

Using Linked Data in Thesaurus Management

Thomas Schandl, Andreas Blumauer, Helmut Nagy

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

1 Introduction and aims

PoolParty is a SKOS Thesaurus Management Tool (TMT) aiming to support the creation and maintenance of thesauri by utilizing Linked Data (LD), text-analysis and an easy-to-use GUI, so domain experts can manage thesauri without knowledge about the Semantic Web. PoolParty has been developed to build various thesaurus based applications like semantic search engines, faceted search/browsing, document similarity recommendation, auto-completion, bookmarking and tag recommendation. In order to achieve this, it is necessary to publish the thesauri built with PoolParty and offer methods of integrating them with various applications [1]. PoolParty does this by offering a RESTful web service interface providing thesaurus querying, indexing, search, tagging and linguistic analysis services in order to integrate them with an organization's CMS, DMS or search engine. PoolParty also publishes SKOS thesauri as Linked Data. When creating thesauri a pattern [2] for generating URIs can be chosen, e.g. combining a base URI with the preferred label of a concept or an incrementing number. PoolParty offers the ability to serve HTML or RDF versions of the concept when its URI is dereferenced according to the Linked Open Data principles (LOD) [3].

In PoolParty any concept from a thesaurus can be looked up in e.g. DBpedia, Geonames, Sindice, FreeBase, WordNet, Yago or any other LD source and can be linked to its counterpart there by importing the external URI and its triples. These data are the basis for PoolParty's goal of enriching a thesaurus with relevant information from Linked Data sources, aiming to (1) ease managing and expanding thesauri and (2) improve retrieval of documents by enhanced tag suggestion and document similarity recommendations. These functionalities are currently researched in the LASSO project funded by Austrian FFG and are described below. Another feature is to synchronize thesaurus data via LD interfaces between different systems, this is described in (3).

2 Using Linked Data to aid thesaurus management

Cost effectiveness is a crucial point for thesaurus management in any environment and Linked Data can be a valuable eco-system to provide additional information to glean metadata for thesaurus concepts. As a consequence a LD lookup service can help to (a) populate thesauri semi-automatically, (b) to provide help to disambiguate

and (c) to categorise new concepts and (d) to enrich concepts with information gained from LD.

- (a) One goal of this thesaurus management functionality is to populate a certain `skos:Concept` with proper narrower concepts from LD sources. Example: A local thesaurus has a concept with the preferred label "Expressionist painters" which in turn has three narrower concepts "Paul Klee", "August Macke" and "Wassily Kandinsky". "August Macke" was already linked to the corresponding DBpedia page by a thesaurus manager utilizing PoolParty's built in look-up functionality, thereby becoming what we call an "anchored concept". This gives us information about the categories and classes August Macke belongs to, e.g. Wikipedia categories "Expressionism" and "German painters". Using this information it is easier to automatically disambiguate Kandinsky and Klee and determine their DBpedia URIs. Then we check which DBpedia, Yago, Wordnet, etc. categories these 3 concepts have in common, e.g. Expressionism or Bauhaus and suggest them as a possible matching DBpedia URI for their super-concept in the thesaurus. A curator still has to pick the right URIs, but he can benefit of smarter suggestions and later use the gained links to LD sources to retrieve more expressionist painters and easily expand the thesaurus.
- (b) In a similar way Linked Data can be leveraged to support the disambiguation of a new concept, which was either manually entered or automatically gleaned from PoolParty's information extraction functionality that analyses documents and finds statistically relevant terms. When "William Forsyth" is added as a narrower concept for "Impressionist Painters", the system has a lot of information from anchored parent and sibling concepts to successfully disambiguate him from the economist and botanists of the same name.
- (c) Concepts that were gleaned from information extraction have to be categorised by humans in order to put them into the right place in a thesaurus. It would be beneficial for a thesaurus curator to receive recommendations, which categories seem to be promising candidates to serve as broader concepts for the new concept. This could be done by looking up its label in e.g. DBpedia look-up service and comparing the graphs of candidate resources from the LOD cloud to the local thesaurus to recommend the proper place and type of the new `skos:Concept`. In this way the system can come up with the suggestion to use the new concept with the label "Paul Cezanne" as a narrower concept of "Expressionist Painters".
- (d) Linked Data can also be used to enrich existing concepts of a thesaurus, e.g. the DBpedia abstract can become a `skos:definition` or alternate names from Geonames can become `skos:altLabels`. Linked Data can also extend a concept, e.g. 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.

3 Using Linked Data for improving document retrieval

Recommending similar documents that are related to a given document is a valuable service for a DMS, CMS or Wiki. Calculating "similarity" between documents by using not only simple tags but also annotation represented by concepts from a thesaurus which have even more metadata from the LOD cloud is a promising approach. Thesauri with anchored concepts derived from services described above can be a solid basis for such improved recommendation services.

Indexing and describing the meaning of a document in the PoolParty means that the following data is indexed:

- a) full-text of the document,
- b) labels of concepts used to tag the document,
- c) labels from concepts related to these tags in the local thesaurus via skos:broader, skos:narrower or skos:related
- d) and labels from categories derived from DBpedia associated with b) and c).

This results in even richer metadata for each tagged document, e.g. a wiki page tagged with anchored concepts "MS Sharepoint", "Atlassian Confluence" and "Enterprise 2.0" is also associated with various other related labels and with DBpedia categories like "Web 2.0", "Online social networking", "Wiki", "Buzzwords", "DMS" to name a few. We call this metadata for a tagged document "semantic profile", which can help to calculate similarity between documents (note that the different aspects b) c) and d) can have different boost values in these calculations).



4 Keeping Linked Data up to date

PoolParty is able to import a SKOS thesaurus from a Linked Data server, additionally it may also receive updates to thesauri imported this way. This feature has been implemented in the course of the KiWi project funded by the European Commission [4]. The KiWi system also contains SKOS thesauri and exposes them as LD. Both systems can read a thesaurus via the other's LD interfaces and may write it to their own store. This is facilitated by special LD 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.

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.

References

- [1] Viljanen, K., Tuominen, J., Hyvänen, 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)
- [2] Dodds, L., Davis, I.: Linked Data Patterns. A pattern catalogue for modelling, publishing, and consuming Linked Data. <http://patterns.dataincubator.org/book/>  (accessed June 19, 2010)
- [3] Berners-Lee, T.: Linked Data. <http://www.w3.org/DesignIssues/LinkedData.html>  (accessed June 21, 2010).
- [4] KiWi project - Knowledge in a Wiki. <http://kiwi-project.eu/> , (accessed June 21, 2010)