

SKOS Thesaurus Management based on Linked Data with Poolparty

Andreas Koller

(punkt. netServices GmbH, Vienna, Austria
koller@punkt.at)

Abstract: The process of building and maintaining thesauri, especially in corporate settings, quite often cannot be justified due to its unfavourable cost-benefit ratio. Combining different approaches to build thesauri like text-mining, machine learning, expert driven thesaurus modelling and linking open data seems to be a promising methodology to overcome this obstacle. This paper gives an overview over a thesaurus building methodology based on Linked Data which was implemented in PoolParty which is a Thesaurus Management Tool (TMT) for the Semantic Web.

Keywords: Semantic Web, Linked Data, Thesaurus, SKOS, RDF

Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1

1 Introduction

Thesauri have been a building block for professional information management in different settings for years. Although advantages are obvious one can derive from thesaurus based content management, tagging systems or search engines, such applications built around professionally managed thesauri are still quite rare. At least four reasons for this phenomenon are frequently enumerated: (1) TMTs are too difficult to learn and to use, (2) TMTs are quite often inflexible in terms of a lack of integration scenarios into existing enterprise information systems, (3) TMTs often support either automatic or manual methods to maintain a thesaurus but scarcely as a combination of those two approaches and (4) companies are rarely knowledgeable about thesaurus building methodologies and/or interesting use cases which can be built around semantic knowledge models like SKOS thesauri.

In this paper we will introduce a methodology and a software tool which helps to manage at least three of the four problems just mentioned above.

2 Introducing PoolParty

PoolParty is a thesaurus management system for the semantic web including text analysis functionalities. The system helps to build and maintain multilingual thesauri providing a simple user interface and a couple of semantic services. PoolParty makes use of RDF/SKOS specification (Simple Knowledge Organisation System)¹

¹ RDF/SKOS specification: <http://www.w3.org/2004/02/skos/>

developed by the World Wide Web Consortium (W3C). PoolParty is a product developed by punkt.netServices GmbH, Vienna/Austria.

On top of the PoolParty service-interface a variety of semantic web applications can be realized, for example:

- semantic search engines
- recommender systems (similarity search)
- corporate bookmarking
- annotation- & tag recommender systems
- auto-complete and faceted browsing.

PoolParty's background functionality is built on top of the SAIL API² which makes it a flexible tool in terms of performance and scalability and the used Triple Store.

Support for viewing, creating and editing SKOS Data is offered via an AJAX-Frontend based on Yahoo User Interface³ (YUI). PoolParty is available as a compound Web Application. Much of its functionalities are also available as Web Services - be it REST or SOAP. PoolParty's Key-Phrase Extraction capabilities come from a modified version of the KEA 5 API, which is extended for the use of controlled vocabularies stored in a SAIL Repository (this module is available under GNU GPL).

3 Building and maintaining thesauri based on Linked Data

The system is fully compliant with W3C SKOS therefore it is able to import existing SKOS thesauri in different formats including RDF/XML, N-Triples or Turtle. PoolParty also provides a lot of functionalities to edit thesauri manually using an easy-to-handle interface: For example, two concepts can easily be merged via drag & drop or all forms of the user-interface are enhanced by auto-complete. In addition to manual editing PoolParty also extracts key-phrases from given documents automatically. Those so called "free concepts" can be inserted into the existing thesaurus and be converted into "approved concepts".

In addition to this semi-automatic approach the tool can also be used as a Linked Data client - making use of data from sources like DBpedia [Auer 08]. One example: The thesaurus manager wants to expand a concept with the preferred label (skos:prefLabel) "Niki Lauda". Therefore a corresponding resource (or category) from DBpedia will be identified, e.g. with the URI http://dbpedia.org/page/Niki_Lauda. This functionality can be realised on top of services like DBpedia URI Lookup⁴ or Sindice [Oren 08]. After the thesaurus manager has confirmed that the proposed resource is identical with the concept represented in the thesaurus some information like the abstract (dbpprop:abstract) can be retrieved and automatically inserted as a definition (skos:definition) of the concept. Moreover also geographical coordinates or narrower

² SAIL API: <http://www.openrdf.org/doc/sesame2/system/ch05.html>

³ Yahoo User Interface: <http://developer.yahoo.com/yui/>

⁴ Dbpedia URI Lookup: <http://lookup.dbpedia.org/>

or sibling concepts can be retrieved, e.g. as instances (skos:subject) of a DBpedia Category, in this case as a Ferrari Formula One driver (http://dbpedia.org/page/Category:Ferrari_Formula_One_drivers).

Iterating between manual editing of a thesaurus, adding concepts by automatic key-phrase extraction (and converting phrases into approved concepts) and the usage of linked data sources to expand an existing thesaurus seems to be a promising approach to build, augment and maintain thesauri in a cost-effective way.

To make thesauri from PoolParty available also as Linked data sources, the system offers SPARQL-Endpoints and Pubby⁵ Linked Data interfaces for all thesaurus projects. Since all concepts can be linked to resources from DBpedia the system also follows the Linked Data Principles⁶ as formulated by Tim Berners-Lee.

4 Conclusio and future prospects

Problems(1) - (3) as mentioned above can be solved partially since PoolParty offers an easy-to-use tool to build thesauri based on a highly interoperable framework called "Semantic Web". It makes use not only of text extraction algorithms but also of linked data mechanisms which helps to maintain and expand existing thesauri in a very effective way.

Especially in corporate settings the quality of Linked Data sources becomes a critical factor since they will be integrated into a corporate thesaurus. This will be one of the big challenges as Chris Bizer⁷ stated: "I think we will see a growing number of applications that use data from the public Web as background knowledge to offer better search capabilities and to augment local content with additional content from the Web of Data. Beside of the classic search engines, there might also be market opportunities for new search engines that specialize on Linked Data. This will allow them to sell access to cleaned views on the Data Web and to become central components within Linked Data applications. Within the corporate market, there is interest in using Linked Data as a lightweight, pay-as-you-go data integration technology."

References

[Auer, 08] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives: DBpedia: A Nucleus for a Web of Open Data. In: Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007), page 722--735. November 2008

[Oren, 08] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, G. Tummarello: Sindice.com: A Document-oriented Lookup Index for Open Linked Data. In: International Journal of Metadata, Semantics and Ontologies(2008)

⁵ Pubby: <http://www4.wiwiw.fu-berlin.de/pubby/>

⁶ Linked Data Principles: <http://www.w3.org/DesignIssues/LinkedData.html>

⁷ Interview with Chris Bizer: <http://blog.semantic-web.at/2009/04/17/chris-bizer-talks-about-the-commercial-opportunities-of-linked-data/>