Final version available at Springer : http://dx.doi.org/10.1007/978-3-319-69459-7_30

Norwegian State of Estate Report as Linked Open Data

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

[1] Statsbygg, Pb. 8106 Dep, 0032 Oslo, Norway
ling.shi@statsbygg.no
[2] SINTEF, Pb. 124 Blindern, 0314 Oslo, Norway
{dina.sukhobok,nikolay.nikolov,dumitru.roman}@sintef.no

Abstract. This paper presents the Norwegian State of Estate (SoE) dataset containing data about real estates owned by the central government in Norway. The dataset is produced by integrating cross-domain government datasets including data from sources such as the Norwegian business entity register, cadastral system, building accessibility register and the previous SoE report. The dataset is made available as Linked Data. The Linked Data generation process includes data acquisition, cleaning, transformation, annotation, publishing, augmentation and interlinking the annotated data as well as quality assessment of the interlinked datasets. The dataset is published under the Norwegian License for Open Government Data (NLOD) and serves as a reference point for applications using data on central government real estates, such as generation of the SoE report, searching properties suitable for asylum reception centres, risk assessment for state-owned buildings or a public building application for visitors.

Keywords: State-owned real estates, Linked Data, Open government data, RDF.

1 Introduction

One significant part of public spending is on buildings and properties needed by public administrations. A State of Estate (SoE) report[1] – a report containing integrated data on state-owned real estates[2] can help the governments use them more effectively.[3] In Norway, such a report is published as an attachment[4] to the proposed parliamentary resolution No.1 every four years by Statsbygg[5] on behalf of the Ministry of Local Government and Modernization[6]. The data collection and quality control pro-

[1] An example of such a report from the UK government can be found at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/200448/SOFTE2012_final.pdf

[2] Real estates can also be called real properties, properties or cadastral parcels if the properties are registered at the national cadastral system.

[3] https://www.theguardian.com/news/datablog/2013/may/21/downsizing-government-estate

[4] https://www.regjeringen.no/contentassets/f4346335264c4f8495bc559482428908/no/sved/stateigedom.pdf

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

[6] https://www.regjeringen.no/en/dep/kmd/id504/

cess has historically been resource demanding and error prone and the result was static and did not reflect the changes after the report was published. Though the report is available online as a PDF file, the data from the report is not easily reusable because of the data format, quality, and lack of semantic descriptions of the complex real property domain. A State of Estate (SoE) business case was introduced in [1] to carry out the reporting task in a more effective way by publishing and integrating government data from both open and proprietary sources: the Norwegian business entity register, cadastral system, building accessibility register and the previous SoE report. Sharing the SoE dataset in a Linked Data format enables data reuse, opens up possibilities for using the SoE data in innovative ways, and helps increase transparency in the government administration.

Our contribution in this paper is the SoE dataset together with the publication process of state-owned real estates in Norway as Linked Open Data – a result of publishing and integrating several cross-domain government datasets in RDF[7]. Complex queries can be run on multiple interlinked source datasets to generate lists of inconsistencies between them. The lists are used to improve the data quality in the source systems and, afterwards, the data quality in the resulting SoE dataset is also improved when the source datasets are updated and republished. Publishing the resulting SoE dataset as Linked Data avoids manual data collection, simplifies the process of SoE report generation, and also promotes innovative services such as risk assessment of state-owned buildings by integrating the dataset with natural hazards datasets.

The structure of the paper is as follows. Section 2 introduces the source datasets. The SoE Linked Data generation process is described in Section 3. Section 4 provides a description of the resulting SoE dataset. Section 5 presents the usage examples of the published dataset. Related work is discussed in Section 6. Section 7 summarizes the paper and outlines further work.

2 Source datasets

This section presents the details on the source datasets and the major challenges on working with and processing the datasets. The following datasets from different sources serve as input for the generation of the SoE dataset:

- The central government organization dataset – a subset of data from the Norwegian Business Entity Register administrated by the Brønnøysund Register Centre[8];
- The cadastral datasets – a subset of data from the Norwegian Cadastral System[9] administrated by the Norwegian Mapping Authority;
- The building accessibility dataset from the Building Accessibility Register[10] administrated by Statsbygg;
- The previous SoE dataset administrated by Statsbygg;

[7] https://www.w3.org/RDF/

[8] https://www.brreg.no/home/

[9] http://www.kartverket.no/en/Land-Registry-and-Cadestre/

[10] https://byggforalle.no/uu/sok.html?&locale=en

- The municipality boundaries dataset administrated by the Norwegian Mapping Authority.

The non-geospatial datasets are prepared by dataset providers in tabular format and the geospatial datasets as shape files[11]. Both tabular format and shape files are supported by the RDF conversion tool used for the generation of the SoE data.

2.1 The central government organizations dataset

The Norwegian Business Entity Register provides a complete dataset that covers all the public and private business entities in Norway. The dataset is assigned the Norwegian License for Open Government Data (NLOD)[12] and available through a Web API on the Brønnøysund Register Centre's website[13] and the Norwegian government's open data sharing platform[14].

