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

licensing conditions, and so on. Such information should be available both to human users as well as machines of the Semantic Web.

Aside from properly linked datasets in RDF format, various organizations have also began publishing open data in whatever format they had it in. The governments of the United States and the United Kingdom have been releasing their governmental data in an open format[4] and other governments are following suit. This provides another source of datasets which have their own unique challenges in classifying and subsequently finding them in that they are released in arbitrary formats with varying amounts of associated metadata. Setting up a uniform schema and vocabulary for annotating these datasets as well as providing effective search tools helps developers find these sets in order to use them for new applications [6].

There are search engines for finding RDF and other datasets, such as ordinary search engines, SWSE [11], Swoogle[5], Watson[6], and others. However, using such systems based on the Google-like search paradigm it is difficult to get an idea of the *whole* cloud of the offered datasets. Furthermore, finding suitable datasets based on different selection criteria such as topic, size, licensing, publisher, language etc. is not supported. To facilitate this, interoperable metadata about the different aspects or facets of datasets is needed, and faceted search (also called view-based search) [19, 9, 12] can be used to provide an alternative paradigm for string-based semantic search.

This paper presents DataFinland, a semantic portal for creating, publishing, and finding datasets based on metadata. In contrast to systems like CKAN[7], the LOD-oriented voiD[8] (Vocabulary of Interlinked Datasets) metadata schema is used to describe datasets with property values taken from a set of shared domain ontologies providing controlled vocabularies with clearly defined semantics. Content is annotated using a web-based annotation tool SAHA 3[9] connected to ONKI ontology services[10] [22, 21] that publish the domain ontologies. SAHA 3 has been integrated with the lightweight multifaceted search engine HAKO[11] [16], which facilitates automatically forming a faceted search and browsing application for taking in and discerning the datasets on offer. The annotation data itself is stored in RDF format, which makes combining the metadata about different datasets from different sources simple. This means that it would be possible to have several annotation projects for different sets of datasets, which could then be combined as needed for searching purposes. As a proof of concept, the system has been applied to describing the LOD cloud datasets and datasets in the Finnish Open Data Catalogue Project[12] complementing the linked open governmental datasets on a national level. The demonstration is available online[13] and the system

[4] http://www.data.gov/ and http://data.gov.uk/
[5] http://swoogle.umbc.edu/
[6] http://watson.kmi.open.ac.uk/WatsonWUI/
[7] http://www.ckan.net/
[8] http://semanticweb.org/wiki/VoiD
[9] http://www.seco.tkk.fi/services/saha/
[10] http://www.onki.fi/
[11] http://www.seco.tkk.fi/tools/hako/
[12] http://data.suomi.fi/
[13] http://demo.seco.tkk.fi/saha3sandbox/voiD/hako.shtml

received the first prize in this year's "Apps4Finland–Doing Good With Open Data" competition.

In the following we will first present the general model and tools for creating and publishing metadata about (linked) datasets, and then discuss the voiD metadata schema and ontology repository ONKI presenting a controlled vocabulary. After this, the annotation tool SAHA for distributed semantic content creation is presented along with the faceted publication engine HAKO. In conclusion, the main contributions of the paper are listed, related work discussed, directions for future research proposed.

2 Overview of the Publication Process

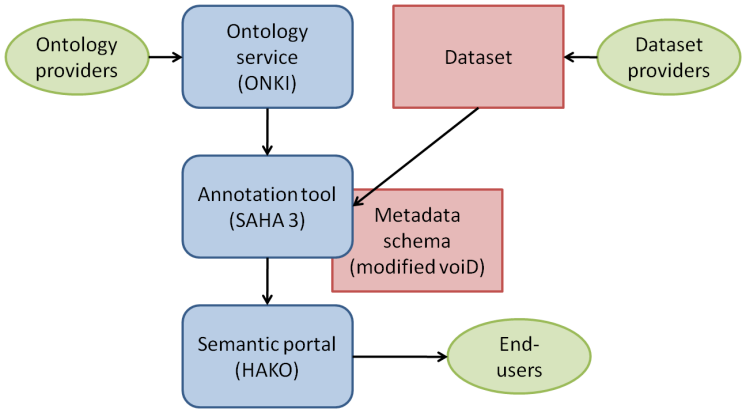

Fig. 1. The distributed process of producing and publishing metadata about (linked) datasets

