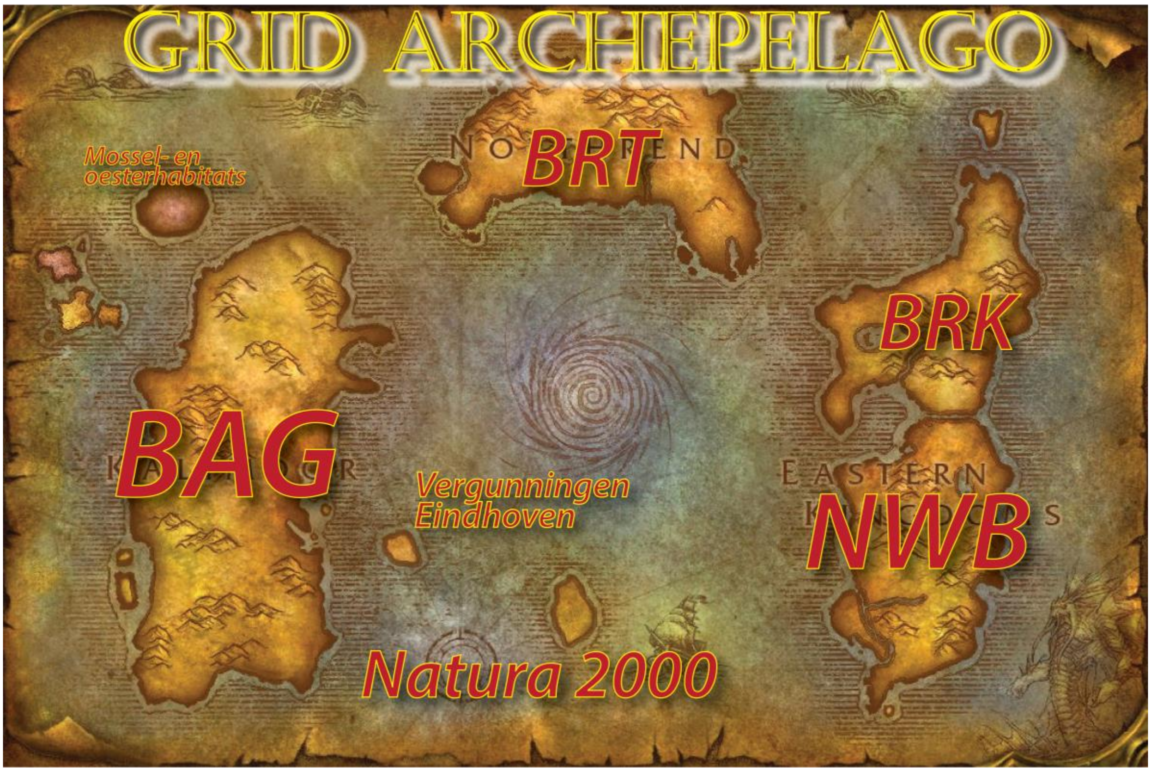
Linking Kadaster's Linked Data

Internship Report



Name: Sam Ubels

Registration number: 3401650

Period of Internship: 2016-10-03 - 2017-03-31

Date final report: 2017-04-06

Erwin Folmer	*Rob Lemmens*
Kadaster	*ITC*
Laan van Westenenk 701, Apeldoorn	

Management summary

This internship report, provides an overview of the research questions, the approach to answering these, the resulst of the research and the conclusion that were drawn. The research set out to answer the question: ‘To what extent does the linking of the datasets hosted by the Kadaster as Linked Data provide added value in the context of the omgevingswet’. The omgevingswet is trying to turn multiple laws into one, which will require the combination of a lot of data, preferably in one platform. The Kadaster is trying to provide this platform, using Linked Data as a method to combine this data. A strategy to link these datasets however has not yet been established. This research has attempted to provide this strategy. Through two iterations of the linking process different approaches were tested and during meetings with the appropriate team members these approaches were discussed and altered. This resulted in suggesting SILK as the appropriate tools to use for linking the data, an approach were the link predicates that are used are, wherever possible, taken from existing ontologies and the links are stored in separate graphs following the standards discussed in the VoID vocabulary. The conclusion that can be drawn is that it appears that there is a lot of added value to be had from interlinking in the context of the omgevingswet.

