Linked Data for the Norwegian State of Estate Reporting Service

Ling Shi[1], Bjorg Elsa Pettersen[1], Dina Sukhobok[2], Nikolay Nikolov[2] and Dumitru Roman[2]

[1] Statsbygg, Pb. 8106 Dep, 0032 Oslo, Norway
{Ling.Shi,BjorgElsa.Pettersen}@statsbygg.no
[2] SINTEF, Forskningsveien 1a, 0373 Oslo, Norway
{dina.sukhobok,nikolay.nikolov,dumitru.roman}@sintef.no

Abstract. The Norwegian State of Estate (SoE) report includes information about all Norwegian state-owned properties and buildings in the public sector and aims to assist government decision makers to allocate resources more effectively. A Linked Data based approach is presented here to increase the transparency in the government administration, improve the report generating process and also the report quality. Cross-domain government data originated from the business entity register, the cadastral system, the building accessibility register and the old SoE report are acquired, prepared, cleaned, transformed to Linked Data format and published. The source datasets are then integrated, augmented and interlinked before the results are published as a SPARQL endpoint, used for data visualization and report generation.

Keywords: Linked Data, data integration, government data

1 Introduction

The SoE report issued by Statsbygg[3] on behalf of the Ministry of Local Government and Modernisation (KMD) is a governmental white paper, providing a complete list of state-owned properties and buildings in the Norwegian public sector. The report has been produced every four years as a result of manual collection of information from multiple sources. The data collection and quality control process has historically been resource demanding and error prone and the result was static and did not reflect the changes after the report was published. This paper describes the publishing and integration of several cross-domain government datasets related to state-owned real estates as Linked Open Data. Sharing the state-owned properties related data in a Linked Data format enables data reuse, opens up possibilities for using the data in innovative ways, and helps to increase transparency in the government administration[4].

In addition, we demonstrate a Web-based application for registration and reporting state-owned properties in Norway. This represents a major improvement compared to the tedious, manual collection of property data.

[3] http://www.statsbygg.no/Om-Statsbygg/About-Statsbygg/

2 Approach and Implementation

Data Sources. Cross-domain government data from different open and proprietary sources as listed below are involved in the Linked Data generation process.

- The central government organization dataset (a subset of data from the Norwegian Business Entity Register administrated by the Brønnøysund Register Centre[4]);
- The cadastral dataset (a subset of data from the Norwegian Cadastral System[5] administrated by the Norwegian Mapping Authority);
- The building accessibility dataset from the Building Accessibility Register[6] administrated by Statsbygg;
- The previous SoE report[7] dataset administrated by Statsbygg;
- The municipality boundaries dataset[8] administrated by the Norwegian Mapping Authority.

The non-geospatial datasets are prepared by dataset providers in tabular format and the geospatial datasets are provided as shape files.

Though the source data comes from the most authoritative sources in the respective domains, it is not always complete and accurate. Data inconsistency between source systems is one of the main challenges in the integration process. Missing values and keys also represent significant barriers for the integration process. Such issues are addressed in the Linked Data generation process.

Ontology description. Property related data are rich in attributes, with both spatial and temporal characteristics. A common ontology model provides the necessary semantic description of the data. We reuse standard and established ontologies such as DBpedia-owl, DUL, GeoSPARQL and schema.org to represent the data. For example, *schema:leiCode*, *schema:legalName*, *dbpedia-owl:type*, *schema:foundingDate* and *schema:parentOrgnization* are used to model organization number, name, type, founding date and parent organization, respectively, in the central government organization dataset.

In addition, we developed the proDataMarket[9][3] ontology to model, among others, the cadastral and building accessibility domain based on existing ontologies and standards. For the cadastral domain, the proDataMarket ontology reuses the Land Administration Domain Model (LADM) defined in ISO 19152:2012[10] standard and cadastral parcel concept specified by the European Union's INSPIRE data specifications[11].

[4] `https://www.brreg.no/home/`
[5] `http://www.kartverket.no/en/Land-Registry-and-Cadestre/`
[6] `https://byggforalle.no/uu/sok.html?&locale=en`
[7] `https://www.regjeringen.no/contentassets/f4346335264c4f8495bc559482428908/no/sved/stateigedom.pdf`
[8] `http://data.kartverket.no/download/content/geodataprodukter`
[9] `http://vocabs.datagraft.net/proDataMarket/`
[10] `http://www.iso.org/iso/catalogue_detail.htm?csnumber=51206`
[11] `http://inspire.ec.europa.eu/data-model/approved/r4618-ir/html/`

Linked Data Generation and Publication. The publication of SoE data as Linked Data was performed with the help of DataGraft [1,2][12]– a cloud-based platform for data cleaning and Linked Data generation. DataGraft facilitates interactive data cleaning and transformation, mapping data to Linked Data ontologies, generating a semantic RDF graph and provisioning data both as RDF dump and as a SPARQL endpoint.