Our solution for the process of producing metadata and publishing the annotated datasets is depicted in Figure 1. The process begins with the publication of a dataset. Metadata for the dataset is produced either by its original publisher or by a third party, using an annotation tool, in our case SAHA 3. A metadata schema, in our case modified voiD, is used to dictate for the distributed and independent content providers the exact nature of the metadata needed. Interoperability in annotation values is achieved through shared ontologies that are used for certain property values in the schema (e.g., subject matter and publisher resources are taken from corresponding ontologies). The ontologies are provided for the annotation tool as services, in our case by the national ONKI Ontology Service (or by SAHA itself). Finally, the metadata about the datasets is published in a semantic portal capable of using the annotations to make the data more accessible to the end-user, be that a human or a computer application. For this part the faceted search engine HAKO is used.

In the figure, we have marked the tools and resources used in our proof-of-concept system in parentheses, but the process model itself is generic.

3 Metadata and Ontologies

From a semantic viewpoint, the key ingredients of general model presented above are the metadata schema and domain ontologies/vocabularies used for filling in values in the schema. As for the metadata schema, the Vocabulary of Interlinked Datasets (voiD), an RDF vocabulary for describing linked datasets [1], seemed like a natural starting point because it addresses specifically problems of representing linked data. It was therefore chosen as a basis in our proof-of-concept system.

The basic component in voiD is a dataset, a collection of RDF triples that share a meaningful connection with each other in the form a shared topic, source or host. The different aspects of metadata that voiD collects could be classified into the following three categories or facets:

1. Descriptive metadata tells what the dataset is about. This includes properties such as the name of the dataset, the people and organizations responsible for it, as well as the general subject of the dataset. Here voiD reuses other, established vocabularies, such as `dcterms` and `foaf`. Additionally, voiD allows for the recording of statistics concerning the dataset.
2. Accessibility metadata tells how to access the dataset. This includes information on SPARQL endpoints, URI lookup as well as licensing information so that potential users of the dataset know the terms and conditions under which the dataset can be used.
3. Interlinking metadata tells how the dataset is linked to other datasets. This is done by defining a linkset, the concept of which is depicted in Figure 2. If dataset :DS1 includes relations to dataset :DS2, a subset of :DS1 of the type void:Linkset is made (:LS1) which collects all the triples that include links between the two datasets (that is, triples whose subject is a part of DS1 and whose object is a part of :DS2).

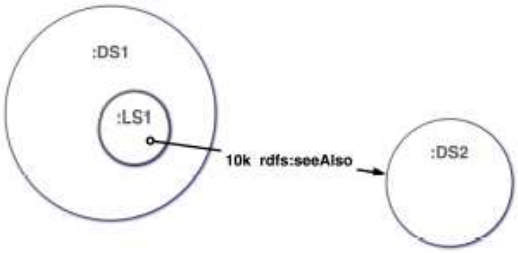

Fig. 2. Modeling interlinking of datasets in voiD [1]

3.1 Extending voiD

In order to facilitate annotating also non-linked open datasets, we made some extensions to voiD. The most important of these was a class for datasets in formats other

than RDF. This `void-addon:NonRdfDataset` is similar to the `void:Dataset` but does not have the RDF-specific properties such as SPARQL endpoint while including a proprety for describing the format of the dataset, `void-addon:format`. The addition of this class also resulted in modifications to most of the voiD properties to include `void-addon:NonRdfDataset` in their domain specifications. Another addition to the basic voiD in our system was `dcterms:language` that facilitates the multi-language applications.

4 From Annotations to Faceted Search

Since the publishing of open data is not done by any central authority, annotating the data should also be collaborative and community-driven. To this end the annotation tools should be easy to use and publishing the results of the annotations should be quick and easy.

