information

MDPI

Article

Kadaster Knowledge Graph: Beyond the Fifth Star of Open Data

Stanislav Ronzhin [1,*], Erwin Folmer [2,3], Pano Maria [3], Marco Brattinga [3], Wouter Beek [4], Rob Lemmens [1] and Rein van't Veer [3]

1 Faculty of Geo-Information Science and Earth Observation, University of Twente, 7514 AE Enschede, The Netherlands; r.l.g.lemmens@utwente.nl
2 Behavioral, Management and Social Sciences, University of Twente, 7522 NH Enschede, The Netherlands; erwin.folmer@utwente.nl
3 Kadaster Dataplatform, Kadaster, 7311 KZ Apeldoorn, The Netherlands; Pano.Maria@kadaster.nl (P.M.); marco.brattinga@ordina.nl (M.B.); rein.veer@kadaster.nl (R.v.V.)
4 Knowledge Representation and Reasoning Group, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands; w.g.j.beek@vu.nl
* Correspondence: s.ronzhin@utwente.nl

Received: 23 August 2019; Accepted: 26 September 2019; Published: 9 October 2019

Abstract: After more than a decade, the supply-driven approach to publishing public (open) data has resulted in an ever-growing number of data silos. Hundreds of thousands of datasets have been catalogued and can be accessed at data portals at different administrative levels. However, usually, users do not think in terms of datasets when they search for information. Instead, they are interested in information that is most likely scattered across several datasets. In the world of proprietary in-company data, organizations invest heavily in connecting data in knowledge graphs and/or store data in data lakes with the intention of having an integrated view of the data for analysis. With the rise of machine learning, it is a common belief that governments can improve their services, for example, by allowing citizens to get answers related to government information from virtual assistants like Alexa or Siri. To provide high-quality answers, these systems need to be fed with knowledge graphs. In this paper, we share our experience of constructing and using the first open government knowledge graph in the Netherlands. Based on the developed demonstrators, we elaborate on the value of having such a graph and demonstrate its use in the context of improved data browsing, multicriteria analysis for urban planning, and the development of location-aware chat bots.

Keywords: linked data; knowledge graph; semantic enrichment; location-aware chat bots; governmental open data

1. Introduction

Coined by Google in 2012 [1], the term knowledge graph (KG), in a broad sense, refers to a graph-based representation of general world knowledge. Although Gartner recognized KGs as an emerging technology climbing the slope of the hype cycle in 2018 [2], the idea to represent knowledge in the form of a graph, where entities are modelled as nodes and the relations between them as edges, in fact, dates back to the early age of computer science. The novelty, however, is in the fact that the data and computing power are now available to make KGs work at scale, which allows us to go beyond a keyword search paradigm in information search and retrieval—"things, not strings" as Google put it [1]. By 2019, KGs were inside every search engine and every speech assistant (e.g., Siri, Alexa, and Cortana). Companies like Amazon, Uber [3] Airbnb [4], Reuters [5], Elsevier, Zalando [6], Blumberg [7], and Siemens [8] are building their KGs or have them in place.

It seems to be a strategic goal for governments to have an intelligent system or an agent that is able to answer the question, "Can I build a shed in my backyard, and if not, what additional requirements do I need to meet?" Apart from question answering, KGs are used in building recommender systems [9] or for image classification [10]. Therefore, expressing their data in a graph-based semantically rich format, such as the Resource Description Framework (RDF) [11], has grown into an almost mainstream activity around the world on many bureaucratic levels, from local to national and international.

Making public data available in the RDF is an important prerequisite for building KGs in governments; however, it does not enable a seamless out of the box reasoning over these data. By taking down the technical barriers between the data silos, it exposes the knowledge gaps between divisions of the government. Therefore, in order to build a KG, these gaps need to be bridged, which is not trivial.

On the one hand, bridging gaps requires their identification. However, how do we identify these gaps if they lie in areas outside of departmental knowledge? This is a chicken-and-egg problem; gaps cannot be identified upfront before constructing a KG, which in turn cannot be created without an identification of the gaps. On the other hand, in contrast to Google, governmental data is used for making legal decisions. This puts additional requirements on the accuracy of the semantic relations between data items that go well beyond the capabilities of "owl:sameAs" [12].

In this paper, we present and discuss our experience of building a KG within Kadaster. We showcase its value by presenting three real-life applications of the graph, namely, improved data browsing, multicriteria analysis for urban planning, and the development of location-aware chat bots. Section 2 briefly introduces the mission of Kadaster and the Kadaster main data assets. The main concepts used in the paper and the related developments are presented and discussed in Section 3, followed by a description of the approach taken at Kadaster in relation to the development of the Kadaster knowledge graph (KKG) in Section 4. Section 5 presents the resulting KG, and Section 6 presents use cases for the graph. Discussion and conclusions are given in Sections 7 and 8, respectively.

2. Kadaster: Context and Data

The Netherlands' Kadaster Land Registry and Mapping Agency (https://www.kadaster.nl/), in short, collects and registers administrative and spatial data on property and the rights involved in ownership. This also includes data on ships, aircraft, and telecom networks. In doing so, Kadaster protects legal certainty.

Kadaster publishes many large authoritative datasets, including several key registers of the Dutch Government (e.g., the Base register of topography (Dutch acronym: BRT) [13], and the Base register of addresses and buildings (Dutch acronym: BAG) [14]). Furthermore, Kadaster is also developing and maintaining "Publieke Dienstverlening op de Kaart" (Dutch acronym: PDOK). PDOK is a web portal where more than 150 spatial datasets coming from different Dutch government organizations are published in several formats.

These data include an incredible number of geospatial objects. These objects are spatially and/or conceptually related but are maintained by different data curators. As a result, these datasets are syntactically and structurally disjoint, and, currently, it requires non-trivial human labor to use them together. For these reasons, Kadaster have made an effort to integrate these data assets into a KG.

3. Open Data in Graphs

Even though there is not a common definition of a KG [15], the term itself was coined by Google in 2012 when referring to a new Web search strategy. In that strategy, Google announced moving from a pure keyword-based search paradigm to a graph representation of knowledge, and phrased it with the slogan "Things, not strings". The authors of [16] and [17] concluded that KGs can be understood as collections of interlinked linked data (LD) sets covering various topical domains. Thus, KGs are a by-product that have emerged as a result of integrating the five-star data from more than one knowledge domain. Therefore, in the following subsections, we start with the overview of the

five-star model of open data, followed by an introduction to the linked data (LD) technology. At the end of the section, there is an overview of the existing KGs, which focusses on their geospatial content.

3.1. Five Stars of Open Data

The five-star model (https://5stardata.info/en/) of open data is often used to classify the technical level of the advancement of a dataset offering (Figure 1). In the model, the first star requires an open license, but without requirements on the data format (e.g., a handwritten document stored in Portable Document Format (PDF) with an open license qualifies as one-star open data). The second star adds to the open license the ability to include structured data (e.g., the Microsoft Excel format (XLS)). When this proprietary format is replaced with an open format, such as comma-separated values (CSV) or eXtensible Markup Language (XML), the dataset receives three stars. The fourth and fifth star require the use of the LD principles [18]. A dataset published in compliance with RDF receives the fourth star. Finally, the fifth star is assigned when the content of a dataset is linked to other RDF-based resources on the Web. The model is meant to advertise the LD technology as an ultimate solution for data reusability.

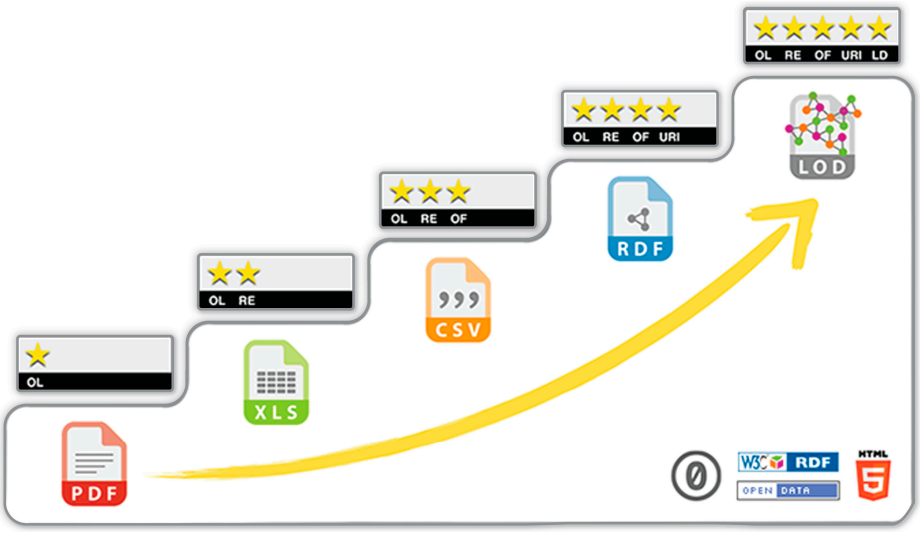


Figure 1. The five-star deployment scheme for open data. A handwritten document published on the Web in Portable Document Format (PDF) with an open license is one-star open data. The second star is given when data is published in a structured way (e.g., the Microsoft Excel format (XLS)). When a proprietary format is replaced with an open format, such as comma-separated values (CSV) or eXtensible Markup Language (XML), the dataset receives three stars. The fourth star data require the use of the LD principles [18]. The fifth star is given when the content of a dataset is linked to other RDF-based resources on the Web. (Source https://5stardata.info/en/).