List of abbreviations

BAG – Basisregistratie Adressen en Gebouwen

BRT – Basisregistratie Topografie

RDF – Resource Description Framework

SKOS – Simple Knowledge Organization System

SPARQL – Simple Protocol and RDF Query Language

OWL – Web Ontology Language

VE – Vergunningen Eindhoven

Table of contents

1 Introduction

1.1 The kadaster

The Kadaster is a Dutch governmental agency that is historically responsible by law for the registration of real-estate ownership and the plots on which this real-estate is build. It registers the owners and the rights that come with this ownership. Furthermore, they are statutorily responsible for the registration of topography, ships and aircraft. On top of these tasks it has become one of the main providers of geo-information in the Netherlands, through the PDOK platform in which they are a partner, and of which they are the administrator. Not only do they provide datasets of which they are the owner, they also serve as a platform for the registrations of other parties. They administer so called 'Landelijke voorzieningen' in which they collect data from, for example, municipalities and publish this as national datasets. On top of providing access to these datasets or excerpts from it the Kadaster also provides custom made datasets, mostly for other governmental agencies through their GIS department. Finally, the Kadaster is also involved in the international activities, advising and sharing knowledge on keeping registries and dealing with geo-information. Currently, they are in the process of creating a new data platform which should be able to deal with changing demands and future responsibilities. On this new data platform, the change is made from serving all data through geo-standards to more general web standards in the form of linked data. The internship will be done within the team responsible for building this new data platform.

1.2 Omgevingswet

With a new law concerning public planning, the 'omgevingswet' which will be valid from 2018 the new data platform may need to provide certain possibilities which are not possible in the current infrastructure. Efficient access to data from multiple (public) organisations will be very useful, and may even be a requirement. As the omgevingswet aims to simplify planning law, questions like 'can I live here' or 'can I build a house here' should become easier to answer. However, to come to an accurate answer to these questions data will be needed from more than one public agency. The current laws and regulations surrounding planning legislation consist of dozens of laws and instruments which in the omgevingswet should be brought back to just one law. This should for better alignment of plans concerning spatial planning, environment and nature and provides more space to municipalities, provinces and waterboards to adjust legislation per their specific needs and context (van Rooijen, 2016). The result of this is that one agency or system should be able to access all data concerned with planning decisions efficiently for this law to work. Linked Data is proposed as one of the techniques that should allow for this to be possible.

In recent years Linked Data has emerged as a new and efficient way to store and combine data in a way which allows it to be both machine and human readable, allowing humans and computers to work in cooperation. In the Netherlands, the

Kadaster is one of the pioneering organisations that is experimenting with this new technique. Through their involvement in the PDOK (Public Services on the Map), they have access to a lot of public data and the move towards using linked data is an effort to serve this data to the public in an easier and more efficient manner. Furthermore, as the name suggests Linked Data allows for the linking of data from multiple sources potentially providing the ideal basis for the combining of data needed for the omgevingswet.

1.2.1 Significance of this topic

Where currently data is often stored and queried per specific dataset and is often stored in varying data formats, Linked Data should bring an end to this, or at least, take away a part of the current limitations. Linked Data standards allow the linking of data between multiple datasets and consequently allows the querying of all datasets from one point with a single query. This is done by using a single format (RDF) to store all data, adding semantics (well-defined meaning) and linking datasets based on overlapping classes or attributes.
Currently, the Linked Data efforts at the Kadaster have reached the point where multiple datasets have been converted to RDF and can be queried from the same endpoint. However, the actual linking of the datasets has not yet been achieved. This internship focussed on taking the first steps in this linking process. As there is a significant number of potential datasets that can be relevant in use cases connected to the omgevingswet it is to be expected that the process of linking datasets is not something that is just a one-time project, but rather a process that might have to be repeated over and over when cases arise that require additional new datasets. Therefore, investing time in researching the right approach to linking Linked Data is an important topic for the Kadaster. Even though there is no prior research on linking key governmental registries, as other cadastral agencies have also just recently started exploring the added value of Linked Data, there is prior research in other fields which might be relevant as a starting point for this internship.

