PoolParty: SKOS Thesaurus Management utilizing Linked Data

Thomas Schandl, Andreas Blumauer

punkt. NetServices GmbH,
Lerchenfelder Gürtel 43, 1160 Vienna, Austria
schandl@punkt.at, blumauer@punkt.at

Abstract. Building and maintaining thesauri are complex and laborious tasks. PoolParty is a Thesaurus Management Tool (TMT) for the Semantic Web, which aims to support the creation and maintenance of thesauri by utilizing Linked Open Data (LOD), text-analysis and easy-to-use GUIs, so thesauri can be managed and utilized by domain experts without needing knowledge about the semantic web. Some aspects of thesaurus management, like the editing of labels, can be done via a wiki-style interface, allowing for lowest possible access barriers to contribution. PoolParty can analyse documents in order to glean new concepts for a thesaurus. Additionally a thesaurus can be enriched by retrieving relevant information from Linked Data sources and thesauri can be imported and updated via LOD URIs from external systems and also can be published as new linked data sources on the semantic web.

Keywords: Semantic Web, Linking Open Data, Thesaurus, SKOS, RDF, Wiki.

1 Introduction

Thesauri have been an important tool in Information Retrieval for decades and still are [1]. While they have the potential to greatly improve the information management of large organisations, professionally managed thesauri are rarely used in content management systems, search engines or tagging systems.

Important reasons frequently given for this are: (1) the difficulty of learning and using TMT, (2) the lacking possibilities to integrate TMTs into existing enterprise information systems, (3) it's laborious to create and maintain a thesaurus, and while TMTs often support either automatic or manual methods to maintain a thesaurus they rarely combine those two approaches, and (4) companies don't have enough knowledge about thesaurus building methodologies and/or valuable use cases utilizing semantic knowledge models like SKOS thesauri.

The TMT PoolParty[1] addresses the first three issues. The demo will show PoolParty's thesaurus management features including document analysis, its Linked Data capabilities and its Wiki interface.

[1] http://poolparty.punkt.at/PoolParty/ - A screencast is available at http://bit.ly/6OqhYZ

2 Use Cases

PoolParty is a tool to create and maintain multilingual SKOS (Simple Knowledge Organisation System)[2] thesauri, aiming to be easy to use for people without a Semantic Web background or special technical skills. Utilizing semantic web technologies like RDF and especially SKOS allow thesauri to be represented in a standardised manner [2]. While OWL would offer greater possibilities in creating knowledge models, it is deemed too complex for the average information worker.

PoolParty was conceived to facilitate various commercial applications for thesauri. In order to achieve this, it needs to publish them and offer methods of integrating them with various applications [3]. In PoolParty this can be realized on top of its RESTful web service interface providing thesaurus management, indexing, search, tagging and linguistic analysis services.

Some of these (semantic) web applications are:

- Semantic search engines
- Recommender systems (similarity search)
- Corporate bookmarking
- Annotation- & tag recommender systems
- Autocomplete services and facetted browsing.

These use cases can be either achieved by using PoolParty stand-alone or by integrating it with existing Enterprise Search Engines and Document Management Systems.

3 Technologies

PoolParty is written in Java and uses the SAIL API[3], whereby it can be utilized with various triple stores, which allows for flexibility in terms of performance and scalability.

Thesaurus management itself (viewing, creating and editing SKOS concepts and their relationships) can be done in an AJAX Frontend based on Yahoo User Interface (YUI). Editing of labels can alternatively be done in a Wiki style HTML frontend.

For key-phrase extraction from documents PoolParty uses a modified version of the KEA[4] 5 API, which is extended for the use of controlled vocabularies stored in a SAIL Repository (this module is available under GNU GPL). The analysed documents are locally stored and indexed in Lucene[5] along with extracted concepts and related concepts.

---

2 http://www.w3.org/2004/02/skos
3 http://www.openrdf.org/doc/sesame2/system/ch05.html
4 http://www.nzdl.org/Kea/index.html
5 http://lucene.apache.org/

4 Thesaurus Management with PoolParty

The main thesaurus management GUI of PoolParty (see Fig. 1) is entirely web-based and utilizes AJAX to e.g. enable the quick merging of two concepts either via drag & drop or autocompletion of concept labels by the user. An overview over the thesaurus can be gained with a tree or a graph view of the concepts.

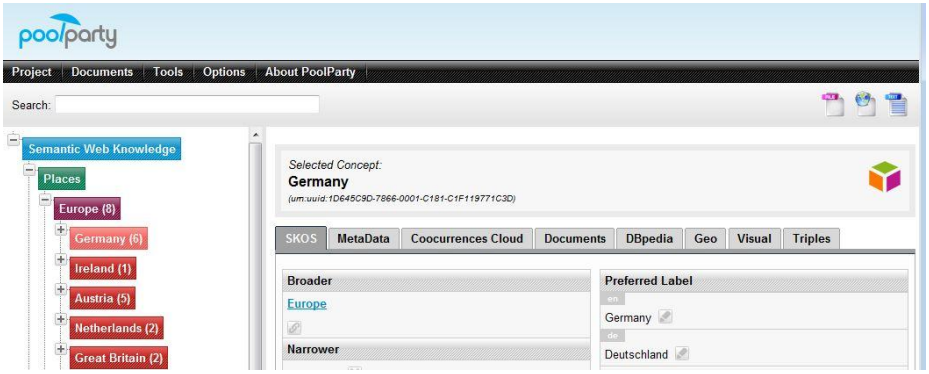

Fig. 1 PoolParty's main GUI with concept tree and SKOS view of selected concept

Consistent with PoolParty's goal of relieving the user of burdensome tasks while managing thesauri doesn't end with a comfortable user interface: PoolParty helps to semi-automatically expand a thesaurus as the user can use it to analyse documents (e.g. web pages or PDF files) relevant to her domain in order to glean candidate terms for her thesaurus. This is done by a key-phrase extractor which is based on KEA. The extractor not only detects concepts within a text which are already part of the thesaurus but also new canditate terms ("free concepts"). These can be approved by a thesaurus manager and can be integrated into the thesaurus, turning them into "approved concepts".