Our solution to facilitating collaborative annotation of distributed communities is based on the SAHA 3 metadata editor and the HAKO faceted search system [16]. In addition, we use the ONKI Ontology Service [21, 22] for providing ontological concepts to annotations. These concepts, organized as hierarchical ontologies, also provide facets classifying the subject matter and some other aspects of the datasets in focus. Using ontologies instead of a free tagging system provides a controlled vocabulary with well defined meanings as well as support for multiple languages. Furthermore, the semantic relations can be used in further reasoning when the number of datasets gets very high.

4.1 SAHA 3 Metadata Editor

SAHA 3 is a metadata editor that allows for easy annotation of varied resources hiding the complexities of RDF and OWL from the end users. It is easily configurable to different schemas and supports distributed, simultaneous annotation in different locations, which is of paramount importance when using it in a community-driven environment such as Linked Open Data. It also functions in a normal web browser needing no special software to be installed.

The process of annotation itself is simple using SAHA 3. When a project has been configured, the annotator is shown the main view of a project, which gives a general overview of it. On the left side all the classes (as denoted by owl:Class) are listed, along with the count of how many instances of that class exist in the project. New instances can be created from this list as can be seen in Figure 3. The instances for any class can be browsed and filtered to find the desired ones.

The resource view, shown in Figure 4, is a basic overview of a resource. There is a resource view for each resource in the model. All the property values of the resource are listed, except those that are configured to be hidden. The data cannot be edited here - to do that the [edit] button must be pressed, which takes the user to the Annotation view.

When annotating an instance, the annotator is provided with a number of fields corresponding to the properties whose domain matches the class of the instance (see

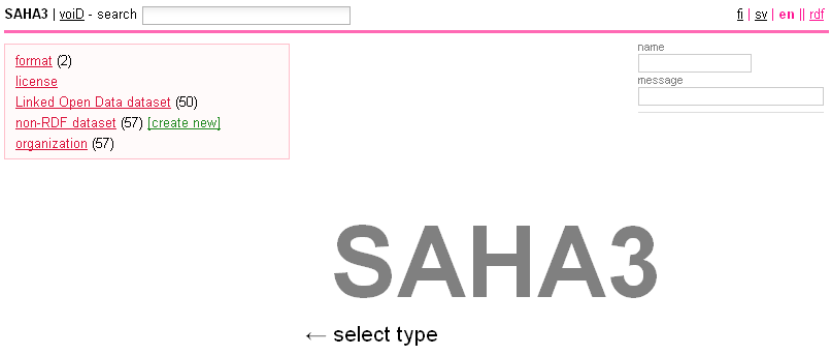

Fig. 3. Overview of a Saha project

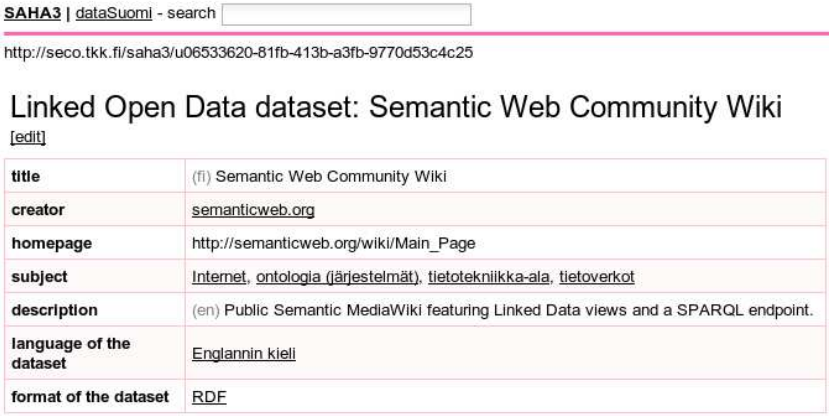

Fig. 4. Resource view of an instance in Saha

Figure 5). Depending on the range of a given property, the field takes in either free text or instances. In the latter case the instances can be either ones defined in the current project or chosen from linked ONKI ontologies. In both cases autocompletion[14][10] is used to aid the annotator.

