DataFinland—A Semantic Portal for Open and Linked Datasets

Matias Frosterus, Eero Hyvönen, and Joonas Laitio

Semantic Computing Research Group (SeCo)
Aalto University and University of Helsinki
http://www.seco.tkk.fi/
firstname.lastname@tkk.fi

Abstract. The number of open datasets available on the web is increasing rapidly with the rise of the Linked Open Data (LOD) cloud and various governmental efforts for releasing public data in different formats, not only in RDF. The aim in releasing open datasets is for developers to use them in innovative applications, but the datasets need to be found first and metadata available is often minimal, heterogeneous, and distributed making the search for the right dataset often problematic. To address the problem, we present DataFinland, a semantic portal featuring a distributed content creation model and tools for annotating and publishing metadata about LOD and non-RDF datasets on the web. The metadata schema for DataFinland is based on a modified version of the voiD vocabulary for describing linked RDF datasets, and annotations are done using an online metadata editor SAHA connected to ONKI ontology services providing a controlled set of annotation concepts. The content is published instantly on an integrated faceted search and browsing engine HAKO for human users, and as a SPARQL endpoint and a source file for machines. As a proof of concept, the system has been applied to LOD and Finnish governmental datasets.

1 Metadata for Linked Datasets

Linked Data refers to data published on the web in accordance with four rules[1] and guidelines [2] that allow retrieving metadata related to data entities, and linking data within and between different datasets. The datasets and their relations are represented using RDF (Resource Description Framework) and entities are identified by Uniform Resource Identifiers (URIs)[2], which allows the use of the Hypertext Transfer Protocol (HTTP) to retrieve either the resources themselves, useful descriptions of them, or links to related entities [3].

The Linked Open Data community project[3] has collected a large number of datasets and mappings between them. However, little metadata about the datasets is provided aside from short, non-uniform descriptions. As the number of linked datasets [8] grows, this approach does not allow for easy understanding of what kind of dataset are offered, who provides them, what is their subject, how they interlink with each other, possible

[1] http://www.w3.org/DesignIssues/LinkedData.html
[2] http://www.w3.org/TR/uri-clarification/
[3] http://linkeddata.org