GEGN.2/2023/63/CRP.63

13 March 2023

English

United Nations Group of Experts on Geographical Names
2023 session
New York, 1 – 5 May 2023
Item 14 of the provisional agenda [*]
Geographical names data management

Availability of Dutch geographical names as linked data

Submitted by the Netherlands [**]

Summary

Kadaster, the cadastre, land registry and mapping agency of the Netherlands, has made various key registers available as linked data in support of the ongoing development and application of linked data technologies within the country's Government. In a recent update to the approach taken in publishing linked data, several of these key registers, including topographical and address registers, were published under shorter time frames and with fewer resources than previous approaches. This update was also a precursor to the development and realization of the Kadaster Knowledge Graph.

The Graph itself, as well as the associated supporting architecture, is completely open standards-based to support the interoperability of this source of information with other sources, both governmental and third-party. Several open key registers are included in the Graph, the most notable for geographical names being the Key Register of Topography (BRT). Geospatial objects with geographical names available in this key register are now linked with those from other registers, providing users with rich, integrated information about geospatial objects from one source.

Technical users and third parties are able to access this information through various standards-based service interfaces, including SPARQL and Elasticsearch. These same services are also used by data browser and viewer applications aimed at providing non-technical users with access to this information in a low-threshold manner, for example Object Viewer and Toponamenzoeker ("topo names finder"). The Kadaster Knowledge Graph forms part of a larger vision for an ecosystem of federated data sources based on semantic technologies.

[*] GEGN.2/2023/1
[**] Prepared by Alexandra Rowland, Kadaster

Availability of Geographical Names as Linked Data

Introduction

Kadaster, the Netherlands' Cadastre, Land Registry and Mapping Agency, has made various key registers available as linked data in support of the ongoing development and application of these technologies for e-government purposes. In a recent update to the approach taken in publishing linked data, several key registers, including the topographical and address registers, were published and are now publicly accessible. Increasing demand for linked data products and the availability of new linked data technologies have highlighted the need for a new, innovative approach to linked data publication by Kadaster in the interest of reducing the time and costs associated with publication. This update was also a precursor to the development and realisation of the Kadaster Knowledge Graph (KKG).

The approach used for linked data publication is completely based on open standards. These standards are used to publish individual key registers as linked data but also to combine these registers using a shared, standardised schema. This design principle supports increased interoperability between systems, both governmental and third-party. Several open key registers are included in this graph, the most notable of which for geographical names is the Key Register of Topography (Dutch acronym: BRT). Geospatial objects with geographical names available in this key register are now linked with geospatial objects contained in other registers, providing users with rich, integrated information about geospatial objects from a single source. The Kadaster Knowledge Graph forms part of a larger vision for an ecosystem of federated data sources based on semantic technologies. In line with the standards-based approach, technical users and third parties are able to access this information through various open standards-based interfaces, including SPARQL, the native linked data query language, and ElasticSearch as a full-text search engine. These same services are also used by data browsers, viewers and other applications; providing non-technical users with access to the information in a low-threshold manner. The ObjectViewer and the Toponamenzoeker are examples of such applications.

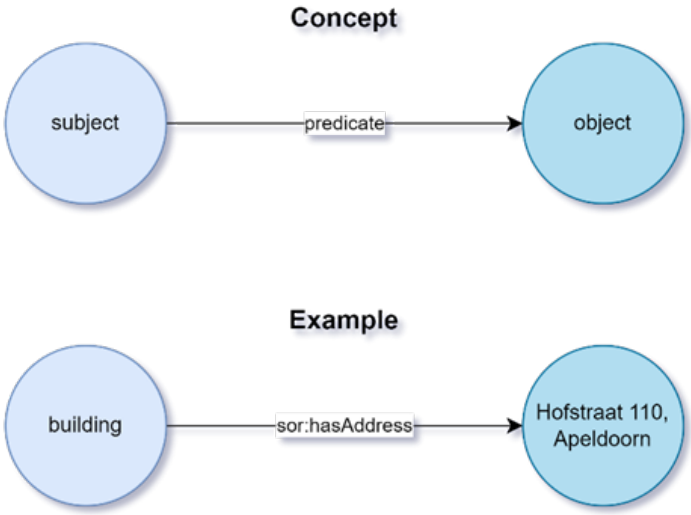
This paper will provide insight into the ongoing activities related to linked data publication at Kadaster with the intention of highlighting the availability of geographical names as linked data and showcasing how these names now form part of the integrated geospatial information accessible to a range of end users. The first two sections of this paper will briefly outline what linked data is and how it is being applied at the Dutch Cadastre and Mapping Agency. The third section will provide examples of geographical names as linked data and how these can be accessed and used by end-users in finding more information about geospatial objects. The final section will conclude with insight into future work planned on the topic of linked data and geographical names.