4.2 HAKO Faceted Search Engine

HAKO is a faceted search engine that can be used to publish a SAHA 3 project as a readily usable portal. The RDF data produced in SAHA 3 is exported into HAKO, which is then configured to produce a portal matching the needs of the end user. The publisher configures the classes whose instances are to be searched and whose properties form the search facets for these instances.

The end result is a semantic portal supporting both faceted search as well as free text search, which is done as a prefix search by default. For machine use, SAHA 3 also has a SPARQL endpoint[14] which can be used to access the metadata from the outside as a service instead of accessing the HAKO portal human interface. The SPARQL interface can be used also internally in SAHA for providing semantic recommendation links between data objects on the human interface.

4.3 DataFinland

DataFinland is the name given for the whole solution of combining SAHA 3 and HAKO search portal with the extended voiD schema for creating, publishing, and finding datasets based on metadata.

When configuring SAHA 3 for voiD, the `dcterms:subject` was connected to the ONKI instance of the General Finnish Ontology (YSO)[15] with over 20,000 concepts. The property `dcterms:license` was linked to an ONKI instance featuring six Creative Commons license types, but the system also allows for the defining of other license types as new instances of a simple license class. Its properties include of a free text description of the license as well as a possible link to a webpage describing the license further. Finally, `dcterms:language` was connected to the ONKI instance of the Lingvoj[16] vocabulary listing of the languages of the world.

The SAHA 3 annotation environment for voiD (depicted in Figure 5) allows for the annotation of both RDF and non-RDF datasets as well as licenses, formats and organizations. Licenses are additional licenses that the user may want to use aside from the ready linked Creative Commons licenses. Formats are simple resources to identify the format of the dataset, e.g. PDF, MS Word Document, etc. Finally, organizations allows for a simple way of describing an organization or a person responsible for a given dataset in the form of a title, free text description and a link to a homepage or a similar information source.

HAKO was configured to search for both RDF and non-RDF datasets and to form facets based on the license, language, format and subject properties. This way the end-user can, for example, limit his/her search to cover only Linked Open datasets by choosing the RDF format. In Figure 6 the user has selected from the facets on the left RDF datasets concerning Information technology industry in the English language. Out of the nine results provided by HAKO, the user has chosen Advogato to see its metadata.

A problem of faceted search with wide-ranging datasets is that facets tend to get very large, which makes category selection more difficult. A solution to this is to use hierarchical facets. However, using the hierarchy of a thesaurus or an ontology intended originally for annotations and reasoning may not be an optimal facet for information retrieval from the end-user's perspective [20]. For example, the top levels of large ontologies with complete hierarchies can be confusing for the end-users. Our planned solution in the future is to provide the annotators with a simple tool for building hierarchies for the facets as a part of the annotation process. Another possible solution would be to use

[14] http://demo.seco.tkk.fi/saha/service/data/voiD/sparql?query={query}
[15] http://www.yso.fi/onki/yso/
[16] http://www.lingvoj.org/