The central government organizations dataset is a subset of the complete list of organizations and, because of the structure of the data, there is no single attribute that can be used to extract all central government organizations. The top level central government organizations (e.g., the parliament, ministries) can be identified using a filter on organizational format that equals to "STAT". The subordinate organizations need to be extracted for each of the top level organizations by using the parent organization attribute. There exist some organizations that are not covered by the above iterative method and the organizations are added from a manual exception list which is a result of domain experts' evaluation of organizations from the previous SoE report compared to the current organization list. Moreover, the central government organizations dataset is subject to changes, though the change frequency is not often. Organizational change after a general election, fusions between organizations and privatization of government organizations and other reasons can have effect on the ownership relationship, which causes further difficulties when reconciling the data.

2.2 The cadastral datasets

The Norwegian cadastral data (including land ownership) are partially open and available as a map-based Web application[15] that presents detailed information of only one cadastral parcel at a time. This distribution of the data does not support programmatic access to cadastral data for all the state-owned real estates and it is not suitable for the purposes of the SoE report generation. The cadastral data are also available through a Web service or a database dump under subscription-based licensing. Thereby, the cadastral datasets suitable for programmatic access are essentially proprietary

[11] https://en.wikipedia.org/wiki/Shapefile
[12] https://data.norge.no/nlod/en/1.0
[13] http://data.brreg.no/oppslag/enhetsregisteret/enheter.xhtml
[14] http://hotell.difi.no/?dataset=brreg/enhetsregisteret
[15] http://seeiendom.no

and closed for public access. The current data export approach is based on a cadastre database dump and returns four sub-datasets:

- The cadastral parcel ownership dataset for state-owned or state-leased cadastral parcels;
- The cadastral parcel geospatial dataset;
- The building dataset of buildings built on state-owned or state-leased cadastral parcels;
- The building geospatial dataset.

Table 1 contains sample records from the cadastral parcel ownership dataset. The organization number (*Org. No.*) is an attribute shared by the central government organization dataset. The *Municipality* is an attribute shared by the municipality boundaries dataset. The *Cadastral ID* is an attribute shared by the other cadastral sub-datasets.

Table 1. Example records from the cadastral parcel ownership dataset

NAME	ROLE	Cadastral ID	Org. No.	Municipality	Date From	Date To
STATENS VEGVESEN	H-HJEMMELSHAVER	0214/107/402/0	971032081	0214	22.10.2014	01.01.1753
STATENS VEGVESEN	H-HJEMMELSHAVER	0214/112/2/0	971032081	0214	23.02.2006	01.01.1753
STATENS VEGVESEN	H-HJEMMELSHAVER	0214/121/9/0	971032081	0214	23.02.2006	01.01.1753
NORSK INSTITUTT FOR SKOG OG LANDSK	F - FESTER	0214/42/1/78	970167641	0214	05.02.1953	01.01.1753

The four cadastral sub-datasets are subject to change due to ownership changes and other kinds of changes related to cadastral parcels and buildings.

2.3 Building accessibility dataset

The Building Accessibility Register covers many aspects of accessibility[16] of public buildings at different levels of a construction: building, floor and room. The building accessibility data is open and available through a Web application[17] which only returns the accessibility data of one building by each search. The dataset is a simplified subset of the building accessibility data and provides primarily five indicators at the

[16] https://en.wikipedia.org/wiki/Accessibility

[17] https://byggforalle.no/uu/sok.html?&locale=en

building level chosen by building accessibility experts. These indicators are *hasStep-FreeMainAccess*, *hasStepFreeSideAccess*, *hasElevator*, *hasHandicapToilet* and *hasHandicapParking*. The dataset is generated through an API-based export from the building accessibility database.

2.4 The previous State of Estate report dataset

The previous State of Estate report dataset was dated 2013-2014 and is open and available as an online PDF file with limited possibility of data manipulation and further processing. The data are also stored in a proprietary relational system at Statsbygg, the access to which was granted for the purposes of the SoE report. Only Statsbygg's properties and buildings are updated after the report was generated and Statsbygg administrates less than 40% of the state-owned buildings. Therefore, this dataset is only used as an alternative reference data source for quality check in this business case as exemplified in Section 3.6. No further updates need to be provided and supported for this dataset.

2.5 The municipality boundaries dataset

The Municipality Boundaries geospatial dataset is one of the open datasets[18] provided by Norwegian Mapping Authority and is downloadable[19] in SOSI[20] format. The downloaded dataset is then converted to a shape format. The dataset covers the whole Norway and contains a national identifier and name of each municipality in addition to the geometry of the municipality's boundary as polygons. In comparison to the other input datasets, the municipality boundaries are relatively stable, but the municipalities could be changed due to administrative reforms.

2.6 Challenges in integrating the source datasets

Open government data's burdens are inherent to large-scale distributed data integration, collective data manipulation and transparent data consumption [2]. Though the data providers are the most authorized actors in their domains in the public sector, none of the datasets are 100% accurate and consistent. The method to extract central government organizations described in Section 2.1 does not cover the whole scope of central government organizations because some organizations are not required to be registered using the specified method and need to be identified and added to the list manually.