What is Linked Data and the Knowledge Graph?

The semantic web, defined as the facilitation of a fully linked, structured and machine-readable web of online data, uses various technologies to formalise the semantics related to data and related domains of knowledge. The development of the semantic web is primarily based on linked data technologies. The foundational technologies include the Resource Description Framework (RDF), a standard for describing web resources to support data interchange on the web, the associated schema for RDF (RDFS) and a range of Web Ontology Languages (OWL). These web technologies formalize the semantics of data using defined ontologies and capture these semantics in triples using unique resource identifiers (URIs) for nodes and relations (Tekli et al., 2013; Guus & Raimond, 2014); enabling location-independent cross-machine readability and interaction (Guus & Raimond, 2014). Objects, places or persons, for example, can be represented using a node and the relationships between these nodes are represented by an edge. This node-edge-node structure, or the subject-predicate-object structure, is known as a triple; a data representation structure which underpins the publication of data as linked data.

An example of this triple structure in the context of geospatial data is provided in the following figure (Figure 1).

Figure 1. Triple Structure (Subject-Predicate-Object) using a Geospatial Example (Rowland, Folmer & Baving, 2022).



Web standards are used in the creation of ontologies which connect nodes to one another. In the figure above, the 'sor:hasAddress' is one attribute in the ontology which represents the connection between the object 'building' and the attribute address 'Hofstraat 110, Apeldoorn'. A single triple can be connected iteratively to other triples, forming an expanded graph data model known as the knowledge graph. These connected triples can be semantically enriched and placed in context through the application of (domain) ontologies and connection to other knowledge bases using web technologies and standards. Triples, and indeed whole knowledge graphs, are stored in scalable graph databases known as triplestores. A knowledge graph combines several data management models in its implementation, including the traditional database model, a graph model as well as, and perhaps most critically, a knowledge base model bearing the formal semantics of numerous knowledge domains (Ontotext, n.d). This was introduced as a means of connecting data from different sources. Indeed, a knowledge graph can be defined as representing 'a network of real-world entities – i.e. objects, events, situations or concepts – and illustrates the relationship between them' (Fu & Sun, 2011 as referenced by Rowland, Folmer & Beek, 2020).

There are many examples of large knowledge graphs, both corporate and open source, that are actively used and expanded, including, but not limited to, Google's knowledge graph, DBpedia and GeoNames. The DBpedia knowledge graph, which started in 2006 and is the core of the Linked Open Data (LOD) movement, is a community-run and open-source project and is exposed and structured by making use of semantic web technologies and standards. Because Wikipedia data is used as its primary data source, the KG is constantly evolving and expanding as users make additions to Wikipedia, but it is also the largest multilingual knowledge graph available (DBPedia, n.d). Similarly, GeoNames is one of the largest knowledge graphs, primarily focused on geographical name data instances, and is available as an open-source dataset. Currently, the body of literature available on the advantages in the formalisation, development and implementation of geographic knowledge graphs is growing. Indeed, researchers often point to the role that geospatial semantics have in improving the interoperability and accessibility of geospatial information (Yan, 2019; Rohnzin et al., 2019) and to the computational advantages of developing geospatial-specific knowledge graphs (Zhu, 2019).

The interoperability of data made available on the web, such as data made available as part of a knowledge graph, is supported by the efforts of the World Wide Web Consortium (W3C), the main standardization organization for the World Wide Web. This Consortium publishes and maintains a variety of open standards, including common linked data standards such as OWL, RDF, SPARQL, PROV and SKOS. OWL is the modelling language used to describe RDF-structured data. PROV is the vocabulary used for representing and exchanging provenance information in RDF and SKOS is the vocabulary used to structure concepts in a taxonomy. SPARQL is the most widely used query language for linked data and allows users to query data flexibly and make use of the returned data in an application given that the user has knowledge about the structure of the data model. The linked data available through a SPARQL API can be returned in a range of formats including JSON-LD, Turtle and N-Quads. These formats are native-linked data serialization formats, each with a different structure.

