Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets

Eero Hyvönen, Jouni Tuominen, Miika Alonen, 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 Linked Data is to aggregate, harmonize, integrate, enrich, and publish data for re-use on the Web in a cost-efficient way using Semantic Web technologies. We concern two major hindrances for re-using Linked Data: It is often difficult for a re-user to 1) understand the characteristics of the dataset and 2) evaluate the quality the data for the intended purpose. This paper introduces the "Linked Data Finland" platform LDF.fi addressing these issues. We extend the famous 5-star model of Tim Berners-Lee, with the sixth star for providing the dataset with a schema that explains the dataset, and the seventh star for validating the data against the schema. LDF.fi also automates data publishing and provides data curation tools. The first prototype of the platform is available on the web as a service, hosting tens of datasets and supporting several applications.

1 Publishing Linked Data

Lots of Linked Data (LD) platforms have emerged on the Web since the publication of the four Linked Data publication principles and the 5-star model[1]. For example, in Life Sciences alone there are LinkedLifeData[2], NeuroCommons[3], Chem2Bio2RDF[4], HCLSIG/LODD[5], BioLOD[6], and Bio2RDF[7].

LDF.fi[8] contributes to the current state-of-the-art of Linked Data publishing [2] as follows: 1) We propose extending the 5-star model[9] into a 7-star model, with the goal of encouraging data publishers to provide their data with explicit metadata schemas and to validate their data for better quality. 2) LDF.fi automates the data publishing process so that not only a SPARQL endpoint but also a rich set of additional data services are generated automatically based on the metadata about the dataset and its graphs. 3) LDF.fi

[1] http://www.w3.org/DesignIssues/LinkedData.html
[2] http://linkedlifedata.com/
[3] http://neurocommons.org/
[4] http://chem2bio2rdf.wikispaces.com/
[5] http://www.w3.org/wiki/HCLSIG/LODD
[6] http://biolod.org/
[7] http://bio2rdf.org/
[8] Our work is funded by Tekes and a consortium of 20 public organizations and companies.
[9] http://5stardata.info/

provides end users with additional tools and documentation for publishing, curating, and re-using the datasets. This paper first explains these ideas, and then presents the actual service available online[10].

2 7-star Linked Data

A major hindrance of re-using a dataset is the difficulty to evaluate how suitable the data is for the application purpose at hand. Datasets often use schemas (vocabularies) for which definitions or descriptions are not available, but are embedded in the data itself. This makes it difficult to figure out the characteristics of the data. Furthermore, given the data and its schema it may be difficult to say how well the data actually matches the schema; there are lots of data quality problems on the Semantic Web[11].

To address these issues, we encourage data publishers by two extra stars:

- The 6th star is given if the schemas (vocabularies) used in the dataset are explicitly described and published alongside the dataset, unless the schemas are already available somewhere on the Web.
- For the 7th star, the quality of the dataset against the schemas used in it must be explicated, so that the user can evaluate whether the data quality matches her needs.

LDF.fi provides supporting tools related to these issues: First, schemas are documented automatically for the human reader by using a schema documentation generator. In our case, the LODE[12] online service is employed. (Other possible tools for schema documentation include SpecGen, Neologism[13], dowl[14], Parrot[15], OWLDoc[16], and OntologyBrowser[17].) Second, in order to find out how schemas are actually used in a dataset, we created a new service http://vocab.at [1]. It analyses a dataset, creates an HTML report that explains vocabulary usage in the data, and reports issues of undefined properties or unresolvable namespaces. The input for vocab.at is either an RDF file, a SPARQL endpoint, or an HTML page with embedded RDFa markup.

3 Automatic Service Generation

LDF.fi tries to automate the process of publishing datasets as far as possible in the following way: The publisher is expected to create an RDF dataset with minimal metadata about it and its schemas. Here an extended version of the new W3C Service Description recommendation[18] and the VoID vocabulary[19] can be used, and the data is stored

[10] http://www.ldf.fi/
[11] http://pedantic-web.org/
[12] http://www.essepuntato.it/lode
[13] http://neologism.deri.ie/
[14] https://github.com/ldodds/dowl
[15] http://ontorule-project.eu/parrot/parrot
[16] http://code.google.com/p/co-ode-owl-plugins/wiki/OWLDoc
[17] http://code.google.com/p/ontology-browser/
[18] http://www.w3.org/TR/sparql11-service-description/
[19] http://rdfs.org/ns/void

into the SPARQL endpoint. Alternatively, a simple JSON object listing the dataset and graph names, human readable labels, and a description of the data can be provided. In the metadata, it is also possible to give an example URI pointing into the dataset, a SPARQL query example for querying the data, and optionally a link to possible visualizations of the dataset. Based on such metadata, LDF.fi generates for each dataset a home page on which the following functionalities are available for re-users:

