Creating and Publishing Semantic Metadata about Linked and Open Datasets

Matias Frosterus and **Eero Hyvönen** and **Joonas Laitio**
Aalto University and University of Helsinki
http://www.seco.tkk.fi/ – firstname.lastname@aalto.fi

Abstract

We present a comprehensive system for producing interoperable metadata for Linked Open datasets and governmental datasets published in various formats.

Introduction

The number of open datasets available on the web is increasing rapidly with the rise of the Linked Open Data (LOD)(Heath and Bizer 2011) cloud and various governmental efforts for releasing public data in various formats, not only in RDF. However, the metadata available for these datasets is often minimal, heterogeneous, and distributed, which makes finding a suitable dataset for a given need problematic. Governmental open datasets are often the basis of innovative applications but the datasets need to be found by the developers first.

There are search engines for finding RDF and other datasets. However, using such systems—based on the Google-like search paradigm—it is difficult to get the general picture of the contents of the *entire* cloud of the offered datasets. Furthermore, finding suitable datasets based on different selection criteria such as subject topic, size, licensing, publisher, language etc. is not supported.

To address the problem, we present a distributed content creation model and tools for annotating and publishing metadata about linked data and non-RDF datasets on the web called DATAFINLAND. At the moment, the most widely used system for annotating and publishing datasets is CKAN[1] by the Open Knowledge Foundation. DATAFINLAND differs from CKAN by being based on semantic web technologies, facilitating more accurate machine processable annotations, semantic interoperability in distributed content creation, semantic search and browsing for human end-users, and RDF-based APIs for machines.

Overview

Figure 1 depicts the generic components and steps needed for producing and publishing metadata about datasets. In the figure, we have marked the tools and resources used in our proof-of-concept system DATAFINLAND in parentheses, but the process model itself is general and different solutions could be substituted.

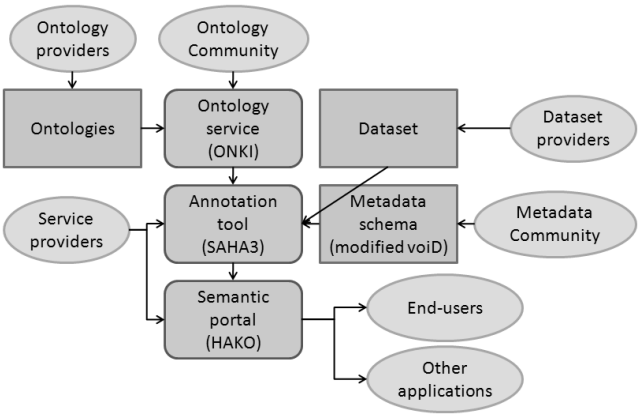

Figure 1: The distributed process of producing and publishing metadata about (linked) datasets

The process begins with the publication of a dataset by its provider (upper right) for which metadata is then recorded. In our application scenario, content creators are not literate in semantic web technologies, and annotate datasets in different, distributed organizations on the web. For this task, a human-friendly metadata editor that hides the complexities of RDF and OWL is needed. The editor should also allow for many simultaneous users without creating conflicts, and the results of annotations, e.g., creating a new organization instance or modifying an existing one, should be instantly seen by every other user. Otherwise multiple representations and URIs for the same object could be easily created.

For DATAFINLAND we used and developed further the SAHA3 metadata editor (Kurki and Hyvönen 2010), which is easily configurable to different schemas, can be used by multiple annotators simultaneously, and works in a normal web browser, therefore requiring no special software to be installed. The support for multiple annotators is made 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, if online dicussion between annotators is needed. SAHA3 is available as open

[1] http://ckan.org/

source at Google Code[2].

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. For example, in the domain of open datasets, CKAN uses free tagging, but always also suggests the use of existing ones by autocompletion. This has the benefit that new tags are easy to add but, at the same time, there is a possibility for sharing them. However, the problem with traditional tag-based systems is that it is easy to end up with several different tags that mean the same thing, and in turn a single tag may end up denoting several different things, because the meaning in tags in not explicitly defined anywhere. This is problematic from both a human and a machine use point of view.