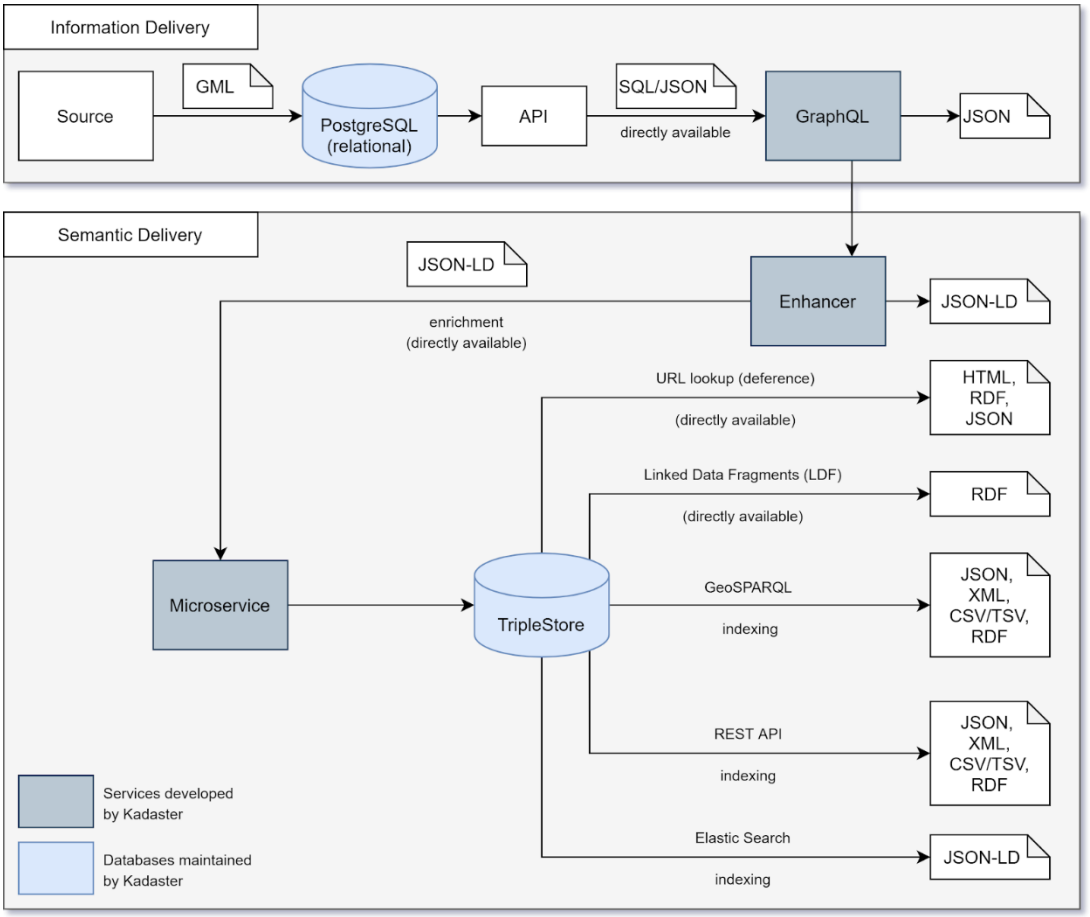
The following section discusses the architecture used in the creation and ongoing development of the Kadaster Knowledge Graph (KKG). The use of open standards published by the W3C in the modelling of the KKG is a central element here and includes OWL-based ontologies, RDF and RDFS for data representation as well as elements of the SKOS and PROV vocabularies for knowledge organization and provenance links between source datasets and the KKG. The resulting linked data is also made available through a SPARQL endpoint, another example of a standards-based implementation within the solution architecture.

What is Linked Data and the Knowledge Graph?

