PROVIDING GEOGRAPHIC DATASETS AS LINKED DATA IN SDI

E. Hietanen, L. Lehto, P. Latvala

Dept. of Geoinformatics and Cartography, Finnish Geospatial Research Institute (FGI), National Land Survey of Finland, Geodeetinrinne 2, FI-02430 Masala, Finland – (eero.hietanen, lassi.lehto, pekka.latvala)@nls.fi

ThS13

KEY WORDS: URI, Geographic Dataset, Linked Data, GeoSPARQL, WFS

ABSTRACT:

In this study, a prototype service to provide data from Web Feature Service (WFS) as linked data is implemented. At first, persistent and unique Uniform Resource Identifiers (URI) are created to all spatial objects in the dataset. The objects are available from those URIs in Resource Description Framework (RDF) data format. Next, a Web Ontology Language (OWL) ontology is created to describe the dataset information content using the Open Geospatial Consortium's (OGC) GeoSPARQL vocabulary. The existing data model is modified in order to take into account the linked data principles. The implemented service produces an HTTP response dynamically. The data for the response is first fetched from existing WFS. Then the Geographic Markup Language (GML) format output of the WFS is transformed on-the-fly to the RDF format. Content Negotiation is used to serve the data in different RDF serialization formats. This solution facilitates the use of a dataset in different applications without replicating the whole dataset. In addition, individual spatial objects in the dataset can be referred with URIs. Furthermore, the needed information content of the objects can be easily extracted from the RDF serializations available from those URIs.

A solution for linking data objects to the dataset URI is also introduced by using the Vocabulary of Interlinked Datasets (VoID). The dataset is divided to the subsets and each subset is given its persistent and unique URI. This enables the whole dataset to be explored with a web browser and all individual objects to be indexed by search engines.

1. INTRODUCTION

Many initiatives to facilitate the integration of spatial data to other available datasets have been introduced. One of the ideas, which lead to better integration, is to provide spatial data as linked data in Semantic Web. Some prominent recent initiatives for geographic linked data are the collaboration of Open Geospatial Consortium (OGC) and World Wide Web Consortium (W3C) to improve interoperability and integration of geospatial information with data on the Web (W3C 2015) and OGC's GeoSPARQL standard, which provides spatial operations for SPARQL and a Resource Description Framework (RDF) vocabulary to describe geometries and topology (OGC 2012).

Different solutions have been introduced to provide existing spatial data as linked data. Schade and Cox (2010) suggest that Geographic Markup Language (GML) can be easily transformed to RDF, because of their similar information structure and the available linking mechanism used in GML. Hereby, automatic tools for transformation have been introduced (e.g. Patroumpas al. 2014). Generic transformation from GML to RDF will lose the domain specific semantics defined in UML (Unified Modeling Language) model. Thus, these tools may provide a possibility for user to define the mappings from the GML schema to existing Web Ontology Language (OWL) vocabularies (e.g. Van den Brink et al. 2014).

In this study, a prototype service to provide a geographic names dataset (Leskinen 2015) as linked data is implemented. At first, persistent and unique URIs are given to all of the dataset objects according to the Public Administration Recommendations in Finland (JUHTA 2015). The OWL ontology is created according to the improved UML/GML (Unified Modeling Language) data model of the original dataset. Then a custom-made on-the-fly transformation process from GML to RDF is implemented to provide the spatial objects from the individual URIs. Existing OGC Web Feature Service (WFS) is used as a source.

The purpose of the study is to find out if such an on-the-fly transformation can be implemented and to find working solution for creating an OWL ontology from the UML/GML data model. In addition, the ways to link the objects to the dataset and to divide the dataset into the subsets in order to make the whole dataset browsable with those links are sought.

2. RELATED WORK

Tschirner et al. (2011) have introduced a SPARQL service, which enables using the INSPIRE Directive (EC 2007) compliant WFS as a source. The idea is to transform SPARQL queries to WFS queries. Then a mapping between GML data model and OWL ontology is used to transform the WFS query results into RDF. To create the mapping between the data models, a general rules for building an OWL ontology according to a GML model are introduced. These rules are used to create the geographic names ontology in this study. Although the used data source is also WFS, there is no SPARQL endpoint implemented in this study. Instead, the implemented prototype provides spatial objects in the RDF format directly from the URIs of those objects.

Jones et al. (2014) did it the other way around. They designed an adapter to provide linked open data of the web from the

WFS. The idea of the adapter is that WFS requests are translated to SPARQL queries and the query results are then transformed to WFS XML documents, which are returned to the client. This allows the GIS applications with a WFS support to access the datasets of geographic linked data.

