Utilizing, creating and publishing Linked Open Data with the Thesaurus Management Tool PoolParty

Thomas Schandl, Andreas Blumauer

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

Abstract. We introduce the Thesaurus Management Tool (TMT) PoolParty based on Semantic Web standards that reduces the effort to create and maintain thesauri by utilizing Linked Open Data (LOD), text-analysis and easy-to-use GUIs. PoolParty's aim is to lower the access barriers to managing thesauri, so domain experts can contribute to thesaurus creation without needing knowledge about the Semantic Web. A central feature of PoolParty is the enriching of a thesaurus with relevant information from LOD sources. It is also possible to import and update thesauri from remote LOD sources. Going a step further we present a Personal Information Management tool built on top of PoolParty which relies on Open Data to assist the user in the creation of thesauri by suggesting categories and individuals retrieved from LOD sources. Additionally PoolParty has natural language processing capabilities enabling it to analyse documents in order to glean new concepts for a thesaurus and several GUIs for managing thesauri varying in their complexity. Thesauri created with PoolParty can be published as Open Knowledge according to LOD best practices.

Keywords: **S**emantic Web, Linking Open Data, Thesaurus, Personal Information Management, SKOS, RDF.

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 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 information systems, (3) it is 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) users don't have enough knowledge about thesaurus building methodologies and/or worthwhile use cases utilizing semantic knowledge models like SKOS thesauri.

The TMT PoolParty[1] addresses the first three issues. A central goal is to ease the process of creating and maintaining thesauri by domain experts, that don't have a strong technical background, don't know about semantic technologies and maybe know little about thesauri. We see an important role for Linked Open Data in this area and equipped PoolParty with the capability to enrich one's own knowledge model with relevant information from the LOD cloud. In combination with several GUIs suited for varying levels of complexity PoolParty allows for low access barriers for creating and utilizing thesauri and Open Data.

PoolParty is a commercial application, but will have a version that can be used free of charge. We are working on such a version that makes use of the Talis platform[2]. In any version the user will have the option to publish thesauri as LOD and license them under various Creative Commons licenses.

2 Use Cases

PoolParty is based on Semantic Web technologies like RDF[3] and SKOS[4] (Simple Knowledge Organisation System) allowing for multilingual thesauri to be represented in a standardised manner [2]. While OWL[5] 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 and non-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.
- Personal Information Management

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.

[1] http://poolparty.punkt.at/
[2] http://www.talis.com/platform/
[3] http://www.w3.org/RDF/
[4] http://www.w3.org/2004/02/skos
[5] http://www.w3.org/TR/owl-ref/

3 Technologies

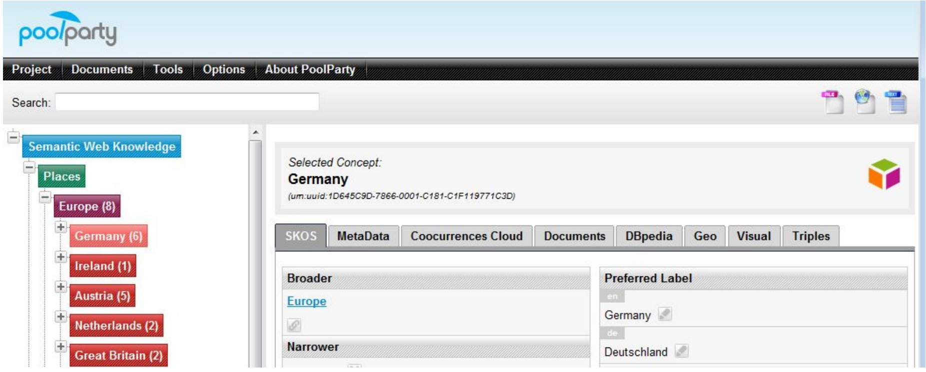
PoolParty is written in Java and uses the SAIL API[6], whereby it can be utilized with various triple stores, which allows for flexibility in terms of performance and scalability.

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[7] 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[8] along with extracted concepts and related concepts.

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.



Semantic Web Knowledge
Places
Europe (8)
Germany (6)
Ireland (1)
Austria (5)
Netherlands (2)
Great Britain (2)

Selected Concept:
Germany
(urn:uuid:1D645C9D-7866-0001-C181-C1F119771C3D)

Broader: Europe

Preferred Label:
en Germany
de Deutschland

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 the key-phrase extractor of KEA. The extracted

[6] http://www.openrdf.org/doc/sesame2/system/ch05.html
[7] http://www.nzdl.org/Kea/index.html
[8] http://lucene.apache.org/

terms can be approved by the user, thereby becoming "free concepts" which later 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 to be picked up the 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.

6 Linked Open Data Capabilities

PoolParty not only publishes its thesauri as Linked Open Data (additionally to a SPARQL endpoint), 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[9] 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 an 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 the 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[10] 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.

[9] http://dbpedia.org/

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

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.

This means each system decides autonomously which data it accepts and there is no risk of a system pushing data that might lead to inconsistencies into an external 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.

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.

7 Personal Information Management utilizing Linked Open Data

An example application that we are currently developing on top of PoolParty web services is a Personal Information Manager (PIM) utilizing Open Data.