Dutch governmental geospatial datasets are organised in siloed key registers. The integration of these is generally poor despite the fact that much of the data contained in these key registers are related to information in other registers. Users who seek to combine these silos for a given purpose often have to resort to the use of specific tooling which supports this integration or downloading whole datasets and performing the integration of entire datasets. For example, attempting to answer the relatively simple geospatial question ‘Which churches were built before 1800 in the city of Amsterdam?’ requires the integration of the Key Register for Topography which includes building types and the Key Register of Addresses and Buildings (Dutch acronym: BAG) which includes the building year of each building. To perform this integration, a user either needs to make use of specific GIS tooling (e.g. QGIS or ArcGIS) or download the whole key register and then perform this integration. One of the solutions to this integration problem is to provide these data silos as linked open data and perform this integration using associated technologies (Rowland et al., 2022).

Although several of Kadaster’s geospatial assets have been available as linked data for a number of years (Folmer & Beek, 2017), the network effects of increased uptake in linked data technologies have demanded an updated, scalable approach to publication. This demand, coupled with the increasing availability of linked data technologies and standards, has been the driving force for innovation of linked data publication. Using this new approach, the Key Register for Addresses and Buildings, as well as the Key Register for Large Scale Topography (Dutch acronym: BGT), were transformed to linked data by a small internal team in 9 and 5 weeks respectively. These are complex linked datasets, each with a complex data model and containing between 800 million and 1 billion triples. Where previous approaches could be lengthy, this approach highlights improved resource and cost-effectiveness, strengthening the business case for linked data within an organisation such as Kadaster (Folmer et al., 2020). All linked datasets, including data models and API documentation, are available in the triple store managed by Kadaster’s Data Science Team (https://data.labs.kadaster.nl). The process of converting relational data for a given geospatial asset to linked data is completed in a number of steps taken during the Extract, Transform and Load (ETL) process. This process is illustrated in the architecture outline in the figure below (Figure 2).

Figure 2. Architecture Supporting the ETL Process which Delivers Linked Data (Rowland et al., 2022)



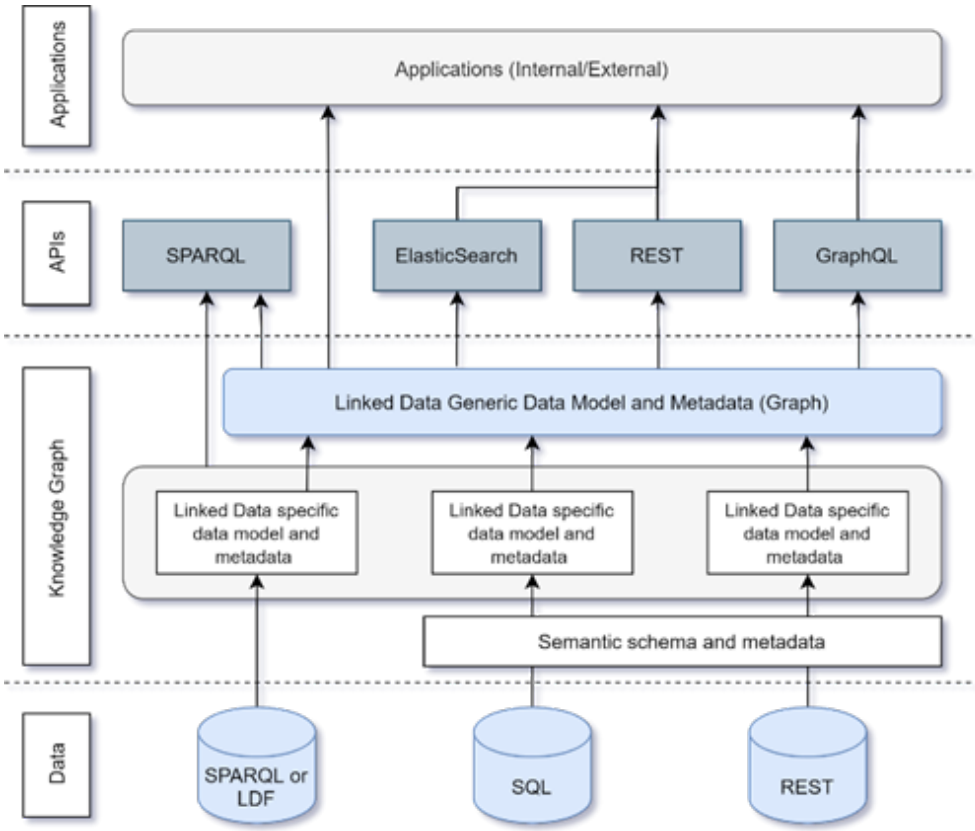
To briefly summarize this process, each key register or dataset is loaded from a relational database source to the PostgreSQL database instance following a Geography Markup Language (GML) indexing step (Brink et al., 2014) and is then made accessible through a GraphQL endpoint. GML is a language defined by the Open Geospatial Consortium (OGC) to express geographical features as XML. GraphQL is a query language supporting the description of graph data in an API. An internally developed microservice, denoted in the figure above as the Enhancer, is then used to query the data and return the results in JSON-LD serialization format. JSON-LD or JavaScript Object Notation in Linked Data format is a data interchange format commonly used for exchanging linked data. This microservice also publishes the resulting data to the triplestore, an instance of TriplyDB, which in turn makes the data available in a number of serialization formats based on the instantiation of services (e.g. a SPARQL service) as shown above. A preceding step to the loading of data into the triplestore is a SHACL validation step which ensures any loaded data complies with a defined data model and ensures that data has not been partially lost or fundamentally changed over the course of the transformation. Indeed, the process is automated as much as possible in order to avoid any human errors in the transformation. SHACL is the Shape Constraint Language used for the validation of data structured as RDF.