There also exist inconsistencies between the source systems which increase the complexity in data integration and impact the quality of the data in the resulting SoE dataset. A common reason for inconsistencies is due to different domain focus. Propman's (Statsbygg's property management system) domain focus is real estate and

[18] http://www.kartverket.no/en/data/Open-and-Free-geospatial-data-from-Norway/

[19] http://data.kartverket.no/download/content/geodataprodukter

[20] http://www.kartverket.no/en/geodataarbeid/SOSI-Standard-in-English/SOSI-Standard-in-English/

building management for Statsbygg while the cadastral system focuses on legal rights and obligations on cadastral parcels and buildings in Norway. State-owned properties abroad are registered in Propman, but they are not available in the Norwegian cadastral system. Buildings with areas of less than 15 square meters are registered in Propman, but it is not mandatory to register them in the Norwegian cadastral system and they do not have corresponding national building identifiers. Furthermore, a physical building can be extended several times and the extensions are assigned new cadastral building numbers and therefore it can be connected to more than one national cadastral building number though Propman treats it as only one building.

Additionally, inconsistencies between systems are often caused by delayed or missing registrations or updates. The ownership change between organizations in the public sector is not always officially registered either to save the registration cost or due to lack of compulsory reporting routines. After a fusion between organizations the business entity register is updated while the previous organization number and name may still be registered in the cadastral system as a real rights holder. There are also examples of old ministries or organizations from earlier government periods registered as owners in the cadastre. Though all cadastral parcels in the previous SoE report have registered national cadastral parcel identifiers, the registration is manual and sometimes includes invalid values. More than 70% of buildings in the previous SoE report lack of national cadastral building identifiers because this attribute was not mandatory.

The above challenges will be addressed and discussed in Sections 3.1 and 3.6 to improve the data quality of both the source datasets and the resulting dataset.

3 Linked Data generation process

The process of generating Linked Data for state-owned properties is shown in Fig. 1.

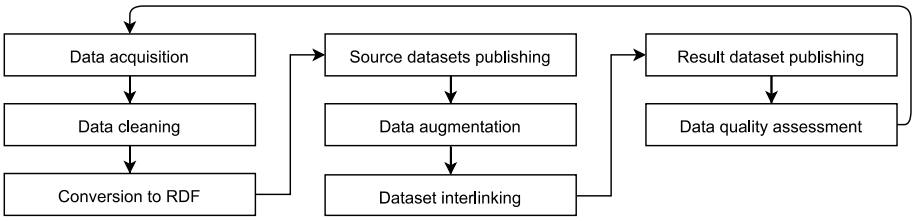

Fig. 1. SoE Linked Data generation process

The data acquisition step collects and prepares the datasets from multiple cross-domain sources as described in Section 2. Data cleaning is introduced in Section 3.1, conversion to RDF is described in Section 3.2, data augmentation using SPARQL CONSTRUCT[21] queries is detailed in Section 3.3, data interlinking is explained in Section 3.4, dataset publishing for the source datasets and result dataset is presented

[21] https://www.w3.org/TR/rdf-sparql-query/#construct

in Section 3.5, and data quality assessment on the result dataset is described in Section 3.6. DataGraft[22] [3][4][5] – a cloud-based platform for data cleaning and Linked Data generation – has been used for generating the Norwegian SoE dataset.

3.1 Data cleaning

The data cleaning step aims to remove syntactic and/or semantic errors in the source datasets by following the tabular data cleaning approach described in [6]. Examples of data cleaning and preparation tasks include:

- The source datasets inherits the number and time-date formatting from their original systems which are Norwegian in this case. The Norwegian decimal separator is comma and it is replaced by point so that the dataset conforms to the decimal formatting used in the RDF conversion tool in DataGraft (e.g., the decimal 300,5 is replaced by 300.5).
- The cleaning process also recognizes and unifies null values for attributes with null values or similar (e.g., converting date values from "", "0", "101", etc., to the unified conventional null date value "17530101").
- Not all records in the previous SoE report dataset have valid cadastral parcel identifier values and many records lack cadastral building identifiers (ref. Section 2.6). A multistep procedure was developed to assign unique identifiers for cadastral parcels and buildings to handle the situation of missing or non-valid national identifiers. The national cadastral parcel identifier is assigned as unique identifier in the first place if it is available; otherwise a unique identifier is generated by concatenating municipality number, cadastral unit number, property unit number and leaseholder number.
- The default separator semicolon may also occur as part of the text in some of the text columns, which will cause wrong mapping of columns. The text columns containing semicolons are first identified and then cleaned by removing the extra semicolons.

3.2 Conversion to RDF

This section describes the process of converting the source datasets to RDF format. In addition to standard and established ontologies such as DBpedia-owl and schema.org, the proDataMarket ontology [7][23] is used as a central reference model for RDF transformation. The proDataMarket ontology reuses the Land Administration Domain Model (LADM) defined in ISO 19152:2012[24] standard and cadastral parcel concept specified by the European Union's INSPIRE data specifications[25].