Our goal is to enable users to create and utilize thesauri without requiring any knowledge about thesauri. We aim to hide the complexity of thesauri and their poly-hierarchical structure and concentrate on presenting the user with listboxes filled with terms from a thesaurus or LOD sources.

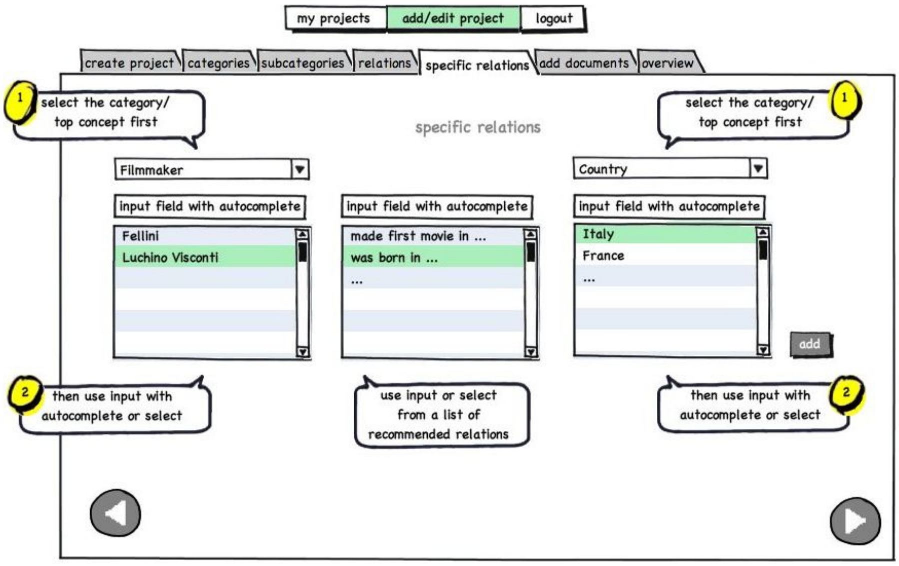
The PIM will be a web based application that makes use of Linked Open Data in order to assist the users with suggestions when they e.g. create categories. A movie expert for example might want to create a knowledge model of filmmakers and the countries they lived in. Upon the creation of a new project, the PIM asks the user to specify a general domain it is about (people, places, things, events, organisations, etc.). After the user selects "people", the system can use e.g. data about categories from YAGO[11], UMBEL[12] or DBpedia that relate to "people" to help refine the user's domain. The system might suggest popular categories or the user can start entering a string like "fil" prompting the system's autocomplete functionality to suggest the DBpedia class "filmmaker". When the user confirms that this is one of the topics his project is about and finishes the same process for the other topics (i.e. countries), several links between the local model and the LOD cloud exist and more specific information can be retrieved to assist the user's work on the thesaurus. The system might suggest possible relations between painters and cities like "made movie in", "was born in" or "lived in", and the user can specify which particular relations are of interest to him.

In a similar way the PIM will assist with creating instance data and suggest particular filmmakers and countries (see mock-up in Fig. 2) from sources like DBpedia that belong to the corresponding classes. In this way the PIM not only helps

[11] http://www.mpi-inf.mpg.de/yago-naga/yago/

[12] http://www.umbel.org/

with rapidly filling the model, but it automatically interlinks it with the LOD cloud in one go.


my projects
add/edit project
logout
create project
categories
subcategories
relations
specific relations
add documents
overview
1 select the category/ top concept first
specific relations
select the category/ top concept first 1
Filmmaker
Country
input field with autocomplete
input field with autocomplete
input field with autocomplete
Fellini
Luchino Visconti
made first movie in ...
was born in ...
...
Italy
France
...
add
2 then use input with autocomplete or select
use input or select from a list of recommended relations
then use input with autocomplete or select 2

The user will also be able add his own classes and instances or use PoolParty's natural language processing service to analyse web pages or documents to glean new concepts for use in the model.

Of course this PIM will not only consume LOD, but it can also publish the user created knowledge models as part of the LOD cloud. In this way LOD can be harnessed to enable the average internet user to create more Open Knowledge. There will be an online version of this PIM that can be used free of charge.

In the upcoming project LASSO funded by the Austrian Research Promotion Agency (FFG)[13] we will do research on algorithms that enable smart interlinking of local data and LOD sources, which will be used for the PIM. Amongst the algorithmic solutions we will pursue are graph based look-up services (e.g. querying LD sources by taking context into account instead of just searching for keywords), network search methods like Spreading Activation and statistical methods such as the Principal Component Analysis.

8 Final Remarks

We have shown how Open Linked Data can help in various ways with easing the creation of knowledge models like thesauri. At OKCon 2010 we will demonstrated PoolParty and session visitors will learn how to manage a SKOS thesaurus and how

[13] http://ffg.at/content.php

PoolParty supports the user in this process. The document analysis features will be presented, showing how new concepts can be gleaned from text and integrated into a thesaurus.

It will be shown how to interlink local concepts from DBpedia, thereby enhancing one's thesaurus with triples from the LOD cloud. Finally the state of the PoolParty PIM tool will be presented.

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