3.2. Linked Data

The LD initiative [18] promotes the use of semantic standards for representing and publishing information on the Web at the data level. This implies that each data element and attribute is individually recognizable, retrievable, and combinable.

This can be achieved by encoding information using the Resource Description Framework (RDF) [11]. This standard is based on mature technologies, namely, the graph data model [19] and the Hypertext Transfer Protocol (HTTP) [20]. The former allows instances and concepts, represented by nodes, to be related to one another by relationships, represented by arcs between the nodes. By adding HTTP Universal Resource Identifiers (URIs) to these data elements (nodes and arcs), they became globally accessible, referenceable, and queryable by the means of the SPARQL Protocol and RDF Query Language (SPARQL) [21].

Many data suppliers, especially ones that publish official government data, such as national mapping agencies, are diving into the world of LD, as they see potential for their authoritative data [22]. Ordnance Survey, the national mapping agency of Great Britain, was one of the first big governmental organizations who pioneered exposing public geospatial data on the Web as LD in 2008 [23]. Even though this was a state-of-the-art development at that time, it relied on unstandardized means for representing data semantics and, as a result, lacked (re)usability. Ordnance Survey Ireland considered the experience of their British colleagues and used standard vocabularies to publish the boundary data of the administrative division at various levels, and to capture the evolution of the administrative boundaries [24]. Another prominent example is the work of the National Geographic Institute of Spain (IGN-E) [25], where they combined the data coming from two governmental institutions and published it as a coherent LD dataset. In the context of Infrastructure for spatial information in Europe (INSPIRE), LD showed potential for the provision of access to European-wide geospatial data [26].

3.3. Knowledge Graphs

Even though there is more than one definition of what a KG is [15], we can name some common properties. Based on the literature [16,17,27], we can conclude the following about a KG:

1. It is a graph composed of statements about the real world.
2. It has both instances and schema.
3. It covers more than one knowledge domain.

The first property clearly indicates that KGs can be built with LD as the guiding principle. KGs include instance-level statements (e.g., things) and statements about the background knowledge (ontologies) needed to understand the meaning of instances. Usually, the volume of instance level information is several orders of magnitude larger than that of the schema level [17]. The reusability of the latter plays a major role in the context of the third property of KG; reusability is required in respect to integrating, dereferencing, and disambiguating across domain knowledge [16].

The construction of KGs can be approached differently. Examples of curated efforts are the oldest KG, Cyc, and its public version OpenCyc, built in the domain of artificial intelligence, as well as WordNet [28], a lexical database that is widely used as a semantic network and as an ontology. With the rise of crowdsourcing, many open KGs were created and maintained collectively online. In this context, Wikidata [29], a collaboratively built KG, operated by the Wikimedia [30] foundation, is an effort to interlink major open sources of structured data such as Geonames [31] and DBpedia [32]. The latter is a KG that was semi-automatedly extracted from Wikipedia. Yet Another Great Ontology (YAGO) is another example of a large knowledge base that was automatically extracted from Wikipedia and WordNet [33].

The above-mentioned graphs, as well as many other LD sets, are available online. They can be accessed and queried in a federated manner. The Linked Open Data (LOD) cloud [34] catalogs LD resources available under open licenses. Even though the interconnectedness of the cloud significantly varies between sets, it contains 1239 datasets with 16,147 links (as of March 2019), which makes it one of the largest distributed KGs, encompassing almost every domain. Figure 2 depicts geo-LOD, the sub cloud of graphs with a strong geographical component. In the figure, the size of the nodes denotes the number of statements the dataset has. As can be seen from the figure, DBpedia and GeoNames are the most interlinked resources. Linked GeoData, an LD version of OpenStreetMap [35], was an ambitious and very promising project to add a spatial dimension to the Web of Data. However, it has not been updated since 2015 (as of March 2019).

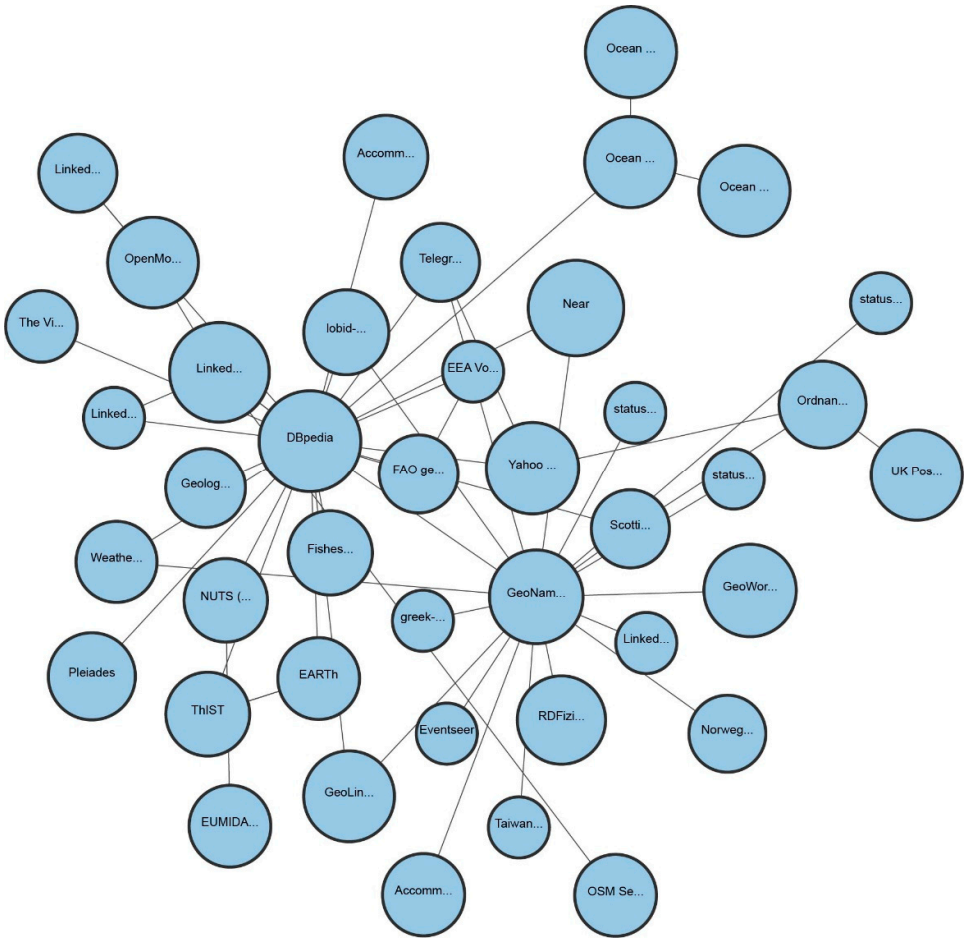


Figure 2. Geo-linked open data (LOD)—sub-cloud of gegraphical resources known by the LOD cloud. DBpedia and GeoNames are the central and most interlinked resources (source: https://lod-cloud.net/).

4. Building the Graph

There is a clear distinction and a big implementation gap between the first three and the last two stars. Most datasets reside at the first three stars, and for many datasets, three stars is the endpoint. The implementation of LD requires several considerations that need to be made. These considerations are of mixed nature and are discussed in this section.

4.1. Moving from the Third Star

Let us imagine that Kadaster registered an object, a building of the Saint Catharine church, erected in 1900 in Eindhoven, with a certain registration ID. Figure 3 depicts this as a plain text at the top of the figure. This information can be decomposed into the following three facts: (1) the object is a church, (2) it has a name, and (3) it was erected in 1900. In Figure 3, these facts are shown as a graph with green rectangles as nodes and arrows as the relations between them.

In Figure 3, the blue circles and rectangles represent the same graph but with all of the arbitrary wording replaced by standardized notions and their URIs. The notions within a collection share the same namespace and are often abbreviated. For instance, in Figure 3 "rdf" is a namespace prefix for the basic RDF vocabulary (http://www.w3.org/1999/02/22-rdf-syntax-ns#). If there is a URI to represent a concept (e.g., bag:AddressableObject), it is depicted as a circle; the literal values are shown as rectangles. This is done to emphasize that only URIs can be linked.