3. DESIGNING THE ONTOLOGY AND URIS

3.1 Expressing the Geometry

The reuse of the ontologies has an important part in the interoperability of datasets. There are many vocabularies available to model spatial information in RDF model. Since OGC's standards are widely used in the Spatial Data Infrastructures (SDI), the choice in this study is to use OGC's GeoSPARQL (OGC 2012) vocabulary. The GeoSPARQL vocabulary supports a wide range of topological relations (e.g. OGC Simple Feature and Egenhofer) and different coordinate reference systems. Also, the division of the feature and geometry in the vocabulary conforms to the ISO 19109 General Feature Model.

The designed class hierarchy of the objects is based on place types defined in the original geographic names dataset (Leskinen 2015). Another option would have been to define the hierarchy according to the area division: provinces, regions and municipalities, but these relations can also be expressed with a GeoSPARQL vocabulary.

3.2 Creating the Ontology

When creating an ontology, it is important to make division between canonical and harmonized data models and data models, which are describing some phenomenon more freely (Cox 2013). In the latter case, it would be practical to follow OWL's design paradigms to create an ontology as expressive as possible. The original geographic names dataset used in this study has a custom-made UML/GML data model. The original model is rather flat and the hierarchy and association relations are mostly hidden in the enumerations. The knowledge available, e.g. metadata documents and the hierarchy of geographic names place type division, is used to improve the UML/GML data model. The goal is to take advantage of OWL's comprehensive features such as hierarchy description and built-in association relations.

After the existing UML/GML data model is modified, the rules introduced by Tschirner et al. (2011) are applied to create the ontology (Figure 1). With this kind of solution, feedback to improve the original data model is also obtained, which might lead to better interoperability between the SDI and the Semantic Web.

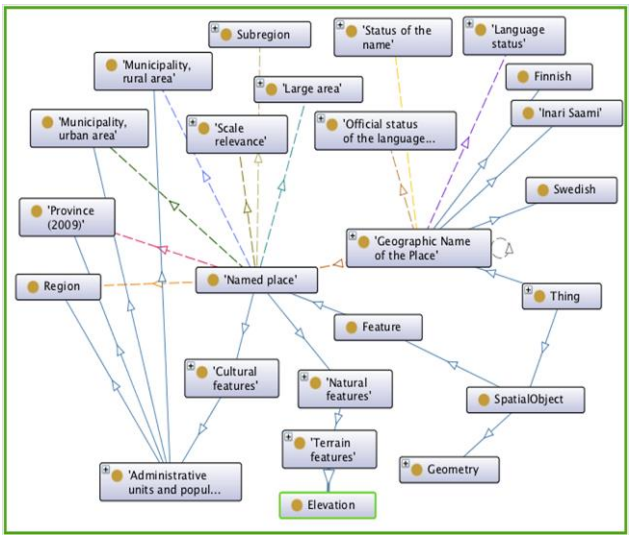

Figure 1. Part of the created OWL ontology. Solid lines depict the class hierarchy. Visualization made with Protégé by Stanford Center for Biomedical Informatics Research

For all the spatial objects, persistent and unique URIs are given according to the public administrative recommendations in Finland. A public redirection service Paikkatiedot.fi for spatial datasets provided by the National Land Survey of Finland (NLSF) is used as a URI domain for spatial objects. All the spatial object URIs contain the */so/* path component. URIs have also been given to the definitions. Those URIs contain the */def/* path component. The HTTP requests to the objects or definitions URIs are redirected to the domain of linked data service provider. In the service domain a content negotiation is used to provide the RDF data in the desired format (Figure 2).

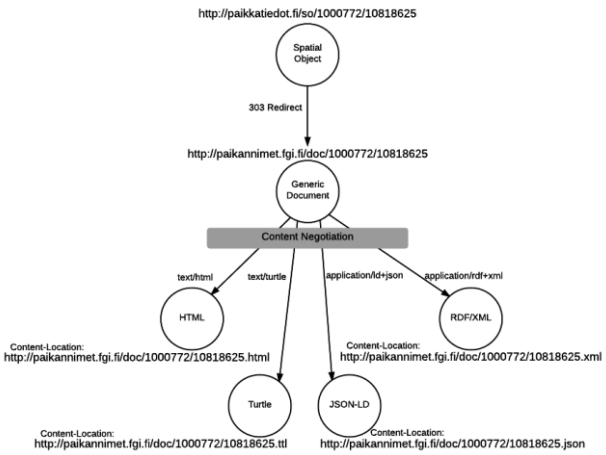

Figure 2. Redirections and the Content Negotiation. Modified from the original figure by W3C (2008), Copyright © 2008 W3C® (MIT, ERCIM, Keio), All Rights Reserved. W3C liability, trademark and document use rules apply.