1.2.2 Related work

A review of literature on linking Linked Data shows that even though Linked Data is being increasingly used in many different domains and the amount of linked data that is available is rising the linking of these datasets remains a difficult task. This is partly because of the lack of incentives and partly because of a lack of user friendly tools. Wölger et al. (2011) did a survey on interlinking methods and tools, this can be taken as a starting point for determining the right approach and selecting some tools to test. Some tools that do seem promising to use and look like they are still being supported are Silk and Limes. The first step in determining the linking approach according to the first review of the literature is finding 'join-points'. Join-points are a point of reference shared by two or more datasets which allows them to be linked on these attributes (Omitola et al., 2010). Join points might be based on background knowledge, or domain specific knowledge, which is available inside a certain organisation but are not clear to non-expert users. Other types of joint-points are matching patterns, for example matching strings (city names) or matching identifiers such as ISBN numbers.

Finally, there is potential overlap based on patterns such as geographical overlap, so matching on the geometries corresponding to the entities. In order for people browsing the data to understand where data is linked it is recommended to establish these links not only on the entity level but also on a conceptual level, by aligning the schema's, which is known as ontology matching (Parundekar, Knoblock, & Ambite, 2010).

1.2.3 Added value of this research

Even though there is certainly some literature on related work, this is still relatively scarce. Especially in the domain of key registries and public organisations. Even though there is some literature on exploring linking methods, as mentioned in the previous section, it is still useful to explore possible tools and try different methods during this internship. The idea of first exploring join points as a first step towards finding semantic overlap is something that has been taken from literature and partly inspires the internship goals mentioned in the following section. This internship aims at providing new insight into applying linking strategies on large datasets, finding the most suitable way in which these linksets can be stored within the infrastructure of a large data platform and further testing of tools where necessary.

1.3 Aim of the internship

The research question for this internship is the following:
'To what extent does the linking of the datasets hosted by the Kadaster as Linked Data provide added value in the context of the omgevingswet'

To answer this question several internship goals have been defined:

- *Providing a clear overview of the semantic overlap that exists between the Linked Datasets*

The different datasets have a lot of concepts that have some form of semantic overlap or 'join-points'. The first goal is to find the classes that have this and then determine what the exact relationship is between these classes.

- *Find a way to query the entities with semantic overlap from the database and establishing links between them*

The second goal is finding the most efficient way to query the entities that overlap from the database. The way to do this is by using SPARQL, the query language of Linked Data. This will require some trial and error, and testing, to find the most efficient query. Furthermore, the SPARQL query language allows for a 'construct' query, which allows to not only find these entities but immediately create a link between them. However, literature review has shown that there are also linking tools that can be used. So, testing is needed to find the most efficient way to create the links.

- *Determine the most useful way to store the generated links*

Whether a construct query will be used or a linking tool the links then have to be stored somewhere on the platform. These links can be stored within the datasets they are concerned with, or somewhere else on the platform. The goal here is to find the best way, keeping in mind that the registries that are maintained by the Kadaster need to remain as uncluttered as possible.

- *Determine an update strategy for the links*

When the links have been stored on the platform, they need to be maintained. This means that links between two entities should still be valid if one of the datasets is updated. This means that some strategy needs to be worked out about when and in what order links should be updated and what should trigger these updates.

- *Determine the added value of the created links*

To determine if the links have added value several possible use cases should be selected and queries should be provided by the end of the project to demonstrate this added value.