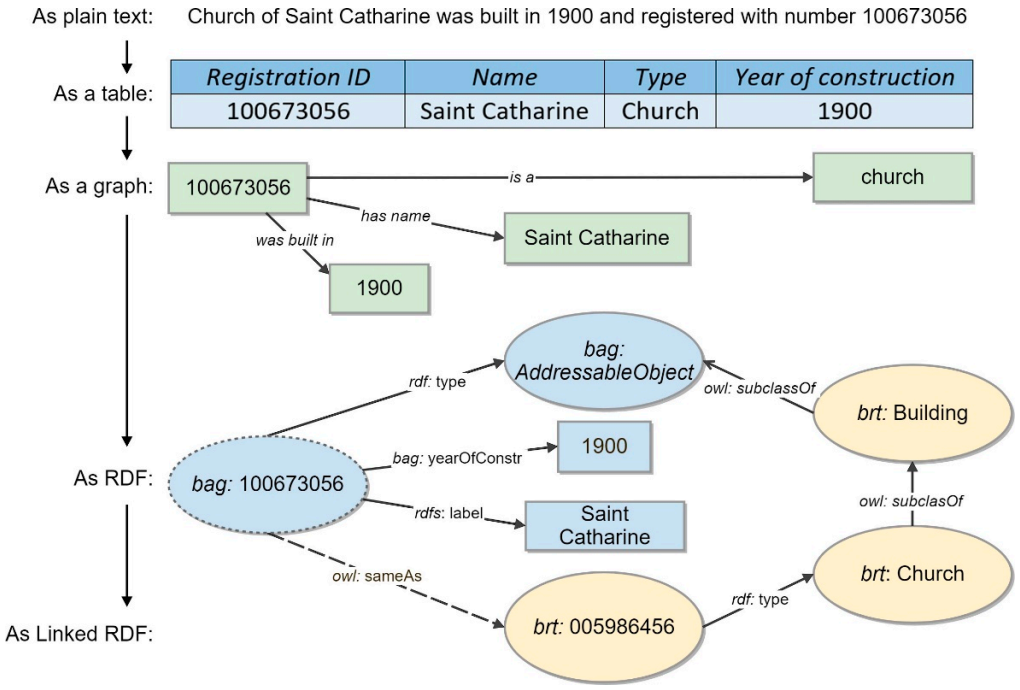


Figure 3. Representing facts from a plain text using the Resource Description Framework (RDF). Green rectangles represent a graph made by decomposing a plain text into facts with arbitrary wording. Blue shapes represent the same graph but expressed with standardized vocabularies and URIs. Yellow shapes represent data items from another dataset (BRT), which are linked to the blue ones (BAG), forming a part of a knowledge graph (KG; source: authors).

The standardization of semantic descriptions and the use of URIs allows for linking data items between datasets, which is a major advantage of four-star data. However, even though The World Wide Web Consortium (W3C) provides actionable cookbooks [36,37] on how to create and publish LD, the designing of vocabularies and URI-patterns requires an understanding of the knowledge domain and intendent use.

In practice, a possible approach would be to narrow down the use case and build the LD to meet the case requirements. This approach was successful for the Ordnance Survey of both Ireland and the United Kingdom [23,24].

We took another approach at Kadaster. Having started publishing LD in 2017, Kadaster pursued its business ambition of having a KG. This was the main goal of creating an LD, and therefore, the intended use case was defined in a generic way. In this context, the main focus was on generating three-star open data from the existing resources as soon as possible. For this reason, it was decided to derive ontologies from the base registry data models. The general transformation rules were defined [38] in order to transform the existing Universal Modeling Language (UML) models to their RDF counterparts, using RDF, the Simple Knowledge Organization System (SKOS), the Web Ontology Language (OWL), and the Shapes Constraint Language (SHACL) [39]. This allowed for the fast prototyping of the three base registers (BAG, BRT, and BRK), consequently linking them.

4.2. Five-Star Data: Linking

The Saint Catharine church in Figure 3 is a building that appears in many datasets. In Figure 3, a dashed arrow represents a relation (*owl:sameAs*) between two representations of the same church in BAG (blue shapes) and in BRT (yellow shapes). Despite the fact the building is classified differently in these graphs (as an addressable object in BAG and as a church in BRT), by linking them together, we can infer additional knowledge (e.g., that a church is an addressable object). In this way, previously disconnected datasets can be linked together via persistent URIs to form five-star open data.

However, to achieve this, somebody has to create the links. On the one hand, the government owns and controls the systems of legal definitions. These systems are often hierarchical, and their structure can be traced from the top-down to identify the precise meaning of the relations between the concepts. On the other hand, this approach does not help in defining instance-level relations, because their numbers and complexity grow very fast with every instance added to a KG. The network effect makes it difficult to foresee and formalise all of the possible relations. Instead, a bottom-up use-case driven approach should be used.

Naturally, Kadaster data are rich with respect to spatial and temporal information. Space and time are fundamental sources of contextual information, and therefore, they allow for linking data instances that lack explicit ontological relations. This is especially relevant in cases when the top-down approach is hindered (or even not possible) because of the existing semantic heterogeneity of the legal definitions and terminology between independent governmental agencies. In this context, the strong spatiotemporal component of Kadaster data is seen as an important competitive advantage [40]. It provides the information dimensions needed for the interrelating data that have very little in common otherwise. In this context, the development of the GeoSPARQL standard [41] by the Open Geospatial Consortium (OGC; http://www.opengeospatial.org/) provided building blocks to formalise the geospatial semantics in the data and enabled geospatial reasoning in SPARQL.

Knowledge about data collection is another source of links. Examples are lookup tables that interrelate object IDs between public data bases. For example, Statistics Netherlands (Dutch: Centraal Bureau voor de Statistiek (CBS)) collects and publishes statistical information about Dutch neighborhoods. However, neighborhoods are not registered by Kadaster, therefore, CBS maintains a look-up table that links municipalities (official administration units maintained by Kadaster) with CBS neighborhoods.

Recent developments in machine learning (e.g., see [42,43]) propose approaches for scalable link predictions in complex KGs [44]. Even though these algorithms perform well on general purpose data, and they can be used, for example, in recommendation systems [9], none of them provide reliable solutions for link prediction based on geospatial information, especially in the context of multi-scale vector geometries.

5. Kadaster Knowledge Graph

The Kadaster Knowledge Graph (KKG) project was aimed at building a first version of a KG and presenting use cases in order to convince management and the owners of datasets about the value of a KG. The scope of the datasets was arbitrarily chosen based on a combination of a pragmatic approach (e.g., datasets already available as linkable "fourth star" data) and a use-case driven approach, where datasets were required by defined use cases.

The KKG was built around the three open base registers from Kadaster and extended with datasets from outside Kadaster. The first version included almost 2 billion triples coming from 12 datasets (see Figure 4), curated by eight organizations and accessible from six data endpoints.

5.1. Data Sources

The datasets used in the KKG are from different categories. The first category contains datasets that are officially released as LD. Table 1 summarizes these datasets, giving details about the approximate size of the sets, their providers, and links for access.

Table 1. Kadaster Knowledge Graph (KKG) data sources that were officially published as linked data (LD). (Source: authors).

	English Name (Dutch Name)	Number of Statements	Web Link	Data Owner (Dutch Name)
1	Base register of addresses and buildings (Basisregistratie adressen en gebouwen (BAG))	~1,000,000,000	bag.basisregistraties.overheid.nl	Kadaster
2	Base register of topography (Basisregistratie topografie (BRT))	~300,000,000	brt.basisregistraties.overheid.nl	Kadaster
3	Base land register (Basisregistratie Kadaster (BRK))	~400,000,000	brk.basisregistraties.overheid.nl	Kadaster
4	Key figures districts and neighborhoods (Kerncijfers wijken en buurten (KWB))	~10,000,000	betalinkeddata.cbs.nl [1]	Statistics Netherlands (Centraal Bureau voor de Statistiek (CBS))
5	Government Web Metadata Standard *(Overheid Web Metadata Standaard (OWMS))*	~10,000	standaarden.overheid.nl/owms/terms	Centre for Official Publications (Kennis - en Exploitatiecentrum voor Officiële Overheidspublicaties (KOOP))
6	Basic geo-information model (Basismodel geo-informatie (NEN3610))	~1000	geonovum.github.io/NEN3610-Linkeddata/	Geonovum

[1] Also available from https://data.pldn.nl/cbs/wijken-buurten.

The second category is presented in Table 2 and consists of datasets that are published in the beta stage and are expected to be released for the public as official LD soon.

Table 2. KKG data sources that are published in the beta stage. (Source: authors).

