Interoperability and Integration: An Updated Approach to Linked Data Publication at the Dutch Land Registry

Alexandra Rowland[1], Erwin Folmer[2], Wouter Beek[3], Rob Wenneker[4]

[1] Kadaster & University of Twente, 7500 AE Enschede, The Netherlands
lexi.rowland@kadaster.nl
[2] Kadaster & University of Twente, 7500 AE Enschede, The Netherlands
erwin.folmer@kadaster.nl
[3] Kadaster & Triply, 1043 BP Amsterdam, The Netherlands
wouter@triply.cc
[4] Kadaster, 7311KZ Apeldoorn, The Netherlands
rob.wenneker@kadaster.nl

Abstract. Kadaster, the Dutch national Land Registry and Mapping Agency, has been actively publishing their base registries as linked (open) spatial data for a number of years. To date, a number of its base registers have been published as linked data and are publicly available. Increasing demand for these services and the availability of new linked data technologies has highlighted the need for new, innovative approach to linked data publication within the organization in order to reduce the time and costs associated with publication. This new approach is novel both in its approach to dataset modelling and architecture implementation and forms part of Kadaster's larger vision for the development of its Knowledge Graph.

Keywords: linked spatial data, knowledge graph, semantic technologies, interoperability, semantic modelling.

1 Introduction[1]

The Dutch Cadastre, Land Registry and Mapping Agency, Kadaster (www.kadaster.nl), is the authoritative source of information relating to administrative and spatial data surrounding property and ownership rights in the Netherlands. Kadaster maintains large registers including several key registers of the Dutch Government such as the Base Register for Addresses and Buildings (Dutch acronym: BAG) and the Base Register Large-Scale Topography (Dutch acronym: BGT); both of which are available as Open Data. The organisation actively publishes and maintains some of its geospatial assets as Linked (Open) Data and, as part of this effort, and in the spirit of continuous innovation, two of these geospatial assets have now been republished as Linked Open Data following a new approach. The intention of this position paper is to outline the problem context which drove the design and im-

[1] Copyright ©2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

plementation of this approach within Kadaster, comment briefly on the effects this has on organisational resources and the outline both the novelty and the architecture that was used to publish these two datasets. The larger vision behind implementation of this approach will also be discussed and illustrated in section 6.

2 Problem Context

Although several of Kadaster's geospatial assets have been available as linked (open) data for a number of years, the network effects of increased uptake in linked data technologies on a broader scale has demanded that an updated, scalable approach to publication be designed. Indeed, increased demand for linked data services is seen both within Kadaster as part of internal innovation processes as well as part of its service delivery to other organisations, both national and European. This demand, coupled with the increasing availability and ongoing development of linked data technologies and standards and the potential this has for smarter, more efficient service delivery, has been the driving force behind the innovation of linked data publication within the organisation.

Guiding the choices made around the technologies, tools and processes chosen for this approach were several key concerns. Firstly, the datasets that were being transformed are complex and the instance data itself voluminous. Based on user requirements of these datasets, this complexity needs to be preserved in any transformation. Secondly, the transformation of these datasets was done in the context of a larger vision (section 6) and time-efficiency was a key concern. This requires that the transformation both happened both quickly and correctly based on a validation technique. Thirdly, this concern for time-efficiency required that, wherever possible, existing tools, libraries and standards were used; a practice in line with general architectural principles. The approach outlined in this paper meets the demands and requirements of this transformation in several ways.

Firstly, this approach makes use of existing community libraries, building on top of open source projects and, therefore, circumventing the need to develop custom, in-house solutions. The use of SHACL, as outlined more extensively below, highlights this reuse of existing community standards. In a similar line, this approach makes use of existing commercial products where they are available in the interest of reducing maintenance costs. Secondly, this approach applies a configuration-over-code principle which ensures that the same pipeline is applied to all linked data publication projects, only configuring components where necessary. Lastly, the implementation of all the relevant components in this design is done with a streaming approach in mind. In practice this means that all linked data models are as close to the source model as possible and that the selected sources are able to support streaming functionality in the interest of real-time data delivery.

In the interest of grounding where the implementation of this approach saw concrete improvements over existing approaches, it is important to note that the BAG and BGT registries were delivered by a small internal team within Kadaster in 9 and 5 weeks, respectively. These are relatively large and complex linked datasets with regards to data model: each containing some 800 million and over 1 billion triples. Where previous approaches could be lengthy, this approach highlights improved cost- and resource-effectiveness, strengthening the business case for linked data within an organisation such as Kadaster [1]. The sections that follow outline the concepts and architecture which support this updated approach to linked data publication within Kadaster; including the standards, technologies and choices made with regard to these during the recent publication of two geospatial datasets.