[22] https://datagraft.io/

[23] Available at http://vocabs.datagraft.net/ and the names start with proDataMarket.

[24] http://www.iso.org/iso/catalogue_detail.htm?csnumber=51206

[25] http://inspire.ec.europa.eu/data-model/approved/r4618-ir/html/

Table 2. Transformation script files for the involved datasets

Input dataset	Source format	Size	Transformation scripts	Output, triples	Dataset is public
Central government organizations dataset	CSV	46KB (217 records)	Central government organization transformation[26]	5049	Yes
Cadastral parcel ownership dataset	CSV	3.83MB (18105 records)	Cadastral parcel ownership transformation[27]	559743	No
Cadastral parcel geospatial dataset	Shape file	4GB (3673391 records)	Cadastral parcel geospatial dataset transformation[28]	~118 M	No
Building dataset	CSV	2.59MB (23651 records)	Buildings built on the state-owned or state-leased cadastral parcels transformation[29]	454192	No
Building point geo-dataset	Shape file	288 MB (4637654 records)	The building geospatial dataset transformation[30]	23188270	No
Building accessibility dataset	CSV	81 KB (665 records)	Building accessibility dataset transformation[31]	19285	Yes
Previous SoE dataset	CSV	1,91 MB (11241 records)	The previous SoE dataset transformation[32]	663219	Yes
Municipality boundaries dataset	Shape file	24,1 MB (428 records)	The municipality boundaries dataset transformation[33]	3424	Yes

The transformation scripts for the datasets are freely available as public transformations on DataGraft for registered users[34]. Table 2 shows the latest transformation scripts for each dataset with links to the actual transformation scripts provided as footnotes.

[26] https://datagraft.io/prodatamarket_publisher/transformations/the-central-government-organization-transformation-1010c106-2254-4c8b-9480-b88d47a41323

[27] https://datagraft.io/prodatamarket_publisher/transformations/the-cadastral-parcel-ownership-transformation

[28] https://datagraft.io/prodatamarket_publisher/transformations/the-cadastral-parcel-geospatial-dataset-transformation

[29]https://datagraft.io/prodatamarket_publisher/transformations/buildings-built-on-the-state-owned-or-state-leased-cadastral-parcels-transformation

[30] https://datagraft.io/prodatamarket_publisher/transformations/the-building-geospatial-dataset-transformation

[31] https://datagraft.io/prodatamarket_publisher/transformations/the-building-accessibility-dataset-transformation

[32] https://datagraft.io/prodatamarket_publisher/transformations/the-historical-soe-dataset-transformation

[33] https://datagraft.io/prodatamarket_publisher/transformations/the-munincipalitiy-boundaries-dataset-transformation

[34] The scripts are currently best visible using Chrome.

The transformation scripts are also published at a third party site[35]. The scripts are coded in Clojure[36] and include two parts: a data cleaning pipeline for data cleaning and preparation tasks; and a part for RDF mapping.

The central government organizations dataset reuses the established vocabulary from *schema.org* and *DBpedia-owl* to map and convert data to RDF. Attributes like the organization number, name, type, founding date and parent organization are mapped to *schema:leiCode*, *schema:legalName*, *dbpedia-owl:type*, *schema:foundingDate* and *schema:parentOrgnization* respectively.

The proDataMarket ontology is the core ontology used to map and transform the four cadastral datasets, the building accessibility dataset and the previous SoE dataset. A cadastral parcel is mapped to *prodm-cad:CadastralParcel* and a building is mapped to *prodm-cad:Building*. Ownership relationships are mapped to *prodm-cad:RealRights* which connect a *prodm-cad:RightsHolder* to *prodm-cad:CadastralParcel* or *prodm-cad:Building*. The *prodm-cad:RightsHolderOrganization* subsumes both *prodm-cad:RightsHolder* and *schema:Organization*, which links automatically the rights holders to the organizations converted using schema.org vocabulary. The geospatial information of cadastral parcels and buildings are mapped to *gsp:asWKT* either as *sf:MultiPolygon* or *sf:Point.* The building accessibility is modelled as indicators on buildings by connecting *prodm-com:Building* to *prodm-com:Indicator* through object property *prodm-com:hasIndicator*.

The geospatial information of municipality boundaries are also mapped to *gsp:asWKT* as *sf:MultiPolygon.* The municipality code is mapped to *au:nationalCode*, and the municipality name is mapped to *rdfs:label*.

3.3 Data augmentation using SPARQL CONSTRUCT queries

The source cadastral datasets cover both the state-owned and non-state-owned properties and buildings. In the data augmentation process SPARQL CONSTRUCT[37] queries are executed on the datasets from RDF conversions in Section 3.2 to generate subsets of state-owned properties and buildings. Table 3 lists the SPARQL CONSTRUCT queries that generate new datasets by only selecting triples related to the state-owned cadastral parcels or buildings. For example, the state-owned or state-leased cadastral parcel geospatial dataset is generated by first executing a CONSTRUCT query to indicate all the state-owned cadastral parcels by integrating the cadastral parcel geospatial dataset with the central government organization dataset. Afterwards, a second CONSTRUCT query is executed to select out triples related to ownership information about the state-owned or state-leased cadastral parcels. CONSTRUCT query can also be used to do calculations such as calculating the area summary for a cadastral parcel that can include one or more land parcels.