	English Name (Dutch Name)	Number of Statements	Web Link	Data Owner (Dutch Name)
7	Spatial planning (Ruimtelijke ordening)	~1,000,000	under construction [1]	Spatial Information Warehouse (Informatiehuis Ruimte)
8	Cultural heritage (Cultureel erfgoed)	~65,000,000	linkeddata.cultureelerfgoed.nl	Cultural Heritage Agency (Rijksdienst voor Cultureel Erfgoed (RCE))

[1] The dataset is not published for the public yet (as of June 2019) but will be available from https://www.pdok.nl/introductie/-/article/ruimtelijke-plannen.

The third category contains datasets published in the experimental environment of Kadaster. See Table 3 for more details.

Table 3. KKG data sources published in the experimental environment of Kadaster. (Source: authors).

	English Name (*Dutch Name*)	Number of Statements	Web Link	Data Owner (*Dutch Name*)
9	Energy labels of buildings in Dordrecht (Dordrecht woning energielabels)	~500,000	data.labs.kadaster.nl/kadaster/energielabels	The Netherlands Enterprise Agency (Rijksdienst voor Ondernemend Nederland (RVO))
10	Base-register of real estate values (Waardering Onroerende Zaken (WOZ))	~160,000,000	data.labs.kadaster.nl/kadaster/woz	Council for Real Estate Assessment (Waarderingskamer)

Finally, two specific linksets were created linking the Base register of addresses and buildings, with two others base registers namely, the Base register of topography (BRT) and Base land register (BRK), based on the spatial relations between the datasets. These linksets were also published in the experimental environment of Kadaster. See Table 4 for more details.

Table 4. Linksets within KKG that link three base registers, namely, BAG, BRT, and BRK. (Source: authors).

	English Name (Dutch Name)	Number of Statements	Web Link	Data Owner (Dutch Name)
11	Linkset BAG–BRK	~11,000,000	data.labs.kadaster.nl/kadaster/bag-brk	Kadaster
12	Linkset BAG–BRT	~10,000,000	under construction [1]	Kadaster

[1] The dataset is not published for the public yet (as of June 2019) but will be available from https://www.data.labs.kadaster.nl/kadaster/bag-brt.

5.2. The Graph

The data sources listed in the previous section comprised the first version of the KKG. Figure 4 provides a network diagram representing the main concepts (rounded rectangles) of KKG and the relations (arrows) between them. Colors denote the datasets from where the classes are originated.

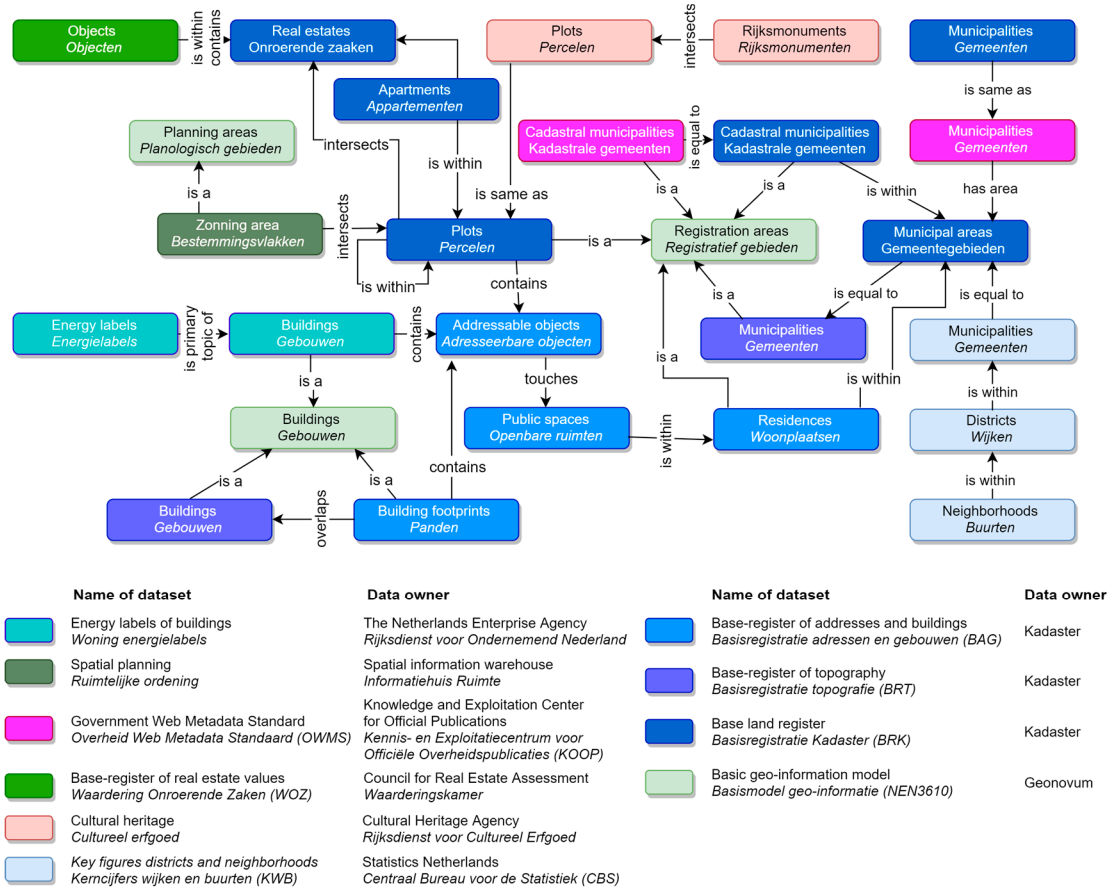
As can be seen from Figure 4, the Kadaster data is in the core of the graph, comprising 12 classes out of a total of 26. The relations and their types are summarized in Table 5. There are 10 different relations used to interrelate concepts of KKG with spatial relations being prevalent (6 out of 10).

Table 5. Relations between concepts of KKG and their types. (Source: authors).

	Relation Name	RDF Term	Relation Type
1	*is a*	rdf:type	thematic relation
2	*is same as*	owl:sameAs	thematic relation
3	*is primary topic of*	foaf:primaryTopic	thematic relation
4	*has area*	bbi:heeftGebeid	thematic relation
5	*is within*	ogc:sfWithin	spatial relation
6	*contains*	ogc:sfContains	spatial relation
7	*overlaps*	ogc:sfOverlaps	spatial relation
8	*touches*	ogc:sfTouches	spatial relation
9	*intersects*	ogc:sfIntersects	spatial relation
10	*is equal to*	ogc:sfEquals	spatial relation

The identification and generation of the thematic links were based on two types of sources, namely, the existing lookup tables (e.g., *same as* links between different representations of municipalities) and

the NEN 3610 standard (e.g., most of the *is a* relations) containing the base model for geo-information in the Netherlands.



Name of dataset	Data owner
Energy labels of buildings *Woning energielabels*	The Netherlands Enterprise Agency *Rijksdienst voor Ondernemend Nederland*
Spatial planning *Ruimtelijke ordening*	Spatial information warehouse *Informatiehuis Ruimte*
Government Web Metadata Standard *Overheid Web Metadata Standaard (OWMS)*	Knowledge and Exploitation Center for Official Publications *Kennis- en Exploitatiecentrum voor Officiële Overheidspublicaties (KOOP)*
Base-register of real estate values *Waardering Onroerende Zaken (WOZ)*	Council for Real Estate Assessment *Waarderingskamer*
Cultural heritage *Cultureel erfgoed*	Cultural Heritage Agency *Rijksdienst voor Cultureel Erfgoed*
Key figures districts and neighborhoods *Kerncijfers wijken en buurten (KWB)*	Statistics Netherlands *Centraal Bureau voor de Statistiek (CBS)*

Name of dataset	Data owner
Base-register of addresses and buildings *Basisregistratie adressen en gebouwen (BAG)*	Kadaster
Base-register of topography *Basisregistratie topografie (BRT)*	Kadaster
Base land register *Basisregistratie Kadaster (BRK)*	Kadaster
Basic geo-information model *Basismodel geo-informatie (NEN3610)*	Geonovum

Figure 4. Network diagram of classes and relations comprising the KKG. Rounded rectangles represent classes with colors representing origin of concepts. (Source: authors).

The construction of spatial relations was done based on the topological analyses of the geospatial features. The relations used to express the topological relations came from the GeoSPARQL vocabulary.

The resultant KG was enriched with outgoing links (approximately 25,000; see https://lod-cloud.net/dataset/bag) to the LOD cloud resources. Moreover, we introduced a dedicated property (http://www.wikidata.org/prop/direct/P5208) to the Wikidata project to stimulate the creation of incoming links (approximately 11,000 links (As of August 2019)) by the community.

6. Use Cases

In this section, we present three applications developed to demonstrate the advantages of the KKG in relation to the improved data browsing (Case 1), multicriteria analysis for urban planning (Case 2), and the development of location-aware chat bots (Case 3).