3 Native Geospatial Sources

There are currently two geospatial assets maintained by Kadaster that have been transformed and published as Linked Open Data using the approach detailed in this paper. Firstly, the BGT was transformed and published in November 2020 and updated in the first quarter of 2021. This asset is a digital map of the Netherlands which includes objects such as buildings, roads, bodies of water and railways. The modelling, updating and maintenance of this dataset is regulated by Dutch law. Secondly, the BAG was transformed and published in February 2021. As the dataset name implies, the dataset includes all buildings and addresses in the Netherlands as well as the attributes associated with these, including house numbers, designations, main- and side addresses. This dataset has a counterpart dataset, namely INSPIRE Addresses[2], that is published based on INSPIRE compliance requirements. Both base registries, including information regarding API availability and querying possibilities, are available in the triple store managed by Kadaster's Data Science Team (https://data.labs.kadaster.nl).

4 Knowledge Model vs. Information Model

The first of the new additions to Kadaster's publication of linked data is the explicit distinction between the Knowledge Model and the Information Model, both composing the larger linked data model. This separation reflects the fact that a linked data model must be able to describe the meaning of the data to the outside world (Knowledge Model), while at the same time describing the organisation-specific aspects (Information Model). This separation allows the Information Model to be optimised towards organisation's internal requirements (including models and processes relating to an asset). At the same time, this also allows the Knowledge Model to be optimised towards efficiently supporting external, community standards of publication required for discoverability and interoperability purposes (W3C, 2014). Since the internal and external aspects are both important for Kadaster's efforts in data publication, this new approach is better able to implement the organisational requirements for linked datasets.

Firstly, an Information Model for a given asset contains the specific and internal information. This information includes the properties of the current information sys-

[2] https://www.pdok.nl/introductie/-/article/adressen-inspire-geharmoniseerd-

tems being used in and around this asset, the organisation-specific rules relating to this asset as well as the asset's technical details. This Information Model is represented using Shapes Constraint Language (SHACL), which serves to constrain a given model based on internally-defined rules and relationships for a given asset. Secondly, a Knowledge Model for the same asset defines any generic and interchangeable knowledge that is both important to retain within the organisation but which should also be shared with others. This Information Model will also makes it easier to reuse external linked data models with an organisation-specific context.

As illustrated in Figure 1A and B, the Information and Knowledge Models are not entirely independent of one another and, indeed, are actually mapped to each other when defining and transforming the model. There are two variants to this mapping process, one being the mapping of datatype properties across the two models and the other being the mapping of object properties across the models. Both variants should be completed over the course of a data model transformation into linked data.

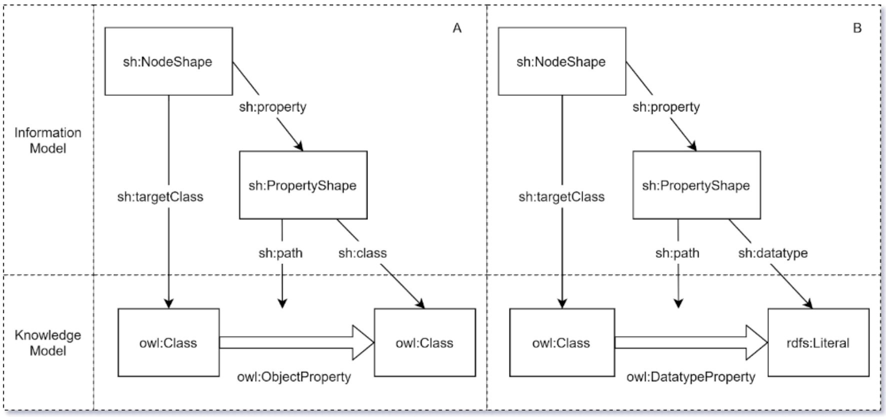

Fig. 1. Model mapping implementation of both object (A) and datatypes (B) properties.

In the first variant (Figure 1A), the process is almost identical except for the fact that SHACL node and property shapes defined for object properties are mapped to the relevant OWL classes and object properties defined in the Knowledge Model. In the second variant (Figure 1B), the datatype properties are defined by mapping the relevant SHACL node and property shapes for each data type to the relevant OWL class, datatype property and RDFS literal defined in the Knowledge Model.

