Semantic Web Techniques for Multiple Views on Heterogeneous Collections: a Case Study

Marjolein van Gendt Antoine Isaac Lourens van der Meij
Stefan Schlobach

Vrije Universiteit Amsterdam
Koninklijke Bibliotheek, Den Haag
Max Planck Institute for Psycholinguistics, Nijmegen

Abstract

Integrated digital access to multiple collections is a prominent issue for many Cultural Heritage institutions. The metadata describing diverse collections must be interoperable, which requires aligning the controlled vocabularies that are used to annotate objects from these collections. In this paper, we present an experiment where we match the vocabularies of two collections by applying the Knowledge Representation techniques established in recent Semantic Web research. We discuss the steps that are required for such matching, namely formalising the initial resources using Semantic Web languages, and running ontology mapping tools on the resulting representations. In addition, we present a prototype that enables the user to browse the two collections using the obtained alignment while still providing her with the original vocabulary structures. This paper is an abbreviated version of a paper accepted at the 10th European Conference on Digital Libraries, ECDL 2006.

1 Introduction

Ontology alignment to facilitate data integration is a prominent issue within the Semantic Web (SW) community. Thesaurus mapping is important in the same fashion for collection integration within the Cultural Heritage domain. The general aim of our project (STITCH[1], funded by NWO, Netherlands Organisation for Scientific Research) is to determine to what extent the current SW techniques can solve heterogeneity issues in the CH sector. Here we report on how we took off-the-shelf, state-of-the-art ontology mappers and evaluated those in a case study concerning Cultural Heritage (CH) thesaurus mapping. In this case study we aligned two thesauri, Iconclass[2] and ARIA[3], to integrate two collections annotated using these vocabularies: the Illuminated Manuscripts collection from the Dutch National Library and the Masterpieces collection from the Rijksmuseum, respectively. Our approach for integrating these collections was first to formalise the vocabularies using SKOS, then to align the thesauri using Falcon [3] and S-Match [1], and finally to visualize the resulting mapping in a faceted browser. A full account of this work can be found in [4].

2 Formalisation

Our case study supplied two controlled vocabularies which needed transformation. For each vocabulary we first performed an analysis of its original structure and its idiomatic elements. This showed that CH thesauri can have very specific features – multiple labels, associative relationships – which could and should be exploited in mapping. Second we constructed a standard representation of the thesauri using SKOS, an RDF vocabulary that is currently being developed within the W3C Semantic Web activity. Because we opted for a standard formalisation, some idiomatic information

[1] http://stitch.cs.vu.nl/index.html
[2] http://www.iconclass.nl
[3] http://www.rijksmuseum.nl/aria/aria_catalogs/index?lang=en

from Iconclass was lost in this step, which is an obvious drawback of using generic methods as proposed by the SW community. Finally, we performed application-specific formalisation in order to let tools like reasoning engines and browsers interpret the semantics of the SKOS models.

3 Collection integration

The formalised versions of ARIA and Iconclass were fed into Falcon and S-Match. Falcon is one of the best performing[4] tools for aligning complex RDFS/OWL ontologies. S-Match has been developed for mapping tree-like structures, and required us to rewrite the SKOS files into a non-standard input format. Mapping thesauri proved to be difficult for both mappers, and the overall results were less than satisfactory. S-Match showed 46% correct mappings for a selected subset of Iconclass (1500 concepts) and the complete ARIA thesaurus (500 concepts). Falcon reached a precision of only 16% and was not scalable regarding complete Iconclass (28.000 concepts, that is matrices of that order of dimension as starting points for Falcon in-memory computation). The main cause for these results is that CH controlled vocabularies have features that make them really different from ontologies. The most prominent differences in our experiment were that some CH thesauri have glosses for describing concepts instead of simple terms and that thesauri structure and semantics are loosely defined compared to full-fledged, formal ontologies.

4 Collection visualization

We implemented a multi-faceted browsing framework to evaluate and explore the results of our mapping effort. This user interface is operational[5] and provides a way to browse the original CH thesauri and retrieve images from both collections. The design of our browser was inspired by the Flamenco search interface framework [2]. Its implementation uses SWI-Prolog and the Sesame RDF repository[6] for storage and querying.

5 Conclusion

The main goal of our research was to evaluate to what extent SW techniques can solve heterogeneity issues when integrating multiple CH collections. The general conclusion is positive: in a relatively short time we managed to implement an integrated browsing environment that was built purely on accepted standards for representing data, and which used existing tools for storage, querying and mapping.

In our research we did notice some problems, though. The most important issues are the balance between standardization and information loss and the need for CH thesauri specific adaptation of ontology mapping techniques: it is for example necessary to have alignment tools properly exploiting rich labeling information while still remaining independent from the case of specific thesauri.

References

[1] Giunchiglia, F., Shvaiko, P., and Yatskevich, M.: Semantic Schema Matching. 13th International Conference on Cooperative Information Systems (CoopIS 2005), Cyprus, 2005.

[2] Hearst, M., English, J., Sinha, R., Swearingen, K. and Yee, P.: Finding the Flow in Web Site Search. Communications of the ACM, 45 (9), 2002.

[3] Jian, N., Hu, W., Cheng, G., and Qu, Y.: Falcon-AO: Aligning Ontologies with Falcon. K-CAP Workshop on Integrating Ontologies, Banff, Canada, 2005.

[4] van Gendt, M., Isaac, A., van der Meij, L., and Schlobach S.: Semantic Web Techniques for Multiple Views on Heterogeneous Collections: a Case Study. In 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), Alicante, Spain, 2006.

[4] In the OAEI - Ontology Alignment Evaluation Initiative 2005

[5] See `http://stitch.cs.vu.nl/demo.html`

[6] Available at `http://www.openrdf.org`