Data cleaning and preparation activities for SoE data included assigning valid cadastral parcel identifiers, unifying null values for attributes with null values, changing decimal formatting and geospatial data conversion. After the source data files were cleaned, they were mapped to the above mentioned ontology and published in DataGraft. Next, the data augmentation was performed with the help of SPARQL CONSTRUCT queries executed on the published data, thus making it possible to infer new data based on known business rules. One example of such a business rule states that *a building is owned by the owner or lessor of a cadastral parcel where the building is built upon.* This helped to select out the state-owned properties and buildings, calculate the area of real estates, infer the building ownership based on the information about owner or lessor of the belonging cadastral parcel, etc. In addition, the published data is interlinked with several central LOD datasets (DBpedia[13], GeoNames[14] and Lenka.no[15]) in order to increase its re-usability to support queries on cross-domain distributed datasets.

The result of the data augmentation and interlinking process is published on DataGraft and is available through a SPARQL endpoint[16] under the Norwegian License of Open Data[17] (NLOD). SPARQL queries can be run on the SPARQL endpoint to assess the data quality of the SoE report dataset.

3 Demonstration

During the demonstration, we will present the process of generating Linked Data from the SoE report using DataGraft, and a web-based application for registration and reporting state-owned properties in Norway. The demo scenario will cover uploading raw SoE data to the DataGraft platform, data transformation and publication as a Linked Data, and the the Web-based application. The application allows users to configure visualizations and browse state-owned properties and relevant data on the map created in CartoDB[18] (see Figure 1). The state-owned properties absent in the previous SoE report are explicitly marked on the

[12] `https://datagraft.io/`
[13] `http://wiki.dbpedia.org/wiktionary-rdf-extraction`
[14] `http://www.geonames.org/`
[15] `http://data.lenka.no/`
[16] `https://datagraft.io/prodatamarket_publisher/sparql_endpoints/norwegian-state-of-estate-report-04693e1f-4060-48c1-8ab9-888a6c95f6d6` - SPARQL querying at this endpoint currently works only in Chrome.
[17] `https://data.norge.no/nlod/en/1.0`
[18] `https://carto.com/`

map (using pink colour), helping users to identify data quality issues such as inconsistencies or missing registrations in the source systems.

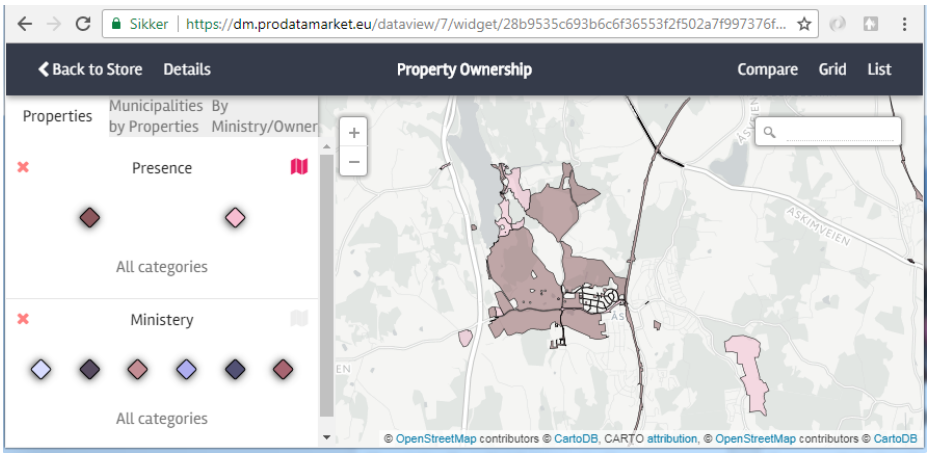

Fig. 1. Visualization of state-owned properties data.

Acknowledgements This work is partly funded by the EC H2020 project proData-Market (Grant number: 644497).

References

1. Roman, D., Dimitrov, M., Nikolov, N., Putlier, A., Sukhobok, D., Elvester, B., Berre, A. J., Ye, X., Simov, A. & Petkov, Y. DataGraft: Simplifying Open Data Publishing. ESWC (Satellite Events) 2016: 101-106.
2. Roman, D., Nikolov, N., Putlier, A., Sukhobok, D., Elvester, B., Berre, A. J., Ye, X., Dimitrov, M., Simov, A., Zarev, M., Moynihan, R., Roberts, B., Berlocher, I., Kim, S., Lee, T., Smith, A., & Heath, T. DataGraft: One-stop-shop for open data management. To appear in the Semantic Web Journal (SWJ) Interoperability, Usability, Applicability (published and printed by IOS Press, ISSN: 1570-0844), 2017, DOI: 10.3233/SW-170263.
3. Shi, L., Nikolov, N., Sukhobok, D., Tarasova, T., & Roman, D. The proDataMarket Ontology for Publishing and Integrating Cross-domain Real Property Data. To appear in the journal "Territorio Italia. Land Administration, Cadastre and Real Estate". n.2/2017.
4. Shi, L., Pettersen, B. E., Østhassel, I., Nikolov, N., Khorramhonarnama, A., Berre, A. J. and Roman, D. (2015, August). Norwegian State of Estate: A Reporting Service for the State-Owned Properties in Norway. In International Symposium on Rules and Rule Markup Languages for the Semantic Web (pp. 456-464). Springer International Publishing.