In an effort to support better validation of the resultant model, a key requirement in the design of this approach, a SHACL validation step has also been applied to the modelling process. This step ensures that the shapes for each object and datatype property in the data model completely validates against the instance data and includes a number of self-defined best practices with regards to the modelling of the Information Model centred on the use of closed node shapes. These best practices are based only on experience with the models being transformed in the context of Kadaster but might be points of note for other applications of this approach. This ensures that the Model is both as specific as is necessary to ensure that there is a meaningful validation of the Model while still allowing correct, but rare, data instances to validate.

Firstly, the use of the appropriate Regex is required to ensure that, for example, properties with type string are not returned empty, that display characters allow for non-English standard characters and have a reasonable number of display characters (see XML Schema Datatypes). Secondly, the definition of a label as a node shape (see skos:label) might be needed to ensure that a label is returned in a number of languages such as, for example, where a city has both a Dutch and an English name. Where this is the case, a shape should be closed such that it is mandatory for both languages to appear in the instance data for validation of the model. Lastly, for streaming ETL and validation purposes, a self-contained record should include an extra triple which relates the SHACL path to the SHACL node shape.

5 Design and Development of Supporting Architecture

The process of converting relational data to linked data for an asset is completed in a number of steps during the Extract, Transform and Load (ETL) process. This process is illustrated in the architecture outlined in the figure below (Figure 2). The first step loads the relational data from the source to a PostgreSQL database following a Geography Markup Language (GML) indexing step. A GraphQL endpoint is then used to access the data delivered through API following the delivery and validation of the data model from the end user. In practice, this step is done by extending the typedefs such that the objects in a data model are correctly described in GraphQL, expanding resolvers in order to allow objects to be queried with the right parameters and, finally, adding the required SQL queries to the relevant resolvers. Note that the approach is not inherently limited to relational data sources, as a GraphQL endpoint may also be able to deliver from other source types.

Once the model is available as a GraphQL endpoint, it can be queried by the Enhancer through a configuration process. Firstly, the enhancer has a set of predefined queries with specific time and/or pagination parameters for each object such that the object is delivered as an endpoint that the microservice can access for delivery of the JSON-LD results. Secondly, a reference to the relevant location of the JSON-LD context should be defined. This was done for each new dataset that went through this ETL process. As is probably evident from the validation discussion in the previous section, SHACL can be used both for the validation of the data model using example data but also for validation of the transformed instance data using the data model. Within this architecture, a SHACL validation step is required to ensure the data delivered from the enhancer is valid.

Finally, the approach makes use of Apache Airflow[3] as a ‘handler’ which guides the data through the entire ETL process. The microservice fetches data from the enhancer and repeats until all data is retrieved as JSON-LD. When all data is validated

[3] https://airflow.apache.org/

and loaded into the Triple Store, which in this case is an instance of TriplyDB[4], various services can be instantiated, including ElasticSearch, a data browser, a SPARQL endpoint for use in data stories. These can be instantiated within the interface of the triple store itself. In the interest of better accessibility of the linked data models, the data models for each base registry is also visualised using the Weaver[5] tool.