The KKG is composed of several key registers including the 1) Key Register for Addresses and Buildings, 2) the Key Register for Large Scale Topography and 3) the Key Register for Topography. When combining the key registers for the construction of the KKG, a central data model is used which simplifies the complexity seen in the siloed key registers. This is done with the goal of making the KKG

more suited for a wider range of user groups, including domain experts and developers as well as interested citizens, researchers and industry experts. In the initial development, the data model used for publication was based on the schema.org specification. Current versions of the KKG use the same architecture defined in Figure 3 but are now using the first version of the Samenhangende Object Registratie (SOR); a data model in development by Geonovum (https://www.geonovum.nl/), a Dutch government foundation which supports the standardisation and management of geospatial information in the Netherlands. This data model, which in English is called the Connected Object Registration, is being developed to improve the way the key registers and other related datasets are connected.

As illustrated in Figure 3, the KKG is realized through the creation of a layer on top of the key registers as linked data and is created by performing mappings between the data models of the source datasets and the SOR data model. As extensively discussed by Rowland et al. (2022), the implementation of these mappings is done through LD views, each of which transforms the data from the key register source dataset to linked data conforming to the SOR model based on predefined SPARQL construct queries. These LD views, in using SPARQL construct functionality, take a part of a data model from a key register and map this to an associated part of the SOR model. This process is a key part of the architecture outlined by the authors as this serves to preserve the provenance and traceability between the source data and that available in the KKG.

Figure 3. Architecture for the Implementation of the Kadaster Knowledge Graph (Rowland et al., 2022)



The KKG is now available for use in the Kadaster triplestore environment and contains approximately 680 million triples. The data in the KKG is updated on a quarterly basis in line with the quarterly updates of the various key registers and an updated notification is posted in the Kadaster triplestore on completion. As highlighted in Figure 3, it is the intention that the KKG is used directly by