A more advanced approach is to use ontologies (Staab and Studer 2009) where indexing is based on language-free concepts referred to by URIs, and keywords are labels of the actual underlying concepts. Defining the meaning behind the index terms in an explicit way, and furthermore by describing the relations between the different concepts, allows for better interoprability of contents and their use by machines. This is important in many application areas, such as semantic search, information retrieval, semantic linking of contents, 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 (Hyvönen et al. 2008b).

DATAFINLAND uses the ONKI Ontology service, which provides a rich environment for using ontologies as web services (Hyvönen et al. 2008a) as well as for browsing and annotation work. ONKI offers traditional web service (WSDL and SOAP) and AJAX APIs for easy integration to legacy applications, such as cataloging systems and search engines, and provides a robust platform for publishing and utilizing ontologies for ontology developers. The simplest way to use ONKI in providing controlled vocabularies for an application is through the Selector Widget. It is an extended HTML input field widget that can be used for mash-ups on any HTML page at the client side with two lines of Javascript code. The widget could be added to, for example, the CKAN web browser based editor, providing then the new possibility of using ontology references as tags in annotations. The ONKI widget provides its user ready-to-use ontology browser functionalities, such as concept finding, semantic disambiguation, and concept (URI) fetching.

Aside from using a controlled vocabulary for describing the open datasets, another important consideration is the choice of annotation schemas that are used. If ontologies define the vocabulary, the schemas can be seen as the topics in the description outlining the information that should be recorded. The aim is to provide a concise, machine usable description of the dataset and how it can be accessed and used.

For DATAFINLAND, we chose the Vocabulary of Interlinked Datasets (VoiD), an RDF schema for describing linked datasets (Alexander et al. 2009), as the starting point for our schema. One of the guiding principles behind the design of VoiD was to take into account clear accessing and licensing information of the datasets resulting in efficient discovery of datasets through search engines. Furthermore, VoiD realized effective dataset selection through content, vocabulary, and interlinking descriptions, and, finally, query optimization through statistical information about the datasets.

Finally, the metadata needs to be made available. In DATAFINLAND, we used HAKO, a faceted search engine, to publish the metadata recorded in the SAHA3 project as a readily usable portal (Kurki and Hyvönen 2010). The RDF data produced in SAHA3 is directly available for HAKO, which is integrated over the same index base as SAHA3. The result is a semantic portal for human end-users supporting faceted search based on the different properties used in the metadata such as language, license or subject. In addition, a complementary traditional free text search engine is also provided. For machine use, HAKO has a SPARQL endpoint which can be used to access the metadata from the outside as a service, in addition to accessing the HAKO portal via the human interface.

Since HAKO and SAHA3 are built on the same index, the search engine is fully dynamic: changes made to the data in SAHA3 are immediately visible in HAKO. This is useful for making real time changes to the data, such as creating new search instances, or updating old metadata, such as the license of a dataset or its other descriptions.

[2]http://code.google.com/p/saha/

References

Alexander, K.; Cyganiak, R.; Hausenblas, M.; and Zhao, J. 2009. Describing linked datasets - on the design and usage of void, the 'vocabulary of interlinked datasets'. In *Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09)*.

Heath, T., and Bizer, C. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, San Francisco, USA.

Hyvönen, E.; Viljanen, K.; Tuominen, J.; Seppälä, K.; Kauppinen, T.; Frosterus, M.; Sinkkilä, R.; Kurki, J.; Alm, O.; Mäkelä, E.; and Laitio, J. 2008a. National ontology infrastructure service ONKI.

Hyvönen, E.; Viljanen, K.; Tuominen, J.; and Seppälä, K. 2008b. Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In *Proceedings of the ESWC 2008, Tenerife, Spain*. Springer–Verlag.

Kurki, J., and Hyvönen, E. 2010. 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/.

Staab, S., and Studer, R., eds. 2009. *Handbook on ontologies (2nd Edition)*. Springer–Verlag.