Faceted Access to Heterogeneous Cultural Heritage Collections using Semantic Web Techniques

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 in these collections. We demonstrate an interface prototype presenting two collections whose vocabularies have been matched applying Knowledge Representation techniques as established in recent Semantic Web research. This prototype features three different views that enable the user to browse the two collections using the alignment, while still providing her with the original vocabulary structures.

1 STITCH general goals and pilot experimentation

STITCH – SemanTic Interoperability To access Cultural Heritage – is a research project within CATCH, a programme funded by the Netherlands Organisation for Scientific Research – NWO. The objective of the project is to evaluate methods and tools from the Semantic Web research area for integrating Cultural Heritage collections.

Our first experiment and implementation aimed at providing integrated access to two heterogeneous collections, the Illuminated Manuscript collection[1] from the Dutch National Library in The Hague, and the ARIA Masterpieces collection[2] from the Rijksmuseum in Amsterdam.

First a conversion to generic Semantic Web formats, such as RDF(S)[3] and SKOS[4], was required. Second, having computer-readable representations, we could align them. We turned to two off-the-shelf ontology mappers (S-Match[1] and Falcon[3]) and evaluated their use for aligning CH thesauri. Third, automatically found correspondences were used in an interface we implemented to browse different vocabularies and to retrieve documents from several collections.

In this demonstration, we will recall the first steps[5], but the emphasis will be on the browsing features offered by our interface.

2 Three faceted views on integrated collections

We implemented a multi-faceted browsing (MFB) interface to evaluate and explore the results of our mapping effort. MFB involves constraining search criteria along – usually orthogonal – aspects of a collection called *Facets*. Here we adapted this paradigm in an atypical way, since we used one category (subject) for defining several facets.

[1] `http://www.kb.nl/kb/manuscripts/`
[2] `http://www.rijksmuseum.nl/collectie/index.jsp?lang=en`
[3] `http://www.w3.org/RDF/`
[4] `http://www.w3.org/2004/02/skos/`
[5] More details on these experiments and the lessons we learned from them can be found in [4].

For searching through the integrated collections we explored three different views on integrated collections: *single*, *combined*, and *merged*. All views take into account the links between the vocabularies established in the automatic alignment process.
The *Single View* presents the integrated collections from the perspective of only one of the collections. Elements of the other collection are found through the links between their subject annotations and the concepts of the current view.
The *Combined View* provides simultaneous access to the collections through their respective vocabularies. This allows us to browse through the integrated collections as if it was a single collection indexed against two vocabularies.
The *Merged View* provides access to the collections through a merged thesaurus combining both original vocabularies into a single one.

3 Technical details

The design of our browser was inspired by the Flamenco search interface framework [2]. It is implemented as a web server written in SWI-Prolog, using its HTTP server and HTML code generation libraries. All data is stored in an external RDF Sesame repository[6] which is accessed using SeRQL queries. The server software runs on Linux and MS Windows operating systems.

The focus for the prototype was on easy configurability and adaptability for experimentation using Semantic Web Techniques. In order to easily adapt configurations to new ideas concerning data modeling, all collection, vocabulary and even website configuration information is stored in the RDF repository. As basic representation blocks, we have instances of a `Facet` class, that we introduced to define all aspects of a single facet, such as the SKOS `ConceptScheme` it uses as structured vocabulary. Website configuration objects – instances of the class `SiteConfiguration` – then group together `Facets` into separate websites. New configurations can thus be specified very easily, which enabled us to define sites that use different sets of alignments between the two vocabularies. This also allows us to seamlessly plug in new collections, vocabularies and alignments. Our prototype thus is a proof-of-concept of the way Cultural Heritage institutes from various contexts can benefit from integrating their collections on the basis of recently established Artificial Intelligence techniques.

The browser and the underlying collections can be found at `http://stitch.cs.vu.nl/demo`. The presentation, a brief explanation of the collection background and the formalisation and alignment process followed by the demonstration of the prototype, will not exceed 25 minutes.

References

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

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