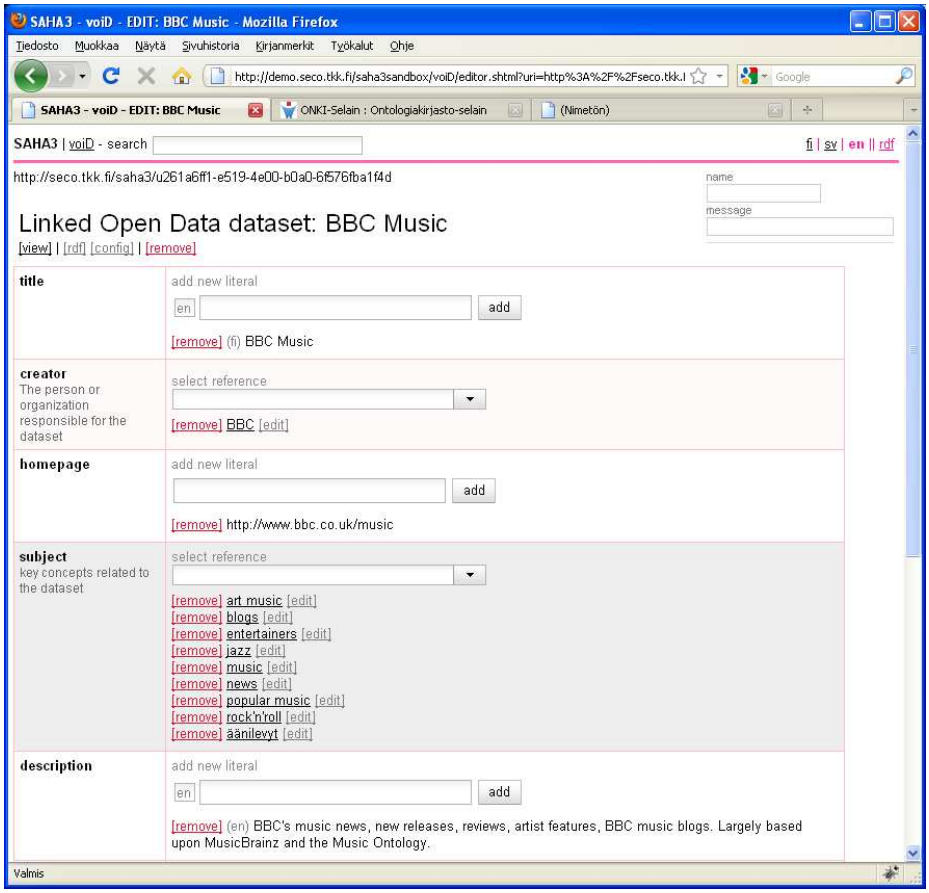

Fig. 5. SAHA 3 annotation environment

some kind of an all-inclusive classification system as the top level of the facets. There has been some discussion of a classification schema for open datasets in the community, but no clear standard has risen. In the future we plan to explore the possibility of using the Finnish Libraries' classification system that is based on Dewey Decimal Classification.

5 Discussion

5.1 Contributions

This paper presented a distributed content creation model for metadata about datasets published on the web. The model emphasizes and supports the idea that metadata should be created in an interoperable way by the actors that publish the actual content.