6.1. Case 1. Data Browsing: Follow Your Nose

Spanning across several knowledge domains, KGs are difficult for comprehending, as they often feature thousands of interlinked concepts used to capture billions of data elements. In this context, an exploratory search [45] is more appropriate for navigating large infrastructures of highly heterogeneous data than classical information retrieval and querying [46].

Simply put, users are not able to formulate a query because of the lack of knowledge about the underlying concepts. The solution to this problem is to allow users to discover concepts and facts in a so-called "follow your nose" manner. In this approach, users learn the concepts by traversing the graph,

and at the same time, they narrow down the search using the learned concepts. Moreover, KGs contain different types of information (e.g., qualitative, quantitative, and location information); therefore, users should be able to use visualisation techniques that fit the type of information, as discussed in the literature [47].

Figure 5 and 6 depict three different panes of the KKG browsing application [48,49] built with the open source Linked Data Theatre toolkit [50]. The graph depicted in Figure 4 can be represented for the purpose of exploration in a graph browser, as in Figure 5. Colors represent the different organizations responsible for data curation. Figure 5 shows a part of the graph containing information related to a certain building (a historical moment) that can be discovered traversing the links. This information includes the related cadastral parcel (blue circle), spatial planning context (green circles), the address and building information (purple), taxation value (orange), and energy label (red).

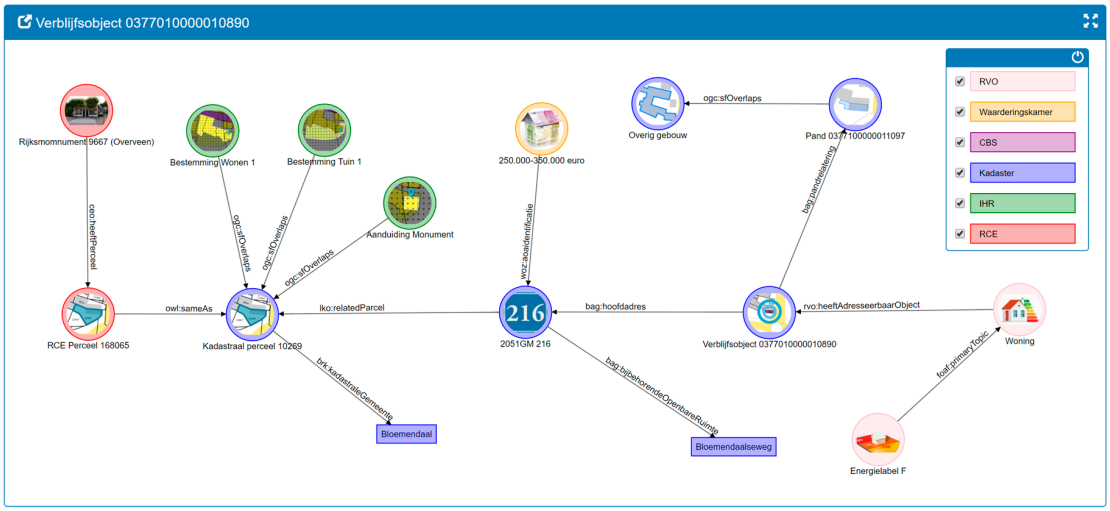


Figure 5. Visualization of a part of the KKG using the Linked Data Theater software. Colors represent organizations curating the data (source: authors).

The graph browser presented in Figure 5 is useful for discovering the relations between the data objects. However, each of the objects, in turn, has their own data attributes. Users can discover them from two dedicated views (tabular and map; see Figure 6), which are linked to the graph browser's view. Figure 6A shows the detailed information about the monument in a tabular view, and 6B provides the location information using the map view.

6.2. Case 2. Urban Planning: Candidate Areas for Urban Development

Urban planning as an interdisciplinary field of study deals with the multicriteria analysis of heterogeneous data sources. Current workflows imply the use of desktop geoinformation systems (GIS; e.g., ArcGIS) to perform the integration of various datasets owned by independent branches of the government. In order to simplify the integration, a research question is commonly reformulated by humans into several sub questions in such a way that each of the sub questions can be answered with the least number of datasets. Eventually, the main answer is synthesized from the answers to sub-questions.

The KKG provides an integrated view over previously disjoint data sources, enabling urban planners to perform an analysis formulating a single data query with all of the needed criteria. As an illustration of this approach, we provide an example of the identification of areas suitable for urban development.

Candidate areas must meet certain criteria, as given in Table 6. Candidate neighborhoods must have houses with a low average price, that were built before 1970, and have low energy labels. These data come from three different data curators, namely, Council for Real Estate Assessment

(Waarderingskamer), Kadaster, and the Netherlands Enterprise Agency (Rijksdienst voor Ondernemend Nederland (RVO)) (see Table 6).

Table 6. Criteria for identification of areas suitable for urban development. (Source: authors).

Criteria	Threshold	Data Provider
Average tax value of a house	<150,000 €	Council for Real Estate Assessment (Waarderingskamer)
Year of construction	Before 1970	Kadaster
Energy efficiency	D (or lower)	The Netherlands Enterprise Agency (Rijksdienst voor Ondernemend Nederland (RVO))

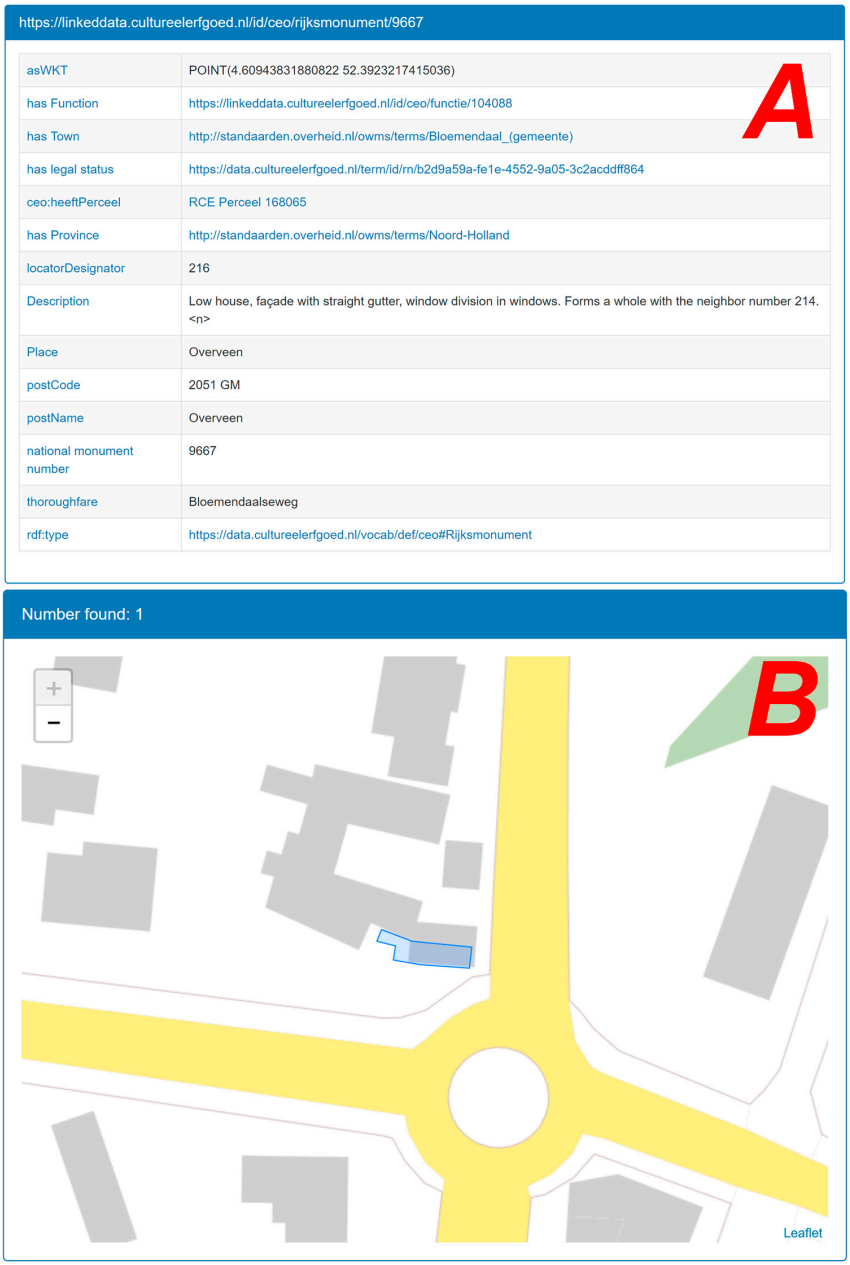


Figure 6. Table (**A**) and map (**B**) view panes of the KKG browsing application. The information presented in the table view (**A**) comes from the cultural heritage datasets curated by RCE, while the location information of the monument presented on a map is originated from BAG, curated by Kadaster (source: authors).