[35] https://zenodo.org/record/834300/files/SPARQLQueriesForLinkedDataGeneartion.pdf

[36] https://clojure.org/

[37] https://www.w3.org/TR/rdf-sparql-query/#construct

Table 3. SPARQL CONSTRUCT queries

Input datasets as RDF		CONSTRUCT queries	Output datasets as RDF		
Name	**Triples**		**Name**	**Triples**	**Public**
Cadastral parcel geospatial dataset; Central government organization dataset	~118 M	1. CONSTRUCT query #1[38] to indicate state-owned or state-leased cadastral parcels 2. CONSTRUCT query #3[39] to select geospatial information about state-owned or state-leased cadastral parcels	The state-owned or state-leased cadastral parcel geospatial dataset	416604	Yes
Cadastral parcel ownership dataset	559743	1. CONSTRUCT query #1 2. CONSTRUCT query #2[40] to select ownership information about state-owned or state-leased cadastral parcels	The state-owned or state-leased cadastral parcel ownership dataset	206962	Yes
State-owned or state-leased cadastral parcel geospatial dataset	416604	CONSTRUCT query #4[41] to calculate the area of each cadastral parcel as summary of the belonging land parcels	The state-owned or state-leased cadastral parcel areas	40476	Yes
Building dataset	454192	1. CONSTRUCT query #1 2. CONSTRUCT query #5[42] to select information about buildings built on state-owned or state-leased cadastral parcels	The state-owned building	267862	Yes
Building dataset; State-owned or state-leased cadastral parcel ownership dataset	~ 1 M triples	CONSTRUCT query #6[43] to generate ownership dataset for state-owned buildings	The state-owned building ownership dataset	15380	Yes
Building geospatial dataset	23188270	CONSTRUCT query #7[44] to generate geospatial dataset of state-owned buildings	The state-owned building geospatial dataset	42295	Yes

[38] https://datagraft.io/prodatamarket_publisher/queries/soe-construct-query-1
[39] https://datagraft.io/prodatamarket_publisher/queries/soe-construct-query-3
[40] https://datagraft.io/prodatamarket_publisher/queries/soe-construct-query-2
[41] https://datagraft.io/prodatamarket_publisher/queries/soe-construct-query-4
[42] https://datagraft.io/prodatamarket_publisher/queries/soe-construct-query-5
[43] https://datagraft.io/prodatamarket_publisher/queries/soe-construct-query-6
[44] https://datagraft.io/prodatamarket_publisher/queries/soe-construct-query-7

SPARQL CONSTRUCT queries are also used to infer triples based on known business rules. For example, the ownership of buildings does not exist in the original datasets, but it can be inferred by using this rule: *The owner or lessor of a cadastral parcel owns the buildings on the cadastral parcel*.

3.4 Interlinking with other datasets

Datasets interlinking is a fundamental prerequisite of the semantic Web [8]. The resulting SoE dataset is linked with several central Linked Open datasets in order to increase its reusability to support queries on cross-domain distributed datasets. Norway as a country is linked to *http://sws.geonames.org/3144096/* from the GeoNames dataset. The municipalities are modelled as administrative units and they have *owl:same* links to both the DBpedia[45], GeoNames[46] and lenka.no (the Norwegian RDF-resources for geographical breakdown)[47]. For example, the municipality of Oslo is linked to *http://dbpedia.org/page/Oslo* at DBpedia on municipality names, and is also linked to *http://data.lenka.no/geo/inndeling/03/0301* at *lenka.no* on the national municipality identifier. The linking triples are produced using SPARQL CONSTRUCT queries. In the result dataset 405 triples are connected to DBpedia, 422 triples are connected to GeoNames and 404 triples are connected to Lenka.no.

3.5 Datasets publishing

Both the RDF conversion results of source datasets from Section 3.2 and the data augmentation results of SPARQL CONSTRUCT queries from Section 3.3 are published via the DataGraft platform. Published data can be queried through the generated SPARQL endpoint or accessed via APIs. The resulting SoE dataset is described in more details in Section 4.

3.6 Data quality assessment using the interlinked datasets

Challenges related to data quality are mentioned in Section 2.6 and a rule-based approach for data quality assessment and improvement has been introduced in [9]. Table 4 shows some examples of quality check scenarios where SPARQL queries are executed on interlinked datasets to identify inconsistencies between source systems. All the queries are shared and freely available. The table also associates possible reasons to the inconsistencies. The results from the SPARQL queries can help the responsible staff control and improve the data quality in the source systems by following the suggested quality improvement strategies. The updated source datasets with better data quality will then be reloaded to the Linked Data generation process to produce an updated resulting SoE dataset with improved quality.