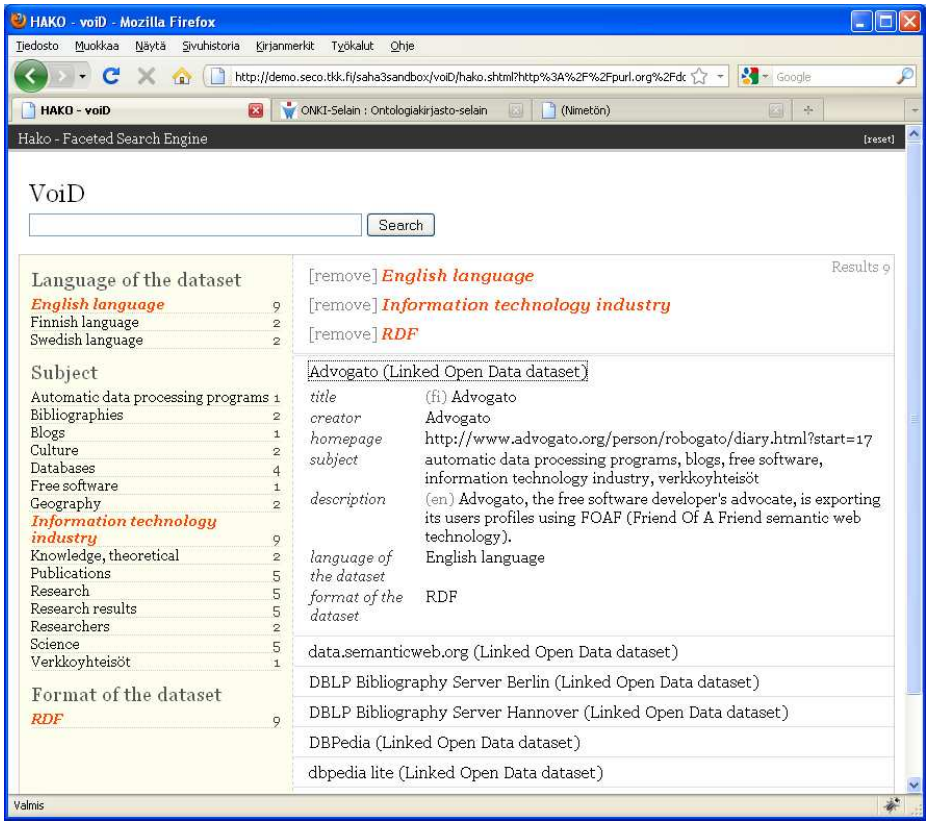

Fig. 6. HAKO faceted search portal

Making metadata interoperable afterwards is usually more difficult and costly. [13] In practice this requires support for using shared metadata schemas and domain ontologies/vocabularies, as well as a shared publication channel, a semantic portal. These facilities are provided in our model by the combination of ONKI, SAHA and HAKO tools.

One of the main challenges of any model dealing with dataset metadata is to motivate dataset publishers to also publish semantically annotated metadata about their content. Our work is driven by the hope that this social challenge can be addressed by making annotating easy by online tools (such as SAHA and ONKI), and by providing the annotators with instant feedback on how their dataset is shown in the final semantic portal (HAKO).

5.2 Related Work

There is a number of tools available for creating voiD descriptions. The voiD editor ve[17] and liftSSM[18], an XSLT script that transforms a semantic sitemap in XML to voiD RDF/XML format, but these allow building only rudimentary descriptions, which should then be added to by manually editing the RDF file.

As for datasets, there are a number of tools for finding Linked Open data. Semantic Web Search Engine[11] (SWSE) takes a free text approach allowing the user to enter a query string and returning entities from Linked Open datasets that match the query term. Searching for whole datasets is not supported.

Aside from search tools intended for human users, there is a number of search indexes intended for applications, including Sindice [18], Watson [5] and Swoogle [7]. These provide APIs supporting the discovery of RDF documents based on URIs or keywords. Sindice is intended for finding individual documents while Swoogle is used for finding ontologies. Watson allows the finding of all sorts of semantic data and features advanced filtering abilities intended for both machine and human users. However, none of these search engines are very good for exploring what sorts of datasets are available or for getting a whole picture of a given domain.

Governmental Open Data is widely published through CKAN[19] (Comprehensive Knowledge Archive Network), a registry for Open Data packages. CKAN provides support for publishing and versioning Open data packages and includes robust API support. However, the metadata about the data packages is recorded utlizing free tagging which does not support hierarchical, view-based search and does not contain semantic relation data between different tags.

Finally, concurrently to our work, an interoperability format for governmental data catalogues based on the dcat RDF vocabulary was proposed in [17]. There, the metadata schema was based on existing metadata used in the data catalogues as opposed to the LOD based voiD. Furthermore, this solution does not contain tools for editing metadata nor link to existing ontologies for use in dataset descriptions. A faceted search using Gridworks in combination with dcat was also proposed in [4].

The distributed semantic content creation and publishing approach, using shared metadata schemas, ontology services, and semantic portals for publication, has been originally developed in the semantic portals of the FinnONTO project [15].

5.3 Future Work

Our intention next is to propose the testing of the demonstrational system in the Finnish open data catalogue project. Another future application prospect is to apply the system for publishing metadata about scientific datasets for research. Additional distributed annotation-publishing projects can be opened with little extra work using the tools presented; proposals are solicited by the authors of this paper.

---

[17] http://ld2sd.deri.org/ve/

[18] http://vocab.deri.ie/void/guide#sec_4_3_Publishing_tools

[19] http://www.ckan.net/

Acknowledgements

This work was conducted as a part of the National Semantic Web Ontology project in Finland[20] (FinnONTO, 2003-2012), funded mainly by the National Technology and Innovation Agency (Tekes) and a consortium of 38 public organizations and companies. Furthermore, we would like to thank Tuomas Palonen for his annotation work on the datasets for the demonstration and Petri Kola and Antti Poikola for fruitful discussions on publishing open datasets.

