Extending Ontologies with Free Keywords in a Collaborative Annotation Environment

Matias Frosterus, Mika Wahlroos, and Eero Hyvönen

Semantic Computing Research Group (SeCo)
Aalto University School of Science, Dept. of Media Technology, and
University of Helsinki, Dept. of Computer Science
http://www.seco.tkk.fi/
firstname.lastname@aalto.fi

Abstract. Semantic web technologies have introduced the idea of annotating content in terms of concepts taken from ontologies. Since concepts are defined in terms of properties and relations to other concepts, descriptions grow up into larger RDF graphs that can be used as a basis for data integration and intelligent information retrieval. Since ontologies do not typically contain all the possible concepts needed for annotation, it is usually necessary to offer the annotator the possibility to introduce new free keywords or tags in addition to the predefined ontology concepts. The problem then is that free keywords/tags do not have ontological connections to the rest of the RDF graph, unless such relations are defined by the annotator. We present a process for integrating free keywords into the ontological framework, and a practical tool implementation of it, discussing the challenges and possibilities introduced by the system. We also describe a case study performed for the Finnish Defence Forces, where the tool is used for creating a faceted semantic search portal featuring the free keywords and the ontological concepts at the same time.

1 Introduction

1.1 Position Statement

A large amount of metadata is being produced through free keywords, or tags, on the web allowing for a robust, easy-to-use, and flexible annotation of content. Ontologies offer an easy way to impose structure and meaning to the free keywords linking the annotated material into the larger framework of the Semantic Web.

1.2 The challenges of free tagging

A common practice in community-based annotation is to allow the users to create the needed terms, or tags, freely when describing objects. This facilitates flexibility in annotations and makes it easier for novice users to describe things. On the other hand, in the professional metadata world (e.g., in museums, libraries, and archives) using shared pre-defined thesauri is usually recommended for enhancing interoperability between annotations of different persons, and enhancing search precision and recall in end-user

applications. Both approaches are usually needed, and can also be supported to some extent by e.g. suggesting the use of existing tags.

A more advanced approach than using thesauri is to use ontologies [6] for harmonizing content indexing. Then indexing is based on language-independent concepts referred to by URIs, and keywords are labels of the actual underlying concepts. Defining the meaning of indexing terms by their properties and relations to other concepts allows for better interoperability of contents and their use by machines. This is important in application areas, such as semantic search, recommending, linking, and automatic indexing. With even a little extra work, e.g. by just systematically organizing concepts along subclass hierarchies and partonomies, substantial benefits can be obtained [2].

Free keywords are needed in many situations:

1. There can be omissions in the ontology that should be added, but are not currently there.
2. Concepts for new things and phenomena that have not yet been added to the ontology may be needed in annotations.
3. The number of concepts, e.g., the names of plants, can be too numerous to be included in the ontology, but can still be needed in annotations.
4. Instance data, e.g., persons, places, events etc. can be needed in annotations.

There is a need for a system that integrates new free keywords into the wider framework of ontologies in an annotation environment. As a solution, we present a system and its implementation for introducing free keywords into ontologies. The next section presents a general overview of the process. After this a specific implementation in a case study done for the Finnish Defence Forces is presented. Finally, we conclude with discussing related and future work.

2 Using Free Keywords in Annotations

Our key problem is how to incorporate metadata with free keywords into an ontology-driven annotation environment in a simple way that does not require ontology modeling knowledge from the annotators. This requires that the free keywords must be turned into a compatible, machine-readable RDF form, and that the relations between the free keywords and the existing ontologies must be established.

The first step in the process depicted in Figure 1 is to go through the free keywords used in the annotations (1) and match as many as possible to existing ontological concepts (2). Keywords should be transformed into the base form and the strings compared to the labels in the ontology.

Keywords that did not match to ontology concepts are then made into RDF objects with the original keyword as the label (3). The class for these should be kept separate from the class of the concepts in the ontology since these have not been approved by ontology developers, and are therefore less reliable than the proper ontological concepts. At this stage, the keyword object can be used in further annotations, and the list can be edited and pruned as needed. However, at this point it does not offer much additional usability compared to existing tagging systems based on using isolated tags.

