# Linked Data Finland: Towards a 7-star Service Platform for Linked Datasets

Eero Hyvönen, Miika Alonen, Jouni Tuominen, and Eetu Mäkelä

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

**Abstract.** The idea of opening data on the Web as Linked Data (LD) is widely adopted in areas such as public government, science, libraries, and cultural heritage. The key idea is to harmonize, integrate, enrich, and re-use existing data repositories in a cost-efficient way via standard APIs in novel applications. This paper concerns two major hindrances for re-using LD: It is often difficult for a re-user to understand the 1) characteristics of the dataset and 2) evaluate the quality of the data for her intended purpose. This paper introduces the "Linked Data Finland" publishing platform LDF.fi addressing these issues. In order to enhance and promote reusability, we propose extending the famous 5-star model of Tim Berners-Lee into a 7-star model: The sixth star requires that the dataset is defined and explained in terms of explicit schemas. Explicit schemas make it possible to explain the re-user the intended characteristics of the data by, e.g., documentation about the schemas, and how the schemas (vocabularies) are actually used in the given dataset. The seventh star is given, if the data has also been validated w.r.t. the schema specifications. The results of the validation may be a human readable document and/or a machine readable reprentation regarding the quality issues found in the data. This paper reports about work in progress, but the first prototype of the platform is already operational on the web as a service http://ldf.fi.

## 1 Use Cases

The platform LDF.fi has two user-groups: 1) application developers, and 2) data publishers. LDF.fi provides the developers with standard Linked Data (LD) services, including machine APIs for downloading datasets, graphs, and data related to individual resources. The cornerstone is a SPARQL endpoint on which the other data services are based on. The developers are also provided with browser-based lookup-services for viewing resource descriptions and for inspecting the Linked Data by browsing in a human readable form, as customary in LD.

For data publishing, the idea is to automate the process as far as possible in the following way: The publisher creates a Service Description of the dataset and its schemas, using an extended version of the W3C recommendation[1]. Based on such metadata, LDF.fi 1) automatically sets up the technical services, such as the SPARQL endpoint, content negotiation, etc., and 2) generates a web page that explains the dataset, schemas, and provides additional related services with ready-to-use interactive query forms, etc., for documenting, inspecting, and validating the data.

## 2 Research Questions

Our research particularly focuses on LD services related to the stars 6 and 7. Our goal is that at LDF.fi the user could more easily get an understanding of how suitable the data is for her application purpose, i.e., 1) what the data actually is and 2) what is the quality of it w.r.t. re-using the data.

Firstly, we encourage publishers, by giving the 6th star, to publish not only datasets but also explicit schemas (vocabularies) used in them, unless the schemas are not already available somewhere on the Semantic Web. For example, in many datasets properties related to a resource are taken from several vocabularies, such as Dublin Core (DC) and FOAF. Even though DC[2] and FOAF[3] are well-specified on the Web, their combination in a particular dataset may not

[1] http://www.w3.org/TR/sparql11-service-description/
[2] http://dublincore.org/
[3] http://semanticweb.org/wiki/FOAF

be, and there may be additional implicit contraints attached to using them regarding, e.g., the cardinality and range of the properties. There may also be dataset specific properties in use in addition to DC and FOAF. We propose that such decisions and practices should be explicated as schemas to the end-users to enhance re-usability.

Schemas can be documented automatically in LDF.fi for the human reader using a schema documentation generator, in our case SpecGen[4]. In the LD world datasets often use schemas (vocabularies) for which definitions or descriptions are not available, but are embedded in the data itself.

In order to find out how schemas are actually used in a dataset, including both published and unpublished schemas, we are creating a service vocab.at[5] that analyzes a given dataset from this perspective and creates an HTML document listing, e.g., statistics of vobaculary usage and raising up issues detected, e.g., if an IRI is not dereferenceable. The input for vocab.at is either an RDF file, a SPARQL endpoint, or an HTML page with embedded RDFa markup. Work on visualizing data and schemas is underway.

To earn the 7th star, the quality of the dataset w.r.t. the schemas used in it must be explicated to the end-user, so that she can evaluate whether the data quality matches her needs. To some extend, the analysis of vocab.at answers also to this question. Additional work on schema based validation is being carried out regarding our data curation tool SAHA[6] on top of a SPARQL endpoint, and is being incorporated in LDF.fi as a service. SAHA will also include a configurable LD Browser that will be available in LDF.fi in addition to the URIBurner[7] service used first.

## 3 Related Work

There are lots of LD platforms on the web: in Life Sciences alone there is LinkedLifeData[8], NeuroCommons[9], Chem2Bio2RDF[10], HCLSIG/LODD[11], BioLOD[12], and Bio2RDF[13]. Several tools have been created for vocabulary documentation, such as SpecGen, neologism[14], ldodds/dowl[15], parrot[16], OWLDoc[17], OntologyBrowser[18], and LODE[19]. Our work aims to extend the state-of-the-art with tools that make re-use of LD easier for end-users.

**Acknowledgements** This work is part of the Linked Data Finland project[20] funded by Tekes and a consortium of 22 public organizations and companies.

[4] https://bitbucket.org/wikier/specgen/wiki/Home
[5] http://vocab.at
[6] http://www.seco.tkk.fi/tools/saha
[7] http://linkeddata.uriburner.com/
[8] http://linkedlifedata.com/
[9] http://neurocommons.org/
[10] http://chem2bio2rdf.wikispaces.com/
[11] http://www.w3.org/wiki/HCLSIG/LODD
[12] http://biolod.org/
[13] http://bio2rdf.org/
[14] http://neologism.deri.ie/
[15] https://github.com/ldodds/dowl
[16] http://ontorule-project.eu/parrot/parrot
[17] http://code.google.com/p/co-ode-owl-plugins/wiki/OWLDoc
[18] http://code.google.com/p/ontology-browser/
[19] http://www.essepuntato.it/lode
[20] http://www.seco.tkk.fi/projects/ldf/