References

1. Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. 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)*, 2009.
2. Chris Bizer, Richard Cyganiak, and Tom Heath. How to publish linked data on the web, 2007.
3. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.
4. Richard Cyganiak, Fadi Maali, and Vassilios Peristeras. Self-service linked government data with dcat and gridworks. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 37:1–37:3, New York, NY, USA, 2010. ACM.
5. Mathieu dÁquin and Enrico Motta. Watson, more than a semantic web search engine, 2010.
6. Makx Dekkers, Femke Polman, Robbin te Velde, and Marc de Vries. Mepsir: Measuring european public sector information resources. final report of study on exploitation of public sector information. Technical report, 2006.
7. Tim Finin, Li Ding, Rong Pan, Anupam Joshi, Pranam Kolari, Akshay Java, and Yun Peng. Swoogle: Searching for knowledge on the semantic web. In *In AAAI 05 (intelligent systems demo*, pages 1682–1683. The MIT Press, 2005.
8. Michael Hausenblas, Wolfgang Halb, Yves Raimond, and Tom Heath. What is the size of the semantic web? In *Proceedings of I-SEMANTICS '08*, 2008.
9. M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Lee. Finding the flow in web site search. *CACM*, 45(9):42–49, 2002.
10. Michiel Hildebrand, Jacco van Ossenbruggen, Alia Amin, Lora Aroyo, Jan Wielemaker, and Lynda Hardman. The design space of a configurable autocompletion component. Technical Report INS-E0708, Centrum voor Wiskunde en Informatica, Amsterdam, 2007.
11. Aidan Hogan, Andreas Harth, Jürgen Umrich, and Stefan Decker. Towards a scalable search and query engine for the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1301–1302, New York, NY, USA, 2007. ACM.
12. E. Hyvönen, S. Saarela, and K. Viljanen. Application of ontology-based techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium, May 10–12, Heraklion, Greece*. Springer–Verlag, 2004.
13. Eero Hyvönen. Preventing interoperability problems instead of solving them. *Semantic Web Journal*, 2010. accepted for pubication.
14. Eero Hyvönen and Eetu Mäkelä. Semantic autocompletion. In *Proceedings of the First Asia Semantic Web Conference (ASWC 2006), Beijing*. Springer–Verlag, 2006.

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

15. Eero Hyvönen, Kim Viljanen, Eetu Mäkelä, Tomi Kauppinen, Tuukka Ruotsalo, Onni Valkeapää, Katri Seppälä, Osma Suominen, Olli Alm, Robin Lindroos, Teppo Känsälä, Riikka Henriksson, Matias Frosterus, Jouni Tuominen, Reetta Sinkkilä, and Jussi Kurki. Elements of a national semantic web infrastructure—case study finland on the semantic web (invited paper). In *Proceedings of the First International Semantic Computing Conference (IEEE ICSC 2007), Irvine, California*. IEEE Press, September 2007.
16. Jussi Kurki and Eero Hyvï¿½en. 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/.
17. Fadi Maali, Richard Cyganiak, and Vassilios Peristeras. Enabling interoperability of government data catalogues. In Maria Wimmer, Jean-Loup Chappelet, Marijn Janssen, and Hans Scholl, editors, *Electronic Government*, volume 6228 of *Lecture Notes in Computer Science*, pages 339–350. Springer Berlin / Heidelberg, 2010.
18. Eyal Oren, Renaud Delbru, Michele Catasta, Richard Cyganiak, and Giovanni Tummarello. Sindice.com: A document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics and Ontologies*, 3, 2008.
19. A. S. Pollitt. The key role of classification and indexing in view-based searching. Technical report, University of Huddersfield, UK, 1998. http://www.ifla.org/IV/ifla63/63polst.pdf.
20. Osma Suominen, Kim Viljanen, and Eero Hyvönen. User-centric faceted search for semantic portals, 2007.
21. Jouni Tuominen, Matias Frosterus, Kim Viljanen, and Eero Hyvï¿½en. Onki skos server for publishing and utilizing skos vocabularies and ontologies as services. In *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, May 31 - June 4 2009. Springer-Verlag.
22. 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.