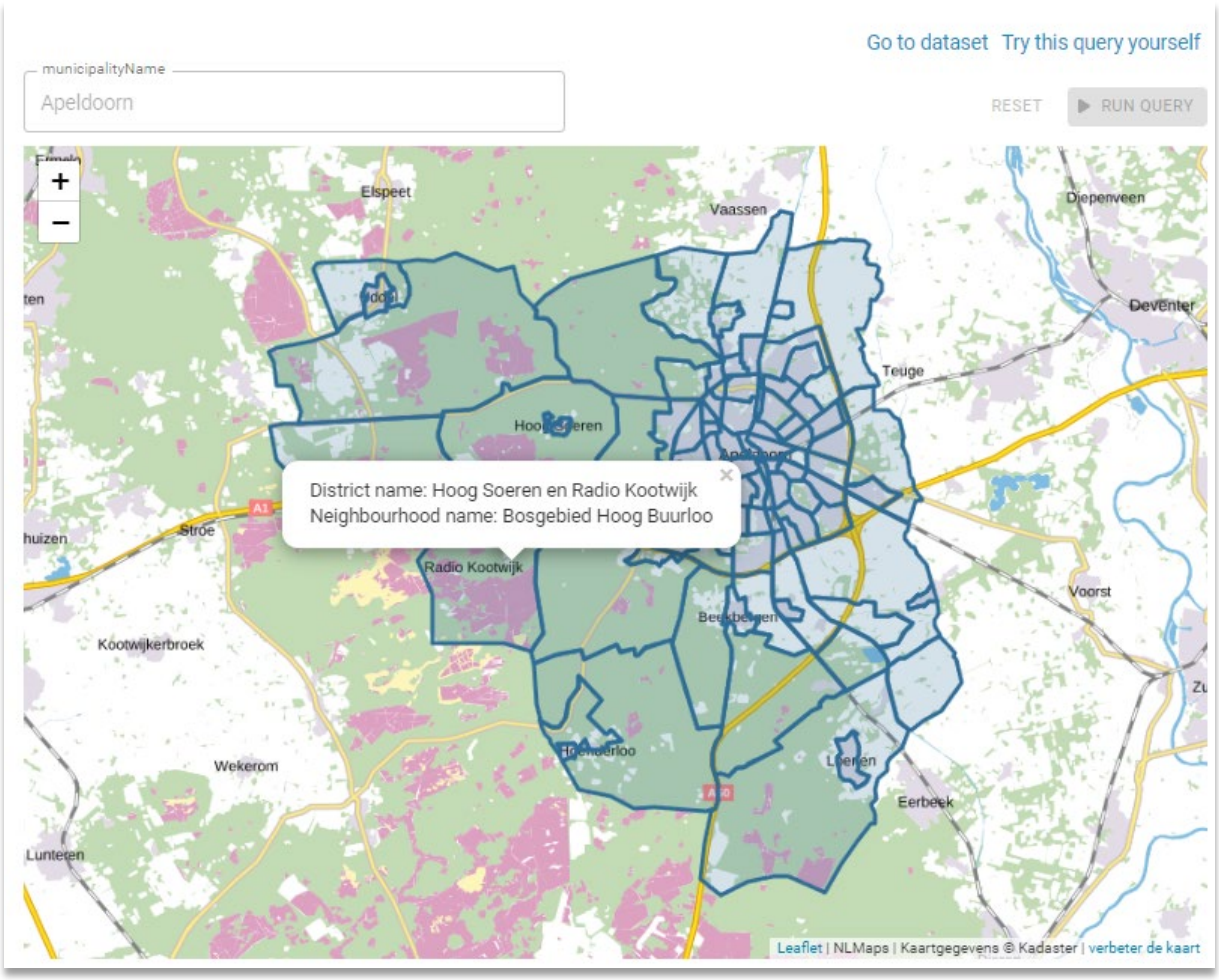
applications which access the data from various service options (e.g. SPARQL, REST or GraphQL) (Rowland, Folmer & Baving, 2022).

Dutch Geographical Names as Linked Data

The key registers contain various examples of Dutch geographical names, the most notable of which are those contained in the Key Register for Topography (Dutch acronym: BRT). Because this key register has been included in the linked data transformation and publication effort described above, these geographical names are also available as linked data and in the Kadaster Knowledge Graph. These names accompany information from associated geographical names in providing integrated data about geospatial objects. The availability of these geographical names as linked data can be demonstrated in a variety of ways as outlined in this section.

The first demonstration is provided based on a data story, a user interface including both text, tables and figures which include examples of Dutch geographical names in the Kadaster knowledge graph. The user interface accesses the knowledge graph through a SPARQL endpoint, querying the dataset live to return the results seen in the environment. In this way, the user is able to interact with the linked data and geographical names live without knowledge of linked data technologies. The demonstration environment is best explored live and can be found here. Where a live query of the linked data is being performed in the data story, users will see the button 'Try this query yourself' on the right-hand side of the results. If clicked, this will direct the user to the underlying SPARQL query and can be changed to visualise more data if necessary. The figure below is an example of one of these queries.

Figure 4. Map Displaying All Neighbourhoods and Districts in a Given Town



In the above query, the user is able to input ‘Apeldoorn’ as the municipality name and return a map of all neighbourhoods and districts in Apeldoorn. If the user clicks on the geometry, the name of both of these is displayed. These data is queried directly from the knowledge graph, demonstrating the existence of geographical names in this dataset. Clicking on ‘try this query yourself’ will result in the display of the SPARQL query below. With this query, the variables queried can be changed in order to display richer or more extensive results on the map.