A persistent and unique URI is also given to the whole dataset. An example of a URI of the spatial object is *http://paikkatiedot.fi/so/1000772/10818625* and the URI for the whole dataset is *http://paikkatiedot.fi/so/1000772/*. The Vocabulary of Interlinked Dataset (VoID, Cyganiak et al. 2011) is used for linking the spatial objects to the dataset. There is a predicate *void:inDataset* in VoID to express that relation. But how to express which objects belongs to the dataset? There is no "the dataset has these objects" kind of relation in VoID and

in addition, there are over 800,000 spatial objects in the dataset. Returning these 800,000 objects when requested the dataset URI, would not be very practical.

The used solution is to divide the dataset into subsets "small enough" by using the *void:subset* property. "Small enough" is defined in this context so that all the objects in a subset should be browsable as an alphabetically ordered list on a web page. The goal is achieved when the objects of the geographic names dataset are divided on the first level by location to approx. 300 municipalities and on the second level according to the approx. 50 place types. Thus, there are less than 3,000 places in the biggest second level subset in this case.

The dataset URIs are defined by adding the municipal code (e.g. 148), the place type code (e.g. 435) or both to the dataset number. For example, the URI for the subset of all the rapids in the municipality of Inari is *http://paikkatiedot.fi/so/1000772:148435/*. Request to the second level subset of geographic names dataset returns the information of the subset as well as the label information and URIs of the objects contained in the subset. This kind of solution makes it possible for the users to browse the whole contents of the dataset with a web browser by following the links.

4. PROTOTYPE IMPLEMENTATION

The service use diagram (Figure 3) depicts the general use flow of the implemented prototype service. The client sends an HTTP request to the *paikkatiedot.fi* domain. The request is redirected to Geographic Names as Linked Data Service. The service parses the URI and makes a WFS Query according to the URI, sends the query to NLSF WFS, creates RDF according to the response and returns the RDF data to the client in desired serialization format. Using WFS as data source guarantees that the data provided by implemented service is up-to-date, because there is no need to replicate the original dataset to a separate triple store.

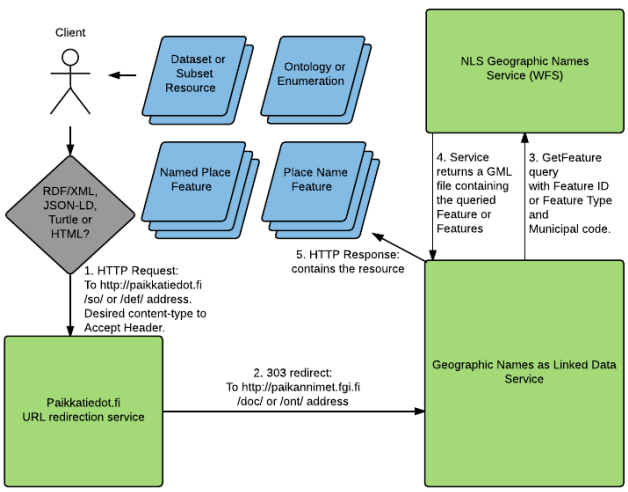

Figure 3 Geographic Names as Linked Data Service use diagram

The Geographic Names as Linked Data Service is implemented with Python programming language on the Ubuntu Server platform. It uses Django Web Framework (www.djangoproject.com) and RDFLib for Python (https://github.com/RDFLib/rdflib). The actual process is divided into two separate processes: a preliminary process and a real time process. The idea of the preliminary process is to reduce the number of HTTP requests to the WFS in the real time process. Data about area instances (municipalities, regions and provinces), named place classes, name of the place classes, enumerations and class hierarchy is stored in RDF format. Preliminary process needs to be re-executed only if there are changes in enumerations of the original data model or in municipality, region and province objects used for areal division expressed with GeoSPARQL topology relations. If there are changes in the original data model (XML Schemas) or in the place type hierarchy, changes in code have to be made.

The real time process is executed with every HTTP request to an object, subset, dataset or definition URI. Only one WFS query to NLSF WFS is made. The data content of the WFS query is transformed into RDF and combined with RDF data created during preliminary process. The returned serialization format is decided according to the *Accept* header of the original HTTP request.

The service provides the data in different serialization formats: RDF/XML, Turtle, JSON-LD and HTML. The information contents of RDF/XML, Turtle and JSON-LD are consistent. The HTML content is meant for human readers and to be viewed with web browsers. In addition to human readable content, RDF content in JSON-LD format is added inside the script tags in HTML using the schema.org vocabulary and its *Place* class (http://schema.org/Place), which allows Google to understand the content (Google 2015).

Browsing the whole content of dataset is possible not only for humans, but also for search engines, which can index all the individual objects in the dataset. Thus, users can find the information content of specific objects or the whole dataset by using Web search engines. This may encourage others to use URIs to reference these objects for example from another datasets, news articles or reports.