[45] http://wiki.dbpedia.org/wiktionary-rdf-extraction
[46] http://www.geonames.org/
[47] http://data.lenka.no/

Table 4. Data quality assessment using interlinked source datasets

No.	The SPARQL query identifies	# of triples	Possible reasons	Suggested quality improvement strategy
1	The owner name difference between cadastral system and business entity register[48]	146	Delayed or missing updates of owner names in the cadastre.	Update the owner names in the cadastre.
2	The state-owned properties that are missing in the previous SoE report[49]	6880	The properties were acquired after the previous report was made.	No actions needed though it reflects partially the quality of the previous SoE report.
			The properties were forgotten to be registered in the previous SoE report.	
3	The state-owned properties from the previous SoE report that are missing in the resulting SoE dataset[50]	2857	The properties were sold to a non-central government organization after the previous report was made.	No actions needed.
			The properties are abroad.	
			There has been organization change with the owner and the owner's organization number is no longer valid in the business entity register.	Update the owner's organization number and name in the cadastre.
			The ownership change between organizations in the public sector is not always officially registered in the cadastre.	Inform the current owner organization to update the ownership in the cadastre.
			The owner's organization is not officially registered as central government organization in the business entity register.	Update the organization in the business entity register if it is applicably or add it to the manual exception list of the central government organization dataset.

4 SoE dataset overview

Table 5 lists the technical details of the resulting SoE dataset. The result is a new dataset of state-owned properties and buildings which contains all the publicly available data published in Table 2 and Table 3. There are a total of 1,223,208 unique tri-

[48] https://datagraft.io/prodatamarket_publisher/queries/soe-query1-the-owner-name-difference
[49] https://datagraft.io/prodatamarket_publisher/queries/soe-query2-missing-soe-records
[50] https://datagraft.io/prodatamarket_publisher/queries/soe-query3-missing-result-soe-records

ples in the dataset. In addition to DataGraft, the dataset dump and documentation are also published at a third-party site Zenodo[51]. The dataset is registered in the datahub.io data catalogue[52]. The resources defined in datahub.io for the Norwegian SoE dataset include the Norwegian SoE SPARQL endpoint, RDF dump, proDataMarket vocabulary, VOID file, Linked Data Generation SPARQL queries, and example SPARQL queries to help users understand and use the dataset.

Table 5. Technical details of the resulting SoE dataset

Name	Norwegian State of Estate Report dataset
URL	https://datagraft.io/prodatamarket_publisher/sparql_endpoints/norwegian-state-of-estate-report-04693e1f-4060-48c1-8ab9-888a6c95f6d6
VOID file	https://datahub.io/dataset/norwegiansoe/resource/f5a83fbb-4324-43c2-a3da-5865b1f2d44e
Data download	https://rdf.datagraft.net/4035596353/db/repositories/norwegian-state-of-estate-report-6/statements
Ontology	http://vocabs.datagraft.net/
Version	1.0
Version date	27.June.2017
License	Norwegian Licence for Open Government Data (NLOD)

Table 6. RDF data sample on cadastral ownership dataset

```
@prefix prodm-cad: <http://vocabs.datagraft.net/proDataMarket/0.1/Cadastre#> .
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix dul: <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc11: <http://purl.org/dc/elements/1.1/> .
@prefix schema: <https://schema.org/> .
<http://vocabs.datagraft.net/proDataMarket/0.1/Cadastre#CadastralParcel/0214121900>
prodm-cad:hasCadastralID "214/121/9/0" .
<http://vocabs.datagraft.net/proDataMarket/0.1/Cadastre#RealRights/0214121900971032081>
  a prodm-cad:RealRights ;
  dbo:type "HJEMMELSHAVER"@no, "OWNER"@en ;
  dul:defines
<http://vocabs.datagraft.net/proDataMarket/0.1/Cadastre#CadastralParcel/0214121900>,
<http://vocabs.datagraft.net/proDataMarket/0.1/Cadastre#RightsHolderOrganization/97103208
1> ;
  dc:source "cadaster" ;
  prodm-cad:hasStartDate "2006-02-23T01:00:00.000+01:00"^^xsd:dateTime ;
  prodm-cad:hasEndDate "1753-01-01T01:00:00.000+01:00"^^xsd:dateTime .
prodm-cad:RealRights rdfs:subClassOf dc11:Rights .
<http://vocabs.datagraft.net/proDataMarket/0.1/Cadastre#RightsHolderOrganization/97103208
1>
  a prodm-cad:RightsHolderOrganization ;
  schema:leiCode "971032081" .
```

[51] https://zenodo.org/

[52] Available at https://datahub.io/dataset/norwegiansoe

Table 6 shows RDF data example that models cadastral ownership described in Table 1. The current plan to update and maintain the dataset is scheduled every 6 months on DatarGraft.io though patch releases are also supported when necessary. Data quality assessment and evaluation are compulsory steps in the publishing process to improve the data quality. In addition to datahub.io as a dataset sharing and user feedback channel, a data marketplace for property data is under development which focuses on selling, sharing and maintaining the datasets.

