using the GATE platform [14] supported by relevant archaeological glossaries and thesauri. STAR focused on English language grey literature reports, while the ARIADNE study processed English, Dutch and Swedish language reports. Subject metadata was derived from named entity recognition of archaeological concepts, such as object, material, together with archaeological contexts and periods (STAR) and numeric date ranges (ARIADNE). The Dutch and Swedish vocabulary resources were mapped intellectually to the AAT providing a multilingual capability for the ARIADNE study. The sequential pipeline architecture starts with domain independent components, such as a tokeniser, sentence splitter and part of speech tagger, proceeding with domain specific rules applied to vocabulary matches in the text. This produces XML format output which is transformed to the RDF format employed for the data extraction in each study. Vlachidis et al. [15] give a detailed description of the NLP methods for the English language STAR study, together with an evaluation of the performance of various NLP pipeline options. A description of the NLP work for the ARIADNE study is provided in [13].

Both projects developed different query builder user interfaces to shield the user from some of the complexity of the metadata framework and from the underlying SPARQL (RDF) implementation. In the ARIADNE case study, the same template based tool [16] was used for data conversion of extracts from the archaeological datasets and also the data resulting from the NLP information extraction in the second project. Query expansion was provided in the demonstrator, based on the AAT's hierarchical relationships and specialised associative relationships.

3 Reflections

As with all projects, KOS-based development efforts involve design choices. Inevitably, with finite resources it is usually impractical to develop parallel implementations to compare major design alternatives and thus not easy to know the consequences of one design choice over another. We reflect on some major design decisions encountered during the two projects, with a view to informing future work.

These include:
- the level of application detail to model in the integration, how much of the source datasets and reports should be extracted, aligned to KOS and expressed as linked data. Should it be a subset or as much as possible?

- the appropriate balance of that application modeling detail, expressed between the ontology and the vocabulary side. How much to handle via the ontology and how much to handle via the thesaurus (or other vocabulary)? How much detail is it worthwhile to model?
- how to mitigate the possibility of creating alternative (valid) ontology mapping expressions of the same underlying semantics from different sources and thus make cross search and interoperability difficult?
- should the native schema of the source datasets be maintained in the resulting integration, or replaced (via the alignment) by the new semantic framework?
- both projects required substantial data cleansing. How should the resulting new information be modeled, what is the relationship with the source dataset, how should it be expressed?
- how to express the information extracted via NLP from texts in an RDF framework. How much certainty to associate with the derived data, what kinds of elements are represented (archaeological texts often refer to types of object or material rather than named specific individual items)?
- how to express results from search over both data and reports, how to express the provenance of the subject metadata extracted and also the method by which it was extracted?

The issues relating to the above points will be outlined and illustrated in the workshop presentation and opened for discussion.

Acknowledgements

This work was supported by the UK Arts and Humanities Research Council [grant numbers AH/D001528/1, AH/H037357/1] and the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193 (the ARIADNE project).

References

1. ISO 25964-2:2013. Information and documentation - Thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies. https://www.niso.org/schemas/iso25964#part2, last accessed 2018/08/26.
2. Isaac A, Waites W, Young J, Zeng M. Eds. Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets. W3C Incubator Group Report. (2011), http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/, last accessed 2018/08/26.
3. STAR Project. http://hypermedia.research.southwales.ac.uk/kos/star/, last accessed 2018/08/26.
4. Heritagedata. Linked Data Vocabularies. http://www.heritagedata.org/, last accessed 2018/08/26.
5. Cripps P, Greenhalgh A, Fellows D, May K, Robinson D. Ontological modelling of the work of the Centre for Archaeology. CIDOC CRM Technical Paper. (2004)

http://old.cidoc-crm.org/docs/Ontological_Modelling_Project_Report_ Sep2004.pdf, last accessed 2018/08/26.

6. Crofts N, Doerr M, Gill T, Stead S, Stiff M. Eds. Definition of the CIDOC Conceptual Reference Model, v5.0.4. (2011), http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.4.pdf, last accessed 2018/08/26.
7. Tudhope D, May K, Binding C, Vlachidis A. Connecting archaeological data and grey literature via semantic cross search. Internet Archaeology, 30, (2011), https://doi.org/10.11141/ia.30.5 (open access), last accessed 2018/08/26.
8. Binding C, Charno M, Jeffrey S, May K, Tudhope D. Template based semantic integration: From legacy archaeological datasets to linked data. International Journal on Semantic Web and Information Systems, 11(1), 1-29 (2015).
9. ARIADNE. ARIADNE Project. http://www.ariadne-infrastructure.eu, last accessed 2018/08/26.
10. CIDOC CRM linked data. https://github.com/erlangen-crm/ecrm, last accessed 2018/08/26.
11. AAT. Getty Vocabularies as Linked Open Data, Getty Vocabulary Program, http://vocab.getty.edu/, last accessed 2018/08/26.
12. Demonstrator. Demonstrator for dendrochronological data integration case study. http://ariadne-lod.isti.cnr.it/description.html, last accessed 2018/08/26.
13. Binding C, Tudhope D, Vlachidis A. A study of semantic integration across archaeological data and reports in different languages. Journal of Information Science, (2018), https://doi.org/10.1177/0165551518789874, last accessed 2018/08/26.. An open access 'author accepted version' is available at https://pure.southwales.ac.uk/files/2683350/Archaeology_integration_JISauthorversion2.docx).
14. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of 40th Annual Meeting of Association for Computational Linguistics, pp 168-175. New Brunswick (2002).
15. Vlachidis A, Tudhope D. A knowledge-based approach to information extraction for semantic interoperability in the archaeology domain. Journal of the Association for Information Science and Technology 67(5), 1138-1152 (2016).
16. STELETO. STELETO open source code, https://github.com/cbinding/steleto/, last accessed 2018/08/26.

** **Open Access** versions of Hypermedia Research Group’s KOS papers are available from https://bit.ly/2ocaHC6