Documents can be searched in various ways – either by keyword search in the full text, by searching for their tags or by semantic search. The latter takes not only a concept's preferred label into account, but also its synonyms and the labels of its related concepts are considered in the search. The user might manually remove query terms used in semantic search. Boost values for the various relations considered in semantic search may also be adjusted. In the same way the recommendation mechanism for document similarity calculation works.

PoolParty by default also publishes an HTML Wiki version of its thesauri, which provides an alternative way to browse and edit concepts. Through this feature anyone can get read access to a thesaurus, and optionally also edit, add or delete labels of concepts. Search and autocomplete functions are available here as well.

The Wiki's HTML source is also enriched with RDFa, thereby exposing all RDF metadata associated with a concept as linked data which can be picked up by RDF search engines and crawlers.

PoolParty supports the import of thesauri in SKOS (in serializations including RDF/XML, N-Triples or Turtle) or Zthes format.

5 Linked Open Data Capabilities

PoolParty not only publishes its thesauri as Linked Open Data (additionally to a SPARQL endpoint)[6], but it also consumes LOD in order to expand thesauri with information from LOD sources. Concepts in the thesaurus can be linked to e.g. DBpedia[7] via the DBpedia lookup service [4], which takes the label of a concept and returns possible matching candidates. The user can select the DBpedia resource that matches the concept from his thesaurus, thereby creating a owl:sameAs relation between the concept URI in PoolParty and the DBpedia URI. The same approach can be used to link to other SKOS thesauri available as Linked Data.

Other triples can also be retrieved from the target data source, e.g. the DBpedia abstract can become a skos:definition and geographical coordinates can be imported and be used to display the location of a concept on the map, where appropriate. The DBpedia category information may also be used to retrieve additional concepts of that category as siblings of the concept in focus, in order to populate the thesaurus.

PoolParty is not only capable of importing a SKOS thesaurus from a Linked Data server, it may also receive updates to thesauri imported this way. This feature has been implemented in the course of the KiWi[8] project funded by the European Commission. KiWi also contains SKOS thesauri and exposes them as LOD. Both systems can read a thesaurus via the other's LOD interfaces and may write it to their own store. This is facilitated by special Linked Data URIs that return e.g. all the top-concepts of a thesaurus, with pointers to the URIs of their narrower concepts, which allow other systems to retrieve a complete thesaurus through iterative dereferencing of concept URIs.

Additionally KiWi and PoolParty publish lists of concepts created, modified, merged or deleted within user specified time-frames. With this information the systems can learn about updates to one of their thesauri in an external system. They then can compare the versions of concepts in both stores and may write according updates to their own store.

Data transfer and communication are achieved using REST/HTTP, no other protocols or middleware are necessary. Also no rights management for each external systems is needed, which otherwise would have to be configured separately for each source.

6 System Demo

In the demonstration session visitors will learn how to manage a SKOS thesaurus and how PoolParty supports the user in the process of creating, editing, relating and merging of SKOS concepts using the web GUI, autocomplete and drag and drop. We will explore different views of concepts (tree, graph, triples and location on a map).

---

[6] Example thesaurus published as Wiki with embedded RDFa to expose linked data: http://bit.ly/aM7LSL

[7] http://dbpedia.org/

[8] http://kiwi-project.eu/

We'll take a tour of the Wiki interface and learn how to use it to edit labels and take a look at the RDFa output exposed in the Wiki.

The document analysis features will be presented, showing how new concepts can be gleaned from text and integrated into a thesaurus. The visitor will learn how to conduct a semantic search function as well as how the similarity recommendations for indexed documents tagged with concepts work.

It will be shown how to interlink local concepts from DBpedia, thereby enhancing one's thesaurus with triples from the LOD cloud. Finally the data synchronisation via LOD will be shown by way of example interactions between the semantic framework KiWi and PoolParty.

7 Future Work

In the course of project LASSO funded by Austria's FFG[9] we will research improved methods of interlinking local thesauri with relevant entities from the LOD cloud. This will enable PoolParty to support thesaurus managers e.g. by semi-automatically populating a thesaurus by looking for related terms on external sites. Tighter interlinking with the LOD cloud can also enable PoolParty to suggest how new concepts could be classified, i.e. recommend possible parent concepts from a thesaurus for a concept in focus.

The synchronisation process via Linked Data will be improved in the ongoing KiWi project. We will implement an update and conflict resolution dialogue through which a user may decide which updates to concepts to accept and to consequently write to the system's store.

Most importantly we will work on integrating PoolParty with existing Enterprise Search Engines, Enterprise Wikis and Content Management Systems.

References

1. Aitchison, J., Gilchrist, A., Bawden, D.: Thesaurus Construction and Use: A Practical Manual. 4th edn. Europa Publications (2000)
2. Pastor-Sanchez, J. P., Martínez Mendez, F., and Rodríguez-Muñoz, J. V.: Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. informationresarch Vol 14 No. 4, Dec 2009. http://informationr.net/ir/14-4/paper422.html
3. Viljanen, K., Tuominen, J., Hyvönen, E.: Publishing and using ontologies as mashup services. In: Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW 2008), 5th European Semantic Web Conference 2008 (ESWC 2008), Tenerife, Spain (June 1-5 2008)
4. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web. Volume 7, Issue 3, September 2009, Pages 154-165

[9] Austrian Research Promotion Agency - http://ffg.at/