Based on this combination of data, we can find areas within a city that might be eligible for new city developments. With KKG, we can create one SPARQL query and get the results for the municipality of Dordrecht in 24 s, utilizing three different data endpoints (Kadaster, RVO, and Council for Real Estate Assessment (Waarderingskamer); (https://labs.kadaster.nl/stories/pdok-knowledge-graph/).

For visualisation, Yet Another SPARQL Graphical User Interface (YASGUI) (http://yasgui.org/) is used as a query visualisation tool, which offers many options, including maps (3D, heatmaps), tables, and google charts. The outcome, the potential areas for new city development in Dordrecht, are shown in Figure 7.

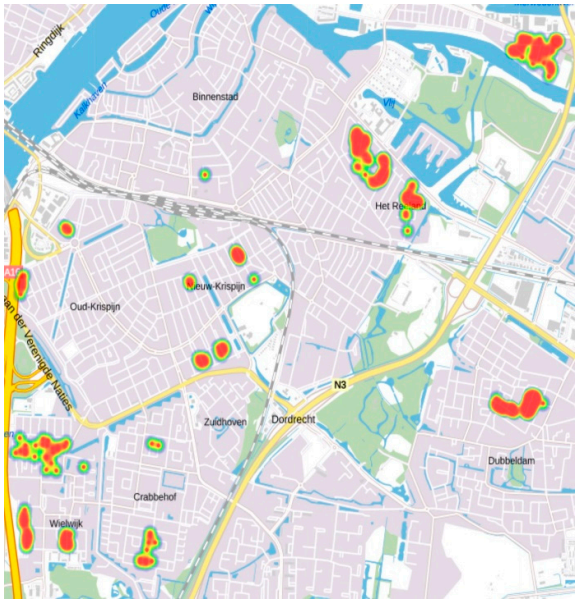


Figure 7. Visualization of the query results on a map showing potential areas for urban development (Source: authors).

6.3. Case 3. Loki: Chatbot for Spatial Questions

Loki (https://labs.kadaster.nl/cases/loki), which stands for location-based Kadaster information (Dutch: Locatie-gebaseerde Kadaster Informatieverstrekking (Loki)), is a proof-of-concept developed to demonstrate the potential of a chatbot in meeting the public demand for approachable geospatial information. For non-experts, the understanding of Kadaster data is difficult, as it requires extensive knowledge from two domains, namely geoinformation and land administration. However, in the modern digital world, citizens increasingly want to know more about their properties, neighbourhoods, or about the Netherlands as a whole. Such users do not know the specific terminology of basic registries—they want to be able to express their information needs in layman terms. This is the target user group and the use case for Loki that uses natural language processing (NLP) to answer user questions with the KKG.

The script given in Listing 1 follows the default scenario, where a user is prompted to specify an object of interest through a series of questions about its address (lines 1–4). Once the object is defined, Loki asks the user to formulate a question about the object of interest.

Line 1 Loki: What is your zipcode?
Line 2 User: 7521 DA
Line 3 Loki: Which house number (including house letter and/or other additions)?
Line 4 User: 135
Line 5 Loki: Ask a Question!
Line 6 User: what is the year of construction?
Line 7 Loki: Also view this address as Linked Open Data (LOD) object http://bag.basisregistraties.overheid.nl/bag/id/pand/0153100000235369
Line 7 Loki: Hint: You can also visualize these results in the BAG Viewer https://bagviewer.kadaster.nl/lvbag/bag-viewer/#?searchQuery = Schietbaanweg 135, Enschede&geometry.x = 255634.59762421434&geometry.y = 472500.0407453137&zoomlevel = 5
Line 8 Loki: The answer is 1953!
Line 9 Loki: Feel free to ask another question

Listing 1. A fragment of a conversation with the location-based Kadaster information (Loki) in which a user learns the year of construction of a certain building. The user inputs a building zip code (line 2) and a house number (line 4) in response to Loki's questions (lines 1 and 3; source: authors)

In contrast to conventional chatbots that have only a chat pane, Loki's interface features a map view, because it deals with geospatial information. Figure 8 provides a screen capture of the conversation given in Listing 1 (lines 6–9). As can be seen from Figure 8, the user interacts with a chat bot using the chat pane on the right part of the screen. The visualisation of the location information in the map assists users, providing an appropriate view to minimise the ambiguity of the conversation.

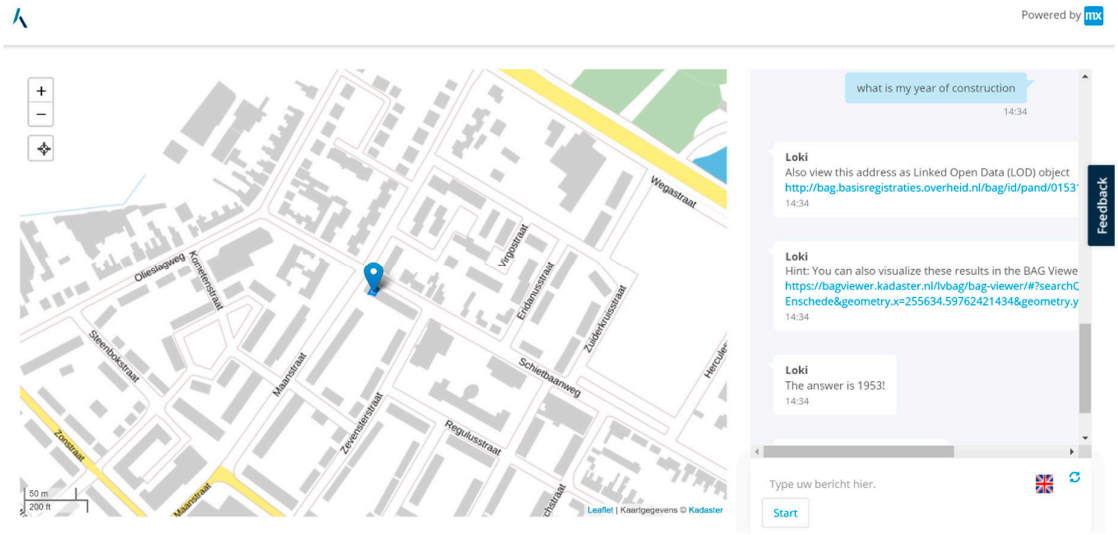


Figure 8. Interface of Loki consisting of a chat pane (on the right) and a map view pane (on the left; source: authors).

In the current version, Loki can answer questions based on the information coming from the BAG, WOZ, BRK, and BRT graphs. Therefore, users can formulate several types of requests, as follows:

- What is the real-estate value of a house?
- What is the year of construction of a house?
- What is the average area of the houses in a street?
- Where is my plot?
- What are the houses in Oranje that were built after 2000?
- What is the oldest house in Haarlem?
- Give me all of the churches built before 1500 in Dordrecht.

Figure 9 illustrates the infrastructure of Loki. The Web interface is available on https://labs.kadaster.nl/. The user input is processed by the chatbot software developed with the use of the open-source python platform Rasa (https://rasa.com/). In the case of speech input, it is first converted to text using speech-to-text application programming interface (API) from Microsoft. The entered questions are converted via natural language processing (NLP) into a SPARQL query on the PDOK SPARQL endpoint. Synonyms are used to convert the layman language into specific terminology used at Kadaster. The result of the query is translated into an answer and (possibly) shown on the map. All of the conversations with Loki and the actions that Loki takes based on them are recorded and stored. These data are used to improve the recognition of the questions asked to Loki and to make an inventory of questions that users are interested in.