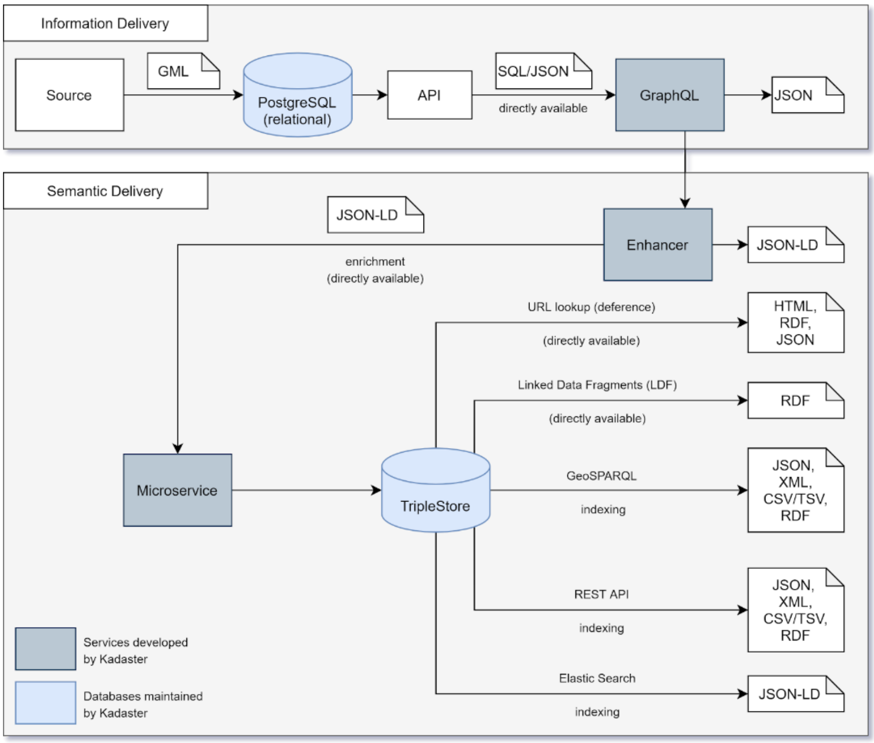

Fig. 2. Architecture supporting the ETL Process which delivers linked (spatial) data.

6 Vision for Geospatial Data Integration

While advancements in linked data technologies and standards, as well as increased demand for these services, initiated the need for an updated approach to Kadaster's delivery of linked geospatial data, this approach is now also at the centre of Kadaster's ambition to deliver a Knowledge Graph (KG) [3]. The contents of the KG are the linked base registries, the digital cadastral map as well as other datasets centred around the theme of a building. The overall vision is to transform these datasets to their linked data versions; keeping the linked data registration as 'close to the source'

[4] https://triplydb.com
[5] https://kadaster.wvr.io/bag2-0/home

as possible in terms of the data model itself while also supporting the improved reuse and findability of Kadaster's geospatial data.

The KG is delivered by creating a further implementation layer on top of these LD registrations based on schema.org specification relating to buildings; for which there are two reasons. Firstly, by layering the vision in this way, provenance of the original datasets is still available to the end user of the KG if necessary. Secondly, making use of the schema.org specifications is done in the interest of reusing existing community standards as well as in the interest of supporting external discoverability and interoperability. Access to the KG is planned to be delivered through REST, GraphQL, GeoSPARQL and ElasticSearch services wherein third party applications make use of these in delivering geoinformation to the end user [4].

7 Conclusion

Kadaster, the Dutch National Land Registry, has recently implemented an updated approach to linked data publication of their geospatial assets in response to growing demand and the pressing need to innovate existing approaches to meet scalability requirements. Building on existing experience with the publication of their base registries as linked data, Kadaster has made use of existing community technologies and standards as well as available commercial products to define an approach which delivers LD assets in a timely, cost-efficient manner and with increased reusability across projects. This approach forms part of a larger vision to deliver a knowledge graph centred around the schema.org 'Building' theme where both this larger vision and the principles applied to this central approach to transformation of the base registries is done in the interest of better geospatial data integration, interoperability and discovery. Although innovative, Kadaster's effort to improve geospatial findability and linkability are not done in isolation and highlight a general need for better spatial interoperability between (European) countries [5] and reusability of this data in various contexts [6].

References

1. Folmer, E., Ronzhin, S., Van Hillegersberg, J., Beek, W., Lemmens, R. Business Rationale for Linked Data at Governments: A Case Study at the Netherlands' Kadaster Data Platform. IEEE. Access 8, 70822-70835, (2020).
2. World Wide Web Consortium. Best Practices for Publishing Linked data. W3C Working Group Note. http://hdl.handle.net/10421/7479. (2014)
3. Ronzhin, S., Folmer, E., Lemmens, R., Mellum, R., von Brasch, T. E., Martin, E., Romero, E. L., Kytö, S., Hietanen, E., Latvala, P. Next generation of spatial data infrastructure: lessons from linked data implementations across Europe. International journal of Spatial Data Infrastructures Research, 14. 83-107. (2019).
4. Rowland, A., Folmer, E., Beek, W. Towards Self-Service GIS-Combining the Best of the Semantic Web and GIS. ISPRS International Journal of Geo-Information 9(12). (2020).

5. Ronzhin, S., Folmer, E., Maria, P., Brattinga, M., Beek, W., Lemmens, R., van't Heer, R. Kadaster Knowledge Graph: Beyond the Fifth Star of Open Data. Information, 10(10). (2019).
6. Bucher, B., Folmer, E., Brennan, R., Beek, W., Hbeich, E., Würriehausen, F., Rowland, L., Maturana, R. A., Alvarado, E., Buyle, R. Spatial Linked Data in Europe: Report from Spatial Linked Data Session at Knowledge Graph in Action. (2021).