1. Links for downloading datasets and graphs are provided (if licensing permits it).
2. Schemas can be downloaded if provided with the data, and links to their documentation are provided (when available).
3. Following forms are created for inspecting the dataset in more detail: 1) Given a URI the corresponding RDF description can be read in various formats (Turtle, RDF/XML, RDF/JSON, N3, N-triples) for human consumption in a browser. The example URI is used as a first choice to try out. 2) Given a URI, Linked Data browsing can be started from it, with the example URI as a starting point.
4. There is a SPARQL query form for querying the service with the given query used as a first example.
5. Links providing Vocab.at analysis reports of the graphs in the dataset are provided. They tell the end-user what schemas (vocabularies) are used in the data, and how they have been used. Issues on data quality are pointed out.
6. SPARQL Service Descriptions of the datasets are provided, if available. LDF uses W3C SPARQL Service Description recommendation for this.
7. Links to visualizations of the data that may give the re-user more insight on how the dataset can be used in applications.
8. Licensing conditions of the dataset are provided as well as a label of 1–7 stars.

4 Data Curation Tools

Data curation refers to activities and processes done to create, manage, maintain, and validate data. In LDF.fi several data curation services are available for analyzing textual data and for creating semantic annotations (semi-)automatically from them:

1. SeCo Lexical Analysis Services[20] can be used for language recognition, lemmatization, morphological analysis, inflected form generation, and hyphenation.
2. ARPA Automatic Text Annotation System[21] can be used for extracting Linked Data from unstructured texts.
3. SAHA[22] tool can be used for investigating and editing LDF.fi datasets interactively in real time. In LDF.fi we modified and extended SAHA to work on top of any standard SPARQL endpoint. SAHA is now used as a Linked Data Browser in LDF.fi in the same vein as, e.g., URIBurner[23]. Using SAHA as an editor service for a dataset requires permission from the LDF.fi team.

[20] http://demo.seco.tkk.fi/las/
[21] http://www.seco.tkk.fi/services/arpa/
[22] http://www.seco.tkk.fi/tools/saha
[23] http://linkeddata.uriburner.com/

In our work, we are also using some external tools, such as the SILK Framework[24] for linking data.

5 The Service

In addition to dataset home pages, the LDF.fi portal includes the following pages available through menu links: *Project* page describes the underlying national Linked Data Finland initiative; *Datasets* lists the datasets in the system and links to their home pages; *Schemas* lists the schemas in the system; *Services* explains what kind of services LDF.fi provides; *Policies* documents URI minting and licensing policies in use; *Documentation* explains dataset documentation features of the portal; *Validation* explains dataset validation features of the portal; *Applications* lists application examples of the portal datasets; *Your Data?* tells how external users can get their data published in LDF.fi.

The first datasets available in LDF.fi include: Finnish DBpedia as a service; various Cultural Heritage datasets including, e.g., BookSampo, whose deployed end-user application[25] has 65,000 monthly users; history datasets Semantic National Biography (6,300 biographies as Linked Data) and events of World War I (in collaboration with University of Colorado Boulder); Finnish Law first time as Linked Open Data; Aalto University Linked Open Data[26]; two Linked Science datasets about ornithological observations and weather data; various ontologies used by the ONKI Ontology Service[27]; a linked news dataset. The LDF.fi service is implemented using a combination of the Fuseki SPARQL server[28] for serving primary data, and the Varnish web application accelerator[29] for routing URIs to pertinent applications as well as content negotiation.

6 Evaluation in a Living Lab Environment

LDF.fi was opened officially in January 2014. The platform is being evaluated by providing the service in an open Living Laboratory environment for data publishers and application developers. References to first data applications can be found in the applications page of the portal[30].

References

1. Alonen, M., Kauppinen, T., Hyvönen, E.: Vocab.at - automatic linked data documentation and vocabulary usage analysis (2013), manuscript, http://www.seco.tkk.fi/publications/submitted/alonen-et-al-vocab.pdf
2. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Morgan & Claypool, Palo Alto, California (2011), http://linkeddatabook.com/editions/1.0/

[24] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/
[25] http://www.kirjasampo.fi/
[26] The service http://data.aalto.fi/ is based on LDF.fi.
[27] http://www.onki.fi
[28] http://jena.apache.org/documentation/serving_data/
[29] https://www.varnish-cache.org/
[30] http://www.ldf.fi/applications/