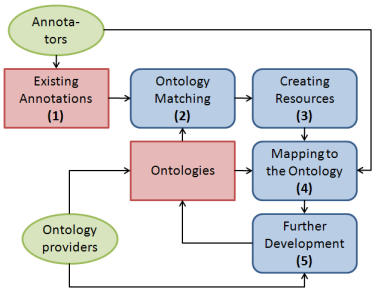

Fig. 1. The process of utilizing free annotations in an ontology-driven annotation environment

In order to take full advantage of using ontologies, the keyword objects should be mapped to the existing ontology (4), typically through the `rdfs:subClassOf` property. Also other relations such as partonomy or equivalence can be used. The keyword objects also do not need to be connected directly to the ontology, but rather can be connected to other keyword objects that are in turn connected to the ontology. When the ontology is developed further (5), the keywords that have been used the most make for prime candidates to be included into the next version of the ontology.

There should be a way, however, for the annotators to keep some keywords out from ontological development if the annotator knows that the keyword will not be of interest to the ontology developers or if the keyword itself is such that it is not wanted to be accessible to the wider public. This latter case is more likely in situations where the annotators are working with sensitive data. The same mechanics can be used by the ontology developers themselves to mark free keywords that they have reviewed but not deemed fit for the ontology.

When new free keywords are needed, the annotator can align them with other ontological concepts straightaway and thus make its meaning explicit within the annotation framework used, leading to less ambiguity. Furthermore, by using literal properties, the annotator can provide detailed explanations of the concept to human readers, and include e.g. labels in different languages, acronyms, and synonyms for the keyword.

A system realizing the process should fulfill the following requirements:

- facilitate finding ontological concepts and free keyword objects for annotations,
- allow the creation of new free keyword objects,
- facilitate the mapping of new free keyword objects to each other and to ontological concepts, and
- instantly show new keyword objects to other annotators and allow their use.

Finally, all of this should be doable without technical expertise, with the application hiding the complexities of the RDF model in the background.

3 Case Study: The Finnish Defence Forces' Norms

The process was implemented in a project done for the Finnish Defence Forces' norms database. The norms comprise of documents describing procedures and regulations as

well as the associated metadata in XML format. The goal of the project was to implement a faceted search portal for the norms utilizing the semantic web technologies.

Metadata about documents included annotations about the subject of the norms using keyword from the Defence Administration's Thesaurus as well free keywords chosen by the annotators. The free keywords contained some spelling mistakes as well as multiples of some keywords (i.e. a singular and a plural form of the same keyword).

For the ontology we used the Finnish Defence Administration's Ontology PUHO[1] which is a domain ontology comprised of concepts relevant to the Finnish Defence Forces developed from the Defence Administration's Thesaurus that has been in use for the annotations of the organization's documents. PUHO extends the General Finnish Upper Ontology YSO[2] so it was also included in the project. For easy use in different applications, the ontology is hosted in the ONKI ontology service[9], which contains several different interfaces for easy integration into other systems and applications.

The metadata was transformed into RDF using a custom conversion process which involved matching keywords present in the metadata with concepts defined in the ontology. Lemmatized forms of the keywords were first obtained in order to identify different inflected forms of the same word, and the lemmatized keywords were then matched with similarly lemmatized labels of ontological concepts using strict string matching. Keywords that did not match the label of any ontological concept were included as new RDF resources with their own URIs.

Once the conversion was ready, the RDF was loaded into the SAHA3 metadata editor [4], which is easily configurable to different schemas, can be used by multiple annotators simultaneously, and works in a normal web browser, therefore needing no special software to be installed. The support for multiple annotators is implemented in a robust way with synchronization and locks which guarantee that the annotators don't interfere with each other's work. The tool also includes a chat channel in case online dicussions between annotators are needed. Using SAHA3, the annotators can collaboratively clean up the free keywords as needed and map them to the ontology, and SAHA3 realizes the requirements set in section 2. SAHA3 is available as open source at Google Code[3].