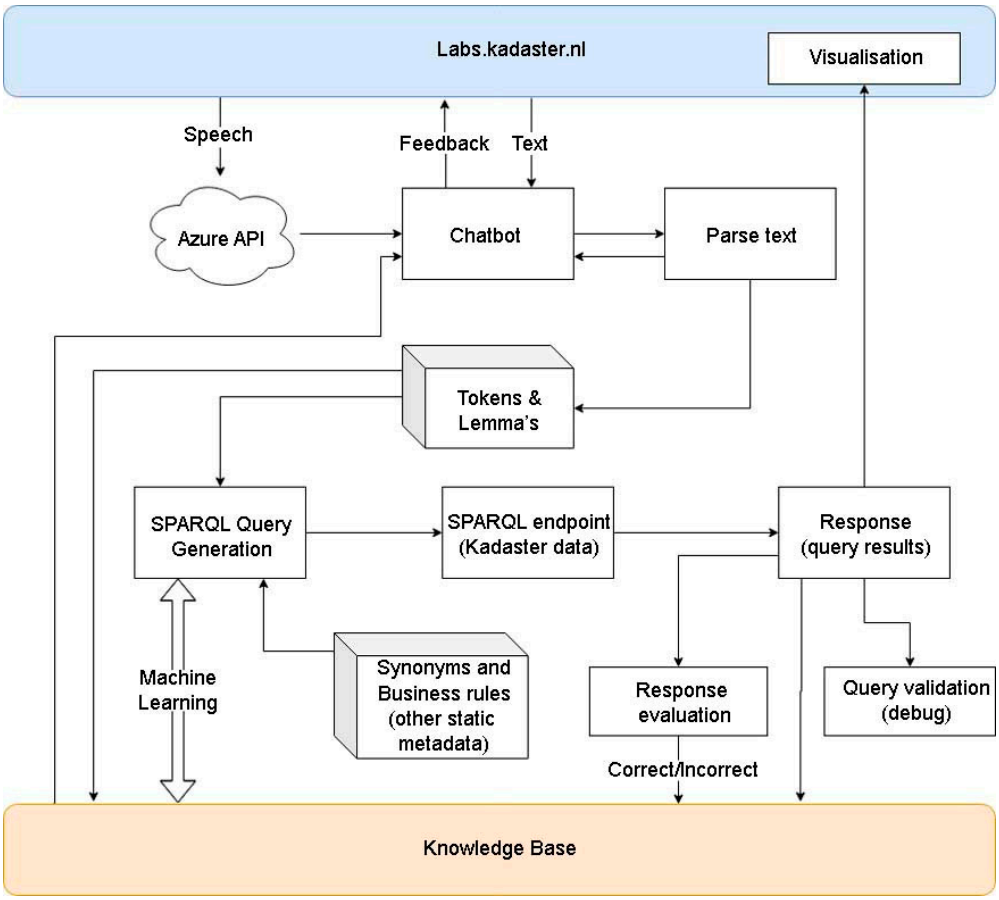


Figure 9. Architecture of Loki (source: authors).

7. Discussion

Many open datasets have been published by governments; for example, there are 11,788 datasets on data.overheid.nl (October 2019) and 939,430 on Europeandataportal.eu (October 2019). However, most of it is three-star data, and being of such quantity, these data have little value for users, as they get lost in endless data catalogs. Therefore, there is a need for KGs or, in general, connected data on a business level within governments.

The KKG project was launched to demonstrate the potential capabilities of a KG built on top of the key registers of Kadaster and using selected sources from seven other governmental organisations.

7.1. Construction and Maintanace: Automation

The construction of the graph started with linking the existing LD sets published by Kadaster. Even though the LD versions of the key registers receive updates (e.g., BAG is updated in real-time),

the constructed links will not be updated. Therefore, the resulting KKG is a static picture and is limited in answering questions about the mutations of data objects.

We experienced that building a KG is a labour-intensive task, even though we launched the project with several large datasets available as linkable data (four stars). Therefore, the automation of the construction and maintenance of the KKG is a key requirement before it can be launched as a product.

The chat bot Loki presented in Section 6.3 demonstrates how the KKG, as a semantic model, can be used in machine learning to provide better answers to human questions. However, machine learning algorithms can be employed for automating KG construction, building user recommended systems, as shown in the literature [9,42,43], and improving the quality of linking.

7.2. From UML to OWL

Creating four-star open data from existing three-star data requires designing an ontology to describe the data first. This task asks for an understanding of the knowledge domain and its intended use, which is often very generic. Therefore, beginners can be paralysed by the number of choices they need to make before they can see the data.

Several works [38,51] promote the approach where target ontologies are generated from existing UML models. One can argue that ÙML modelling is different from OWL; UML is object-oriented and application-oriented, whereby OWL was designed with an open world assumption in mind. However, generating target ontologies from existing UML models might be seen as an actionable recipe that helps to cope with the difficulties related to the need of creating an initial ontology. Once the LD is generated, the underlying ontology can evolve and be adopted to a use case.

7.3. Sources of Links: Space and Case

Finding semantic relations between data items, followed by the materialisation of links, took the major part of the time spent on the construction of the KKG. Legislation is a good source for ontological links, as governmental data collection is based on legislation. However, creating instance-level relations is more difficult; therefore, the spatial component, if present, helps to interlink datasets that have very little in common. In general, inferring topological relations from the geospatial information is a way to work around the ambiguities of the entities in KGs.

Linking with external datasets should be approached, depending on the use case. This will help in defining the scope of linking and will allow for setting up quality requirements for links. In practice, reducing the errors in linking is costly, and without a use case, the required quality level is unknown.

7.4. Access to Governmental Object-Based Intelligence

The notion of object-based intelligence (OBI) is used in the domain of defence and security [52]. It refers to the practice of collecting intelligence information about physical or intangible objects, such as persons, things, events, or places. In the OBI approach, conceptual objects are created to store all of the information and intelligence produced about those people, places, and things. The object becomes a single point of convergence for all of the information related to the subject of interest.

We employed the concept of OBI to emphasise a paradigm shift in the understanding of the governmental data ecosystems that occurred in the wake of the KG construction. KG brings information dispersed over public registers but, in fact, describing the same real-life objects into one view, where they can be seen together as a whole. As a result, we do not need to think about our information needs in terms of datasets and registers anymore. Instead, we can treat information in a more human-like manner, referring to things holistically, regardless of the limitations of the scope of a register. In this context, it is important to ensure outgoing linking with external resources in LOD and provide means for community (e.g., create dedicated properties in Wikidata) to create incoming links from major LD resources. This will improve the discoverability of the KG resources and will provide context for other data.

8. Conclusions and Future Research

Although KGs have existed for a long time, there seems to be a new momentum, spotted by Gartner, adding KGs to the hype cycle in 2018. However, the timing is appropriate, as many governments have published large silos of three-star open data, and there is the technology and expertise to build KGs. The focus moved from publishing data to making data discoverable and useful for answering questions.

One of the main lessons learned from this case study was that building KGs involves a lot of work. One of the main reasons for this is the lack of knowledge about the possible relations. This knowledge was scattered across teams and departments of data owners. Better documentation of data collection and data semantics by data owners can help in overcoming this difficulty.

The generation of initial ontologies from existing UML models can help with faster prototyping and with the testing of the datasets. Ontologies can be improved in later stages.

Establishing spatial relations between objects is another fast way to interlink data on an instance level. For linking with datasets that are lacking a spatial component, it is more appropriate to use a use-case driven approach, as it helps in setting up clear requirements for the linking, including quality.

Location-aware chat bot Loki and the graph browser illustrated how novel interfaces built on top of and powered by KGs can help data providers to redefine the patterns of data consumption for the public.

The lack of knowledge on how to manage KGs and the linksets is a limiting factor for five-star open government data. Therefore, bridging this gap is one of the important future research directions. In this context, it would be interesting to know how to share ownership and responsibility for the maintenance of a KG. In relation to a reduction of costs, further research is needed on the use of machine learning for building KGs.

Author Contributions: Writing—original draft, review & editing S.R. and E.F.; visualization S.R.; supervision E.F., W.B. and R.L.; project administration E.F.; Software P.M., W.B. and M.B.; data curation W.B., M.B. and R.v.V.; investigation W.B., R.L. and R.v.V.

Funding: This research received no external funding.

Acknowledgments: The authors express their gratitude towards Kadaster for their support in performing research on knowledge graphs.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Singhal, A. Introducing the Knowledge Graph: Things, Not Strings. Google Blog Post. 2012. Available online: https://www.blog.google/products/search/introducing-knowledge-graph-things-not/ (accessed on 1 August 2019).
2. Gartner. Gartner Identifies Five Emerging Technology Trends That Will Blur the Lines between Human and Machine. Available online: https://www.gartner.com/en/newsroom/press-releases/2018-08-20-gartner-identifies-five-emerging-technology-trends-that-will-blur-the-lines-between-human-and-machine (accessed on 1 August 2019).
3. Hamad, F.; Liu, I.; Zhang, X. Food Discovery with Uber Eats: Building a Query Understanding Engine. Uber Engineering. 2018. Available online: https://eng.uber.com/uber-eats-query-understanding/ (accessed on 1 August 2019).
4. Chang, S. Scaling Knowledge Access and Retrieval at Airbnb. Airbnb Engineering and Data Science. 2018. Available online: https://medium.com/airbnb-engineering/scaling-knowledge-access-and-retrieval-at-airbnb-665b6ba21e95 (accessed on 1 August 2019).
5. Song, D.; Schilder, F.; Hertz, S.; Saltini, G.; Smiley, C.; Nivarthi, P.; Hazai, O.; Landau, D.; Zaharkin, M.; Zielund, T.; et al. Building and querying an enterprise knowledge graph. *IEEE Trans. Serv. Comput.* **2017**, *12*, 356–369. [CrossRef]
6. Kari, K. The Art of Ontology: Introducing Semantic Web Technologies at Zalando. 2018. Available online: https://jobs.zalando.com/tech/blog/semantic-web-technologies/index.html?gh_src=4n3gxh1 (accessed on 1 August 2019).