5. FUTURE WORK

Providing the individual dataset objects from the URIs of those objects is only a part of an integrated linked data service. Thus, SPARQL endpoint, other search capabilities and the possibility to download the whole data content in RDF format would be needed for making this a comprehensive solution.

The solution presented in the paper is custom-made and it cannot be directly applied to other datasets without additional programming work. The next step is to develop a general solution with a possibility to configure the service for different data models and datasets.

6. CONCLUSION

A prototype of a Geographic Names as Linked Data Service is implemented in this study. Using the WFS as a data source guarantees the data provided by the implemented service is concurrent with the original dataset.

The improving of the original data model with comprehensive properties of OWL and the providing of the data and its

metadata on the Web as linked data improve the usability and the accessibility of the dataset.

The introduced solution enables all the individual data objects to be retrieved from the persistent and unique URIs of these objects. This solution facilitates the use of a dataset in different applications without replicating the whole dataset. In addition, individual spatial objects in the dataset can be referred with URIs. Furthermore, the needed information content of the objects can be easily extracted from the RDF serializations available from those URIs.

The created linking between the dataset, its subsets and the data objects makes it possible for humans to browse the whole dataset with a web browser. Search engines can also index all the individual objects of the dataset.

REFERENCES

Cyganiak, R., Zhao, J., Alexander, K., Hausenblas, M. 2011. Vocabulary of Interlinked Datasets (VoID). Web page. URL: http://vocab.deri.ie/void (Accessed: 15 Mar 16)

EC 2007. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). URL: http://inspire.ec.europa.eu/

Google 2015. About schema.org. Web page. URL: https://developers.google.com/structured-data/schema-org (Accessed: 17 Mar 16)

Jones, J., Kuhn, W., Keßler. C., Scheider, S. 2014. Making the web of data available via web feature services. *Connecting a Digital Europe Through Location and Place.* pp. 341–361. Springer International Publishing. URL: http://carsten.io/jones-kuhn-kessler-scheider-agile2014.pdf.

JUHTA – Advisory Committee on Information Management in Public Administration 2015. JHS 193 Unique identifiers of the geographic information. Public Administration Recommendations in Finland. In Finnish. URL: http://www.jhs-suositukset.fi/web/guest/jhs/recommendations/193.

Leskinen, T. 2015. A Data Repository for Named Places and Their Standardised Names Integrated With the Production of National Map Series. *7th International Cartographic Conference proceedings*. 15 p. URL: http://icaci.org/files /documents/ICC_proceedings/ICC2015/papers/21/516.html.

OGC, Open Geospatial Consortium 2012. OGC GeoSPARQL - A geographic query language for RDF data. Version 1.0. OGC Implementation Standard. Edit. Matthew Perry and John Herring. OGC 11-052r4. Standard, 75 p. URL: http://www.opengis.net/doc/IS/geosparql/1.0.

Schade, S., Cox, s. 2010. Linked Data in SDI or How GML is not about Trees. *Proceedings of the 13th AGILE International Conference on Geographic Information Science - Geospatial Thinking*. 10 p. URL: https://agile-online.org/Conference_Paper /CDs/agile_2010/ShortPapers_PDF/73_DOC.pdf.

Tschirner, S., Scherp, A., Staab, S. 2011. Semantic access to INSPIRE. *Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web 798*. pp. 75–87. CEUR-WS.org. URL: http://ceur-ws.org/Vol-798/paper7.pdf.

Patroumpas, K., Alexakis, M., Giannopoulos, G., Athanasiou, S. 2014. TripleGeo: an ETL Tool for Transforming Geospatial Data into RDF Triples. *EDBT/ICDT Workshops.* p. 275-278. URL: http://www.dbnet.ece.ntua.gr/pubs/uploads/TR-2014-2.pdf

Van den Brink, L., Janssen, P., Quak, W., Stoter, J.E. 2014. Linking spatial data: semi-automated conversion of geo-information models and GML data to RDF, *International Journal of Spatial Data Infrastructures Research*, Vol.9, 59-85, URL: http://repository.tudelft.nl/view/ir/uuid:ae1167c7-d246-4255-9419-800ac8d805e8/

W3C 2008. Cool URIs for the Semantic Web. W3C Interest Group Note 03 December 2008. Ed. Leo Sauermann and Richard Cyganiak. Web page. URL: http://www.w3.org/TR/2008/NOTE-cooluris-20081203/ (Accessed: 15 Mar. 16).

W3C 2015. Spatial Data on the Web Working Group Charter. Version 1.10. Ed. Phil Archer. Web page. URL: http://www.w3.org/2015/spatial/charter (Accessed: 10 Mar. 16).