5 Application scenarios and use cases

Generating the SoE report is the main application scenario of the resulting SoE dataset. The report includes a table-based list of all state-owned properties and buildings and also as aggregated list grouped by municipalities and organizations – both lists can be generated by SPARQL queries. Fig. 2 presents an example of visualization of Norwegian state-owned properties per county, both as a pie chart and on the map. There are 11342 state-owned properties in Norway and the total area is 4043 million square meters as shown in the figure.

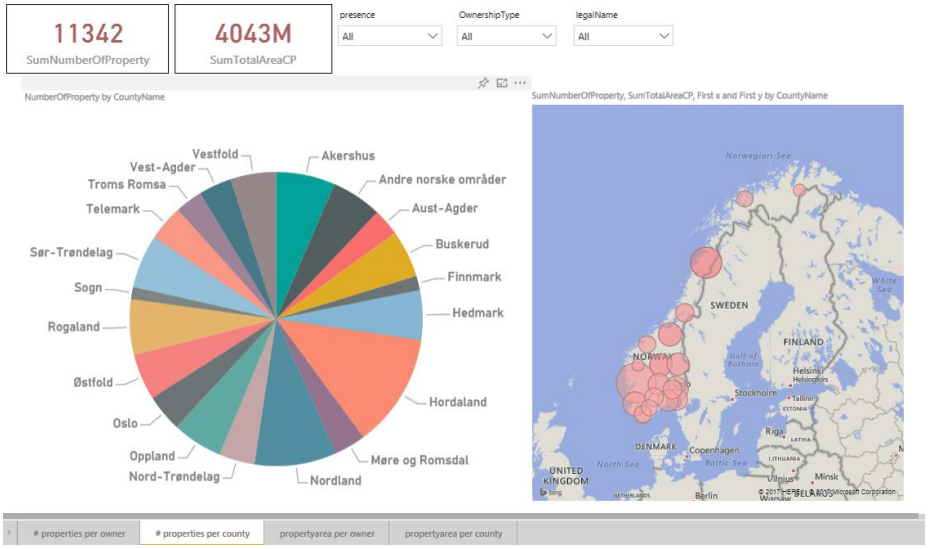

Fig. 2. The visualization of Norwegian state-owned properties per county

In addition, the dataset can also be integrated with other contextual datasets to generate added value as follows.

Risk and vulnerability analysis of state-owned buildings. The dataset can be integrated with natural hazard datasets such as flood continuance and storm mean hours. Flood continuance map presents the areas in Norway that can possibly flood and it reveals areas where the danger of flooding needs to be further assessed. The geospa-

tial dataset is provided by The Norwegian Water Resources and Energy Directorate as an open source dataset[53]. It is converted to RDF and published at DataGraft and accessible via the data endpoint[54]. Examples of queries that are enabled by such integration between the SoE dataset and the flood continuance dataset include:

- How many buildings owned by Statsbygg have flood risk in Norway?[55]
- Which state-owned buildings have the flood risk in Ås municipality?[56]

The result of the risk and vulnerability analysis helps the property owners to take proactive actions in maintenance and thereby reduce the damage and cost when natural hazard occurs.

Searching properties suitable for asylum reception centres. This is a demanding task that has significant effect on local communities. The demographical statistics and relevant geographical data can be integrated with the Norwegian SoE dataset to help identifying state-owned properties suitable for asylum reception centres.

Visitor application for public buildings. A subset of the Norwegian SoE dataset can be integrated with cultural heritage data, traffic data, weather data and other contextual data to provide input to a public building application for visitors.

6 Related work

State-owned properties and buildings are shared as a downloadable text reports in the U.K. and Norway and similar reports are available for other countries. However, no systematic transformations of such reports to Linked Open Data exist to date.

Cadastral data including land ownership data is one of the core datasets used in the LOD generation of the SoE dataset. Openness of land ownership data is presented at the OKFN's website[57]. The statistics from 2015 show that the land ownership data is 100% open in Denmark and Uruguay, and only partially open in several other countries including, for example, Norway with 45% openness and Spain with 5% openness. There is no known effort to publish the land ownership data as Linked Data in Norway. There are few known efforts in Europe or worldwide to transform and publish cadastral data as Linked Data, and even fewer as Linked Open Data. One related research is [10] that developed a process to generate, integrate and publish geospatial Linked Data from several Spanish national datasets including the administrative units from the Spanish cadastre, and this process methodology has been applied in [11] to integrate two cadastre datasets for a city in Colombia. Neither of the Spanish and Colombian datasets from the research is open to the public.