7. Bloomberg. Bloomberg Launches "Ready-to-Use" Data Website to Help Firms Derive Value and Enterprise-Wide Efficiencies. 2018. Available online: https://www.bloomberg.com/company/announcements/bloomberg-launches-ready-to-use-data-we (accessed on 1 August 2019).
8. Hubauer, T.; Lamparter, S.; Haase, P.; Herzig, D.M. Use Cases of the Industrial Knowledge Graph at Siemens. In Proceedings of the International Semantic Web Conference (P&D/Industry/BlueSky) 2018, Monterey, CA, USA, 8–12 October 2018; Available online: http://iswc2018.semanticweb.org/sessions/use-cases-of-the-industrial-knowledge-graph-at-siemens (accessed on 1 August 2019).
9. Li, H.; Liu, Y.; Mamoulis, N.; Rosenblum, D.S. Translation-based sequential recommendation for complex users on sparse data. *IEEE Trans. Knowl. Data Eng.* **2019**. [CrossRef]
10. Marino, K.; Salakhutdinov, R.; Gupta, A. The more you know: Using knowledge graphs for image classification. *arXiv* **2016**, arXiv:1612.04844.
11. RDF 1.1 Concepts and Abstract Syntax. Available online: https://www.w3.org/TR/rdf11-concepts/ (accessed on 1 August 2019).
12. Beek, W.; Schlobach, S.; van Harmelen, F. A Contextualised Semantics for owl: SameAs. In *International Semantic Web Conference*; Springer: Cham, Germany, 2016; pp. 405–419.
13. Overheid BRT. Basisregistratie Topografie (BRT). Available online: https://brt.basisregistraties.overheid.nl/ (accessed on 1 August 2019).
14. Overheid BAG. Basisregistratie Adressen en Gebouwen (BAG). Available online: https://bag.basisregistraties.overheid.nl/ (accessed on 1 August 2019).
15. Ehrlinger, L.; Wöß, W. Towards a Definition of Knowledge Graphs. In Proceedings of the SEMANTiCS Posters and Demos Track, Leipzig, Germany, 13–14 September 2016; Volume 48.
16. Wilcke, X.; Bloem, P.; de Boer, V. The knowledge graph as the default data model for learning on heterogeneous knowledge. *Data Sci.* **2017**, *1*, 39–57. [CrossRef]
17. Paulheim, H. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semant. Web* **2017**, *8*, 489–508. [CrossRef]
18. Linked Data: Design Issues. Available online: http://www.w3.org/designissues/linkeddata.html (accessed on 1 August 2019).
19. Silberschatz, A.; Korth, H.F.; Sudarshan, S. Data models. *ACM Comput. Surv.* **1996**, *28*, 105–108. [CrossRef]
20. Hypertext Transfer Protocol—HTTP/1.1. Available online: https://tools.ietf.org/html/rfc2616 (accessed on 1 August 2019).
21. Pérez, J.; Arenas, M.; Gutierrez, C. Semantics and Complexity of SPARQL. In *International Semantic Web Conference*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 30–43.
22. Folmer, E.; Beek, W. Kadaster Data Platform—Overview Architecture. In *Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings*; ScholarWorks@UMass: Boston, MA, USA, 2017; Volume 17, Article 23; Available online: http://scholarworks.umass.edu/foss4g/vol17/iss1/23 (accessed on 1 August 2019).
23. Goodwin, J.; Dolbear, C.; Hart, G. Geographical linked data: The administrative geography of Great Britain on the semantic web. *Trans. GIS* **2008**, *12*, 19–30. [CrossRef]
24. Debruyne, C.; Meehan, A.; Clinton, É.; McNerney, L.; Nautiyal, A.; Lavin, P.; O'Sullivan, D. Ireland's Authoritative Geospatial Linked Data. In *International Semantic Web Conference*; Springer: Cham, Germany, 2017; pp. 66–74.
25. de León, A.; Saquicela, V.; Vilches, L.M.; Villazón-Terrazas, B.; Priyatna, F.; Corcho, O. Geographical linked data: A Spanish use case. In Proceedings of the 6th International Conference on Semantic Systems, Graz, Austria, 1–3 September 2010; ACM: New York, NY, USA, 2010; p. 36.
26. 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. *Int. J. Spat. Data Infrastruct. Res.* **2019**, *14*, 84–106.
27. Ballatore, A.; Bertolotto, M.; Wilson, D. A structural-lexical measure of semantic similarity for geo-knowledge graphs. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 471–492. [CrossRef]
28. Fellbaum, C. *WordNet: An Electronic Lexical Database*; MIT Press: Cambridge, MA, USA, 1998.
29. Wikidata. Available online: https://www.wikidata.org/wiki/Wikidata:Main_Page (accessed on 1 August 2019).
30. Wikimedia. Available online: https://www.wikimedia.org/ (accessed on 1 August 2019).
31. Geonames. Available online: http://www.geonames.org/ (accessed on 1 August 2019).

32. Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; Hellmann, S. DBpedia-A crystallization point for the Web of Data. *Web Semant. Sci. Serv. Agents World Wide Web* **2009**, *7*, 154–165. [CrossRef]
33. Suchanek, F.M.; Kasneci, G.; Weikum, G. Yago: A large ontology from wikipedia and wordnet. *Web Semant. Sci. Serv. Agents World Wide Web* **2008**, *6*, 203–217. [CrossRef]
34. The Linked Open Data Cloud. Available online: https://lod-cloud.net/ (accessed on 1 August 2019).
35. Auer, S.; Lehmann, J.; Hellmann, S. Linkedgeodata: Adding a spatial dimension to the web of data. In *International Semantic Web Conference*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 731–746.
36. Data on the Web Best Practices. W3C Recommendation. Retrieved. Available online: https://www.w3.org/TR/dwbp/ (accessed on 1 August 2019).
37. Spatial Data on the Web Best Practices. W3C Working Group Note. Available online: https://www.w3.org/TR/sdw-bp/ (accessed on 1 August 2019).
38. Brattinga, M.; Maria, P. The geospatial knowledge graph: From traditional UML defined datasets to Linked Data. In Proceedings of the Semantics 2019 Conference, Karlsruhe, Germany, 11 October 2019; Available online: https://2019.semantics.cc/geospatial-knowledge-graph-traditional-uml-defined-datasets-linked-data (accessed on 17 September 2019).
39. Knublauch, H.; Kontokostas, D. Shapes Constraint Language (SHACL). 2017. Available online: https://www.w3.org/TR/shacl/ (accessed on 17 September 2019).
40. Black, J. On the Derivation of Value from Geospatial Linked Data. Ph.D. Thesis, Faculty of Physical Sciences and Engineering, University of Southampton, Southampton, UK, 2013. Available online: https://eprints.soton.ac.uk/358899/ (accessed on 1 August 2019).
41. Battle, R.; Kolas, D. Geosparql: Enabling a geospatial semantic web. *Semant. Web J.* **2011**, *3*, 355–370.
42. Dettmers, T.; Minervini, P.; Stenetorp, P.; Riedel, S. Convolutional 2d knowledge graph embeddings. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
43. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2013; pp. 2787–2795.
44. Liu, Y.; Li, H.; Garcia-Duran, A.; Niepert, M.; Onoro-Rubio, D.; Rosenblum, D.S. MMKG: Multi-modal Knowledge Graphs. In *European Semantic Web Conference*; Springer: Cham, Germany, 2019; pp. 459–474.
45. Marchionini, G. Exploratory search: From finding to understanding. *Commun. ACM* **2006**, *49*, 41–46. [CrossRef]
46. Janowicz, K.; van Harmelen, F.; Hendler, J.A.; Hitzler, P. Why the data train needs semantic rails. *AI Mag.* **2014**, *36*. [CrossRef]
47. Brunetti, J.M.; Auer, S.; García, R.; Klímek, J.; Nečaský, M. Formal linked data visualization model. In Proceedings of the International Conference on Information Integration and Web-Based Applications & Services, Vienna, Austria, 2–4 December 2013; ACM: New York, NY, USA, 2013; p. 309.
48. PDOK Knowledge Graph Browser. Available online: http://linkeddata.ordina.nl/pdkg/resource?subject (accessed on 17 September 2019).
49. Use Case: PDOK Knowledge Graph. Available online: https://labs.kadaster.nl/cases/pdok-knowledge-graph (accessed on 17 September 2019).
50. Linked Data Theatre. Available online: https://github.com/architolk/Linked-Data-Theatre (accessed on 17 September 2019).
51. van den Brink, L.; Janssen, P.; Quak, W.; Stoter, J. Linking spatial data: Semi-automated conversion of geo-information models and GML data to RDF. *Int. J. Spat. Data Infrastruct. Res.* **2014**, *9*, 59–85.
52. Johnston, C.; Wright, E.C., Jr.; Bice, J.; Almendarez, J.; Creekmore, L. Transforming defense analysis. *JFQ Jt. Force Q.* **2015**, *79*, 12–18.