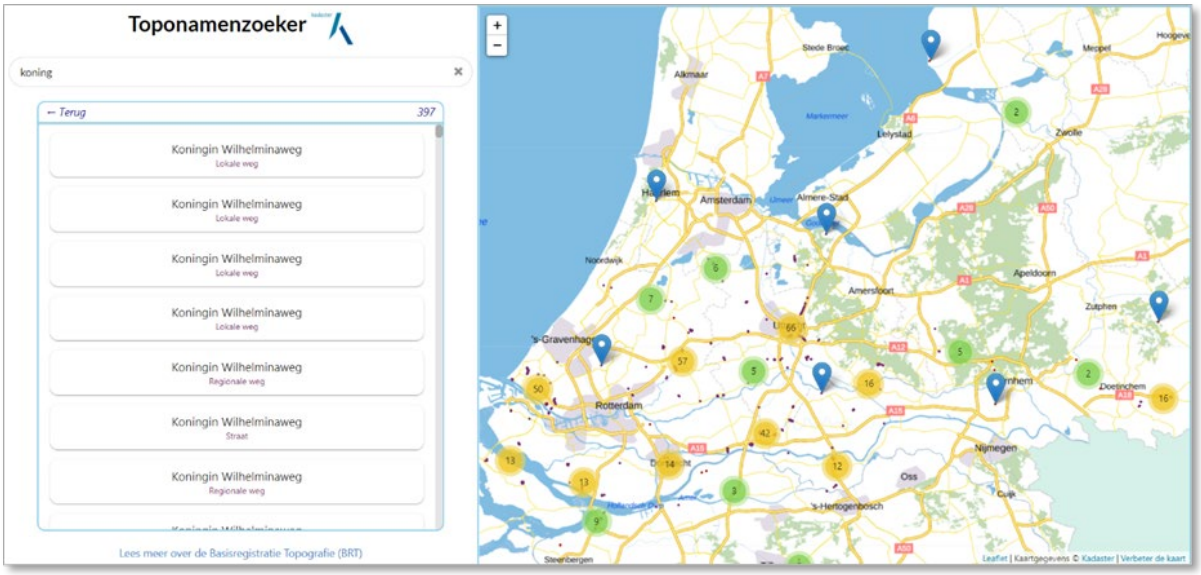
Figure 5. SPARQL Query which Results in the Map Shown in Figure 4.

```
PREFIX sd: <http://www.w3.org/ns/sparql-service-description#>
PREFIX sdo: <https://schema.org/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
select distinct
    ?neighbourhoodGeo
    ?neighbourhoodName
    ?districtName
    (strdt(concat('<h6>District name: ',str(?districtName),'</h6><h6>Neighbourhood name: ',str(?neighbourhoodName),'</h6>'),rdf:HTML) as ?
neighbourhoodGeoLabel)

WHERE {
  ?municipality
    a <https://data.labs.kadaster.nl/cbs/wbk/vocab/Gemeente>;
    rdfs:label ?municipalityName .
  ?district
    a <https://data.labs.kadaster.nl/cbs/wbk/vocab/Wijk>;
    rdfs:label ?districtName;
    geo:hasGeometry/geo:asWKT ?districtGeo;
    geo:sfWithin ?municipality .
  ?neighbourhood
    a <https://data.labs.kadaster.nl/cbs/wbk/vocab/Buurt>;
    rdfs:label ?neighbourhoodName;
    geo:hasGeometry/geo:asWKT ?neighbourhoodGeo;
    geo:sfWithin ?district .
}
```

The above data story is only one example of the ways in which it is possible for users to interact with linked data containing geographical names. While the data story is intended to support the querying of linked data live in a web page, other applications have been developed that support users in browsing this information based either on search functionality or on browsing data interactively on a map. The ability to search for geographical information based on geographical names is the primary functionality of the Toponamenzoeker application. This application supports users in searching 200,000 unique geographical names from topographical files and maps from the Kadaster, including municipal names, place names, water, roads and buildings. The figure below (Figure 6) illustrates this search functionality and the types of results that can be expected through this application. Here, the word ‘koning’ or king has been used to find the location of all geographical names containing this search term. The application itself can be found here.

Figure 6. Toponamenzoeker Linked Data Browser, using the search term 'Koning'. Retrieved from: https://labs.kadaster.nl/demonstrators/namen-app/#/.