1.4 Reading guide

This report consists of five chapters. The current one, explaining the context and the aim of the research. The second chapter will provide insight into the methods used during this research, the approach that was adopted and the outcomes that are expected. Chapter three presents the results that were obtained, leaving the analysis of the results to chapter four. Both chapters also shed some light on the limitations and relevant side notes. Chapter four further provides recommendations for further research into linking strategies. The final chapter provides a conclusion and a review of the research question.

1.5 Linked Data and SPARQL

To understand the significance of this internship project and the approach taken the reader should have a basic understanding of a couple of concepts related to Linked Data. This section provides a short technical background.

1.5.1 Linked Data

Linked Data refers to the recommended best practices for exposing, sharing and connecting RDF data via dereferenceable URIs. RDF is the Resource Description Framework, a way to describe information or data. The way data is described in RDF is through object – predicate – subject triples. The subject is the piece of data that is being described, the predicate describes a relationship, and the subject is another piece of data. So, if a book is described on the Semantic Web it might look like '*book – hasTitle – carrie*', '*book – hasAuthor – Stephen_King*'. As mentioned before RDF describes data via dereferenceable URIs. These are Uniform Resource Identifiers, and the fact that they are dereferenceable means they can be looked up online providing more information on this specific entity. So, if someone would look up the URI for '*book*' the triples described above would show, providing more information on '*book*'. All parts of the triple are then again dereferenceable allowing the user to look up who Stephen King is and what the predicate hasAuthor means. Other people online can link pieces of data to these same URIs creating a large network of data.

1.5.2 Ontologies

The way structure is added to the semantic web is by using ontologies to model the data. Ontologies are collections of classes and properties defining structure to data in a specific domain. For example, an ontology about books might describe a class such as 'Book' which can have properties such as hasAuthor, hasGenre, hasTitle. The hasGenre property might then have a pre-defined selection of possible values (adventure, thriller, etc.), all identified by a specific URI to which people can refer. If all organisations that produce data on books adhere to the URIs and structure defined in this ontology it is possible to take data from different data providers and use it all in the same application as it is already in the same format and structure.

1.5.3 (Geo)SPARQL

Simple Protocol and RDF Query Languague (SPARQL), was first standardised in 2008, and its most recent version SPARQL 1.1 was standardised in 2013. SPARQL allows the querying of RDF knowledge graphs that are exposed through a SPARQL endpoint. SPARQL makes use of triple patterns like those described in RDF documents however it allows for the replacement of some of the resources by variables. It allows for type of queries (select, ask, describe & construct) of which select and ask are relevant to this research. The select query allows the user to specify of which variables used in the query the bound value should be returned. This data is returned in the form of a raw table. The construct query allows for the specification of a specific triple pattern consisting of one or more variables, and constructing this new triple patterns based on the values bound to the variables in the query. This is the most basic way of linking two datasets.

To be able to extend queries on Linked Data with the possibility to ask questions about topological relations the GeoSPARQL extension was developed by the Open Geospatial Consortium (OGC). It's an attempt to link the Geo oriented standards of the OGC to the Semantic Web. It provides a vocabulary that allows to describe geographic data and furthermore provides different sets of topological query functions which can be used in queries that request this geographic data. The sets of topological functions are the Simple Features, Egenhofer and RCC8 spatial relations (OGC, 2012). Finally, it also provides a set of basic geographical operators such as buffer, intersection and union. This functionality combined with a construct query can be used to create links based on geographical patters rather than overlapping semantics.

1.5.4 Prefixes

In many of the serialisations of RDF as well as in many SPARQL queries prefixes are used to make it more readable. Prefixes are simple shorthand for describing RDF resources that have the same URI pattern. For example, all classes in the ontology for the Kadaster dataset Basisregistratie Adressen en Gebouwen (BAG) start with http://bag.basisregistraties.overheid.nl/def/bag#. By defining at the start of a document that the prefix bag represents said URI pattern the rest of the classes can simply be described as bag:Pand or bag:Wooplaats, instead of typing the entire URI for every resource.

2 