For the publication of the metadata, SAHA3 is integrated with the multi-faceted search portal generator HAKO that provides easy access to the datasets from different faceted viewpoints. The facets are built automatically based on the properties of the metadata according to a simple configuration description, and the faceted search application is complemented by free text search. HAKO works in a normal web browser allowing easy access to the data from anywhere. For machine use, SAHA3 and HAKO have two machine APIs: one for using the content as an ONKI ontology service [7] for annotation work, and one for using the content via a SPARQL end-point, which can be used by other applications to access all the metadata as needed.

In our case, one of the facets was the subject of the norms featuring both the ontological concepts from PUHO as well as the new free keyword objects. The hierarchical

[1] http://onki.fi/en/browser/overview/puho
[2] http://www.seco.tkk.fi/ontologies/yso/
[3] http://code.google.com/p/saha/

facet contains both types of concepts integrated so that a user does not see a difference between them since the inner workings of the system are of no interest to the user.

4 Discussion and Related Work

This paper presented a process of bringing free keyword annotations into the framework of an ontology-driven annotation system, detailing the different steps necessary as well as the requirements for the tools that facilitate this process. A case study where this was done was presented and the tools used.

Folksonomies and ontologies have been combined before [3, 8, 5] but much of the focus has been on blogs and similar domains where the annotations have been done by the public within a completely free framework, as opposed to professional annotators working with free keywords in tandem with a controlled vocabulary. Others have built domain ontologies based on partially controlled and partially free tagging data and discussed the need to merge future development of the controlled tag vocabulary with the ontology [1]. Our work is more focused on the process of bringing the free keywords into the ontological framework as opposed to using them to build new ontologies or to permanently extend existing ones.

In addition to processes for manually defining relations between isolated tags and ontological concepts, ontologies have also been derived from folksonomies using automatic or semi-automatic methods based on machine learning [3]. Much of the work has focused on discovering implicit semantic relations between tags based on statistical analysis of connections between users, tags, and the objects tagged by the users. The focus of our work is on relatively sparse free keyword data which may not lend itself well to using statistical analysis of the tagging data as the primary technique.

Next, our goal is to try to devise ways to facilitate mapping the free keywords into the ontology easier by trying to reason possible relations from their usage alongside the ontology terms. This could also be used to find out relations between the keywords themselves. We also intend to evaluate the benefits of the system described in the case study from the perspective of practical use cases in document management and search of the norms database.

Acknowledgements This work is part of the National Semantic Web Ontology project in Finland[4] FinnONTO (2003–2012), funded currently by the National Technology and Innovation Agency (Tekes) and a consortium of 35 public organizations and companies.

References

1. Mihai Codescu, Gregor Horsinka, Oliver Kutz, Till Mossakowski, and Rafaela Rau. OSMonto - an ontology of OpenStreetMap tags. In *State of the map Europe (SOTM-EU) 2011*, 2011.
2. Eero Hyvönen, Kim Viljanen, Jouni Tuominen, and Katri Seppälä. Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In *Proceedings of the ESWC 2008, Tenerife, Spain*. Springer–Verlag, 2008.

[4] http://www.seco.tkk.fi/projects/finnonto/

3. Hak Lae Kim, Simon Scerri, John G. Breslin, Stefan Decker, and Hong Gee Kim. The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, pages 128–137. Dublin Core Metadata Initiative, 2008.
4. Jussi Kurki and Eero Hyvönen. Collaborative metadata editor integrated with ontology services and faceted portals. In *Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece*. CEUR Workshop Proceedings, http://CEUR-WS.org, 2010.
5. Alexandre Passant. Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. In *ICWSM'2007*, 2007.
6. S. Staab and R. Studer, editors. *Handbook on ontologies (2nd Edition)*. Springer–Verlag, 2009.
7. Jouni Tuominen, Matias Frosterus, Kim Viljanen, and Eero Hyvönen. ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services. In *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, 2009. Springer–Verlag.
8. Céline Van Damme, Martin Hepp, and Katharina Siorpaes. FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies. In *Bridging the Gep between Semantic Web and Web 2.0 (SemNet 2007)*, pages 57–70, 2007.
9. Kim Viljanen, Jouni Tuominen, and Eero Hyvönen. Ontology libraries for production use: The Finnish ontology library service ONKI. In *Proceedings of the ESWC 2009, Heraklion, Greece*. Springer–Verlag, 2009.