A more recent application, developed as an extension to the Toponamenzoeker, allows users to search for information about geospatial objects based on a given address. The results of this query are shown on a map where information about building years, floor size and the type and usage of the building is shown in a pop-up window associated with the geometry of the object in question. Here, geospatial names associated with addresses, most notably street names, are the primary input for such a query. The results of searching for 'koningin' and choosing a relevant address are shown in the figure below (Figure 7). The application can be accessed here.

Future Work

The Kadaster Knowledge Graph as well as the surrounding architecture and associated applications has been successful in highlighting the potential that linked data technologies have in improving access to geospatial data, including geospatial names. In its current form, the Kadaster Knowledge Graph only contains open geospatial data. Going forward, it is the intention that the graph is expanded to include closed data which can only be accessed following authentication but that does provide individuals, largely Kadaster employees and other government officials, with this authorisation to harness the potential of this integrated data for richer analysis. Additionally, it is also the intention that the graph is expanded to include links to third-party data sources, including statistical information, cultural heritage information and information in support of the energy transition. Geographical names play an important role in each of these domains and as integrated information becomes increasingly accessible, play an important role in connecting, visualising and exploring this information.

The Group of Experts is invited to:

a. take note of linked data developments in the Netherlands and how geographical names support search and discovery applications for geospatial data;

b. identify mechanisms to support member states in making national geographical names data available as linked data.

Bibliography

A. Rowland, E. Folmer, T. Baving. “The Knowledge Graph as Interoperability Foundation for an Augmented Reality Application: The Case at the Dutch Land Registry”. In proceedings from the 9th International Kaleidoscope Conference: Extended Reality - How to Boost Quality of Experience and Interoperability, Accra Ghana. 2022. Available online: https://www.itu.int/pub/T-PROC-KALEI-2022.

A. Rowland, E. Folmer, W. Beek, R. Wenneker, “Interoperability and Integration: An Updated Approach to Linked Data Publication at the Dutch Land Registry”. ISPRS International Journal of Geo-Information. 2022.

A Rowland, E. Folmer, W. Beek. “Towards Self-Service GIS - Combining the Best of the Semantic Web and Web GIS”. International Journal of Geo-Information. 2020.

DBpedia. About. Available online: https://wiki.dbpedia.org/about (accessed on 15 October 2020).

Folmer, E., Ronzhin, S., Van Hillegersberg, J., Beek, W., Lemmens, R. (2020). Business Rationale for Linked Data at Governments: A Case Study at the Netherlands’ Kadaster Data Platform. IEEE. Access 8, 70822-70835.

Folmer, E. & W. Beek, ‘Kadaster Data Platform - Overview Archicture,’ Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings, vol. 17, no. 1, Sep. 2017. doi: https://doi.org/10. 7275/R5N58JJ0. [Online]. Available: https://scholarworks.umass.edu/foss4g/vol17/iss1/23.

J. Tekli, A.A Rjeily, R. Chbeir, G. Tekli, P. Houngue, K. Yetongnon, M.A Abebe, “Semantic to Intelligent Web Era. In proceedings of the 5th International Conference on Management of Emergent Digital Ecosystems, Luxembourg. Association for Computing Machinery, New York, NY, USA. Vol 20. 2013.

L. Van den Brink, P. Janssen, W. Quak, J. Stoter, “Linking Spatial Data: Automated Conversion of Geo-Information Models and GML data to RDF”. International Journal of Spatial Data Infrastructure. Vol. 9, p.p. 59-85. 2014.

Ontotext. What is a Knowledge Graph? Available online: https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/ (accessed on 15 October 2020)

Rohnzin, S.; Folmer, E.; Maria, P.; Brattinga, M.; Beek, W.; Lemmens, R.; van’t Veer, R. Kadaster Knowledge Graph: Beyond the Fifth Star of Open Data. Information 2019, 10, 310.

S. Guus, Y. Raimond, “RDF 1.1 Primer”. W3C Working Group Note. 2014. Available online: http://www.w3.org/TR/rdf11-primer/ (accessed on 20 April 2022).

Yan, B. Geographic Knowledge Graph Summarisation. 2019. Available online: https://www.iospress.nl/book/geographic-knowledge-graph-summarization/ (accessed on 11 October 2020).

Zhu, Y. Geospatial semantics, ontology and knowledge graphs for big Earth data. Big Earth Data 2019, 3, 187–190.