STW Thesaurus for Economics Web Service applied to library applications

Especially in the realm of digital libraries we normally expect positive effects from the use of controlled vocabularies: With regard to indexing, we expect a higher level of precision by categorizing literature (or other resources of scientific purpose) by means of subject-specific concepts. With regard to retrieving, we expect a better matching by suggesting those concepts which are related both to search input and indexed terms, so that the work of librarians (or other scientific staff responsible for professional indexing including researchers themselves) and the demand of users will converge.

In this paper we outline the technical infrastructure necessary to support and to optimize this scenario. This infrastructure consists of

- a SKOS-representation of the STW Thesaurus for Economics (recently published as Linked Data),
- a web service (http://zbw.eu/beta/stw-ws) for requesting and delivering its concepts and their relations, offering a simple, REST-oriented API for resources like “concepts”, “narrower” or “synonyms”,
- a thin PHP wrapper which translates the resource requests to SPARQL queries, hiding their complexity from the application programmer, and
- a Joseki server operating directly on a memory model of the RDF/XML file and alternatively a SPARQLite server with a Jena TDB triple store and LARQ (Lucene free text indexing for SPARQL) to execute the queries.

We exploit it in the two fields introduced above:

In the field of retrieval it can

1) support (even cross-lingual) retrieval by an optional query expansion with synonyms (in environments where only full text retrieval is available), and
2) support retrieval in holdings indexed using a controlled vocabulary by providing the user with links to searches for narrower, related etc. terms

STW with its 5,800 bilingual concepts, 17,000 synonyms and 25,000 relations is well suited for this.

For a large scale evaluation of the thesaurus coverage and for a simulation of the effects of query enrichment in association with “real life” queries we examined 1.15 million “simple” queries mined from the logs of a representative bibliographic database in economics (http://econis.eu). As we found out, 50% of the (already normalized) queries only occurred once in two years (only less than 1% showed up 10 or more times). In case of queries with multiple occurrences we looked up the thesaurus and found 14% direct matches. The main reason for this rather low rate was that users combined multiple concepts within one query. We therefore will present different approaches to deal with this, especially a) syntactical splitting of search strings, and b) an application of full text matching techniques to the lookup itself. For the identified thesaurus concepts the synonyms, broader, narrower and related terms are evaluated statistically.

In the field of indexing, the thesaurus web service can assist users with indexing or “tagging”. In our DSpace repository for scientific articles in economics, indexing is performed by authorized end users submitting and uploading their articles. These users, e.g. scientific staff from an economics faculty, generally have no experience in applying controlled vocabularies. To support them we will offer an autosuggest service for thesaurus terms very similar to the one already in use on the STW web site (http://zbw.eu/stw). Finally, we will stretch this service to another well-known terminology, the “Journal of Economic Literature” (JEL) classification.