[53] http://nedlasting.nve.no/gis/

[54] https://rdf.datagraft.net/4035596353/db/repositories/statsbygg_data-2

[55] https://datagraft.io/prodatamarket_publisher/queries/soe-query4-floodriskstatsbygg

[56] https://datagraft.io/prodatamarket_publisher/queries/soe-query5-floodrisk-as

[57] http://global.census.okfn.org/dataset/land accessed 20.March.2017

The Norwegian government implemented PSI-directive in the Norwegian law on 1st January 2009 [12]. One of the central Norwegian LOD projects was Semicolon[58] which aimed to improve interactions in the Norwegian public sector. The project presented a status report on open and linked data in Norway. Examples of datasets available in LOD formats included business entity register, registry of municipalities and counties, central register of parties and their income, travel information for the public transportation system in Oslo, city bike stand status, etc., as described in [13]. The PlanetData project[59] made a report on Norwegian LOD extensions [14], which included two business cases and six updated or new datasets in RDF. Though the business entity register has been published as RDF in the aforementioned projects, the endpoint is not stable enough to be reused for central government organizations dataset in the SoE case.

7 Summary and outlook

This paper introduced the SoE dataset (containing information about state-owned properties integrated from a variety of relevant sources), together with the process of generating a Linked Data representation of the dataset. Several government datasets over multiple systems/databases including the cadastral system and business entity register of Norway are integrated. The Linked Data generation process includes data cleaning, source dataset publishing, conversion to RDF, data augmentation using SPARQL CONSTRUCT queries and dataset publishing. The technical details of the resulting SoE dataset were presented. The data quality challenges were discussed in detail and SPARQL queries on interlinked source datasets were proposed to improve data quality. New services can also be generated by integrating the resulting SoE dataset with other contextual data.

In terms of future work, the preparation of source datasets currently includes several manual steps and it can be automatized to a certain level. The data cleaning process can be more user friendly by introducing visualization tools for semantic and syntactic error checks. The experience and process methods can further be tested with similar datasets from other countries.

Acknowledgements. The work has been funded by the European Commission under the proDataMarket project (H2020-ICT-2014-1 644497). The authors would like to thank Bjørg Pettersen and the proDataMarket consortium for additional contributions.

References

1. 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

[58] http://www.semicolon.no/

[59] http://www.planet-data.eu/

Markup Languages for the Semantic Web (pp. 456-464). Springer International Publishing.
2. Ding, L., Lebo, T., Erickson, J.S., DiFranzo, D., Williams, G.T., Li, X., Michaelis, J. Graves, A., Zheng, J.G., Shangguan, Z. and Flores, J.. (2011). TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web, 9(3), 325-333.
3. Roman, D., Nikolov, N., Putlier, A., Sukhobok, D., Elvesæter, B., Berre, A., Ye, X., Dimitrov, M., Simov, A., Zarev, M. and Moynihan, R. 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.
4. Roman, D., Dimitrov, M., Nikolov, N., Putlier, A., Sukhobok, D., Elvesæter, B., Berre, A., Ye, X., Simov, A. and Petkov, Y. (2016, May). DataGraft: Simplifying Open Data Publishing. ESWC (Satellite Events) 2016: 101-106.
5. Roman, D., Dimitrov, M., Nikolov, N., Putlier, A., Elvesæter, B., Simov, A. and Petkov, Y. (2011). DataGraft: A Platform for Open Data Publishing. In the Joint Proceedings of the 4th Inter-national Workshop on Linked Media and the 3rd Developers Hackshop. (LIME/SemDev@ESWC 2016).
6. Sukhobok, D., Nikolov, N., Pultier, A., Ye, X., Berre, A., Moynihan, R., Roberts, B., Elvesæter, B., Mahasivam, N. and Roman, D. (2016, May). Tabular data cleaning and linked data generation with Grafterizer. ESWC (Satellite Events) 2016: 134-139.
7. L. Shi, N. Nikolov, D. Sukhobokb, T. Tarasovac and D. Roman. 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.
8. Bizer, C., Heath, T., Ayers, D. and Raimond, Y. (2007, June). Interlinking open data on the web. In Demonstrations track, 4th european semantic web conference, innsbruck, austria.
9. Shi, L., and Roman, D. Using rules for assessing and improving data quality: A case study for the Norwegian State of Estate report. In the Proceedings of the Doctoral Consortium, Challenge, Industry Track, Tutorials and Posters @ RuleML+RR 2017, hosted by International Joint Conference on Rules and Reasoning 2017 (RuleML+RR 2017), London, UK, July 11-15, 2017.
10. Vilches-Blázquez, L. M., Villazón-Terrazas, B., Corcho, O., and Gómez-Pérez, A. (2014). Integrating geographical information in the Linked Digital Earth. International Journal of Digital Earth, 7(7), 554-575.
11. Saavedra, J., Vilches-Blázquez, L. M., and Boada, A. (2014). Cadastral data integration through Linked Data.
12. Lov om rett til innsyn i dokument i offentleg verksemd. §9(2006).
13. Roman, D. and Norheim D. (2012). An Overview of Norwegian Linked Open Data. In proceedings of the Fourth International Conference on Information, Process, and Knowledge Management (eKNOW) (pp. 93-96).
14. Roman, D., Mjelva , J. K., Norheim D., Grønmo, R. NorthPole Report on Norwegian LOD extensions. http://www.planet-data.eu/sites/default/files/D12.2.2.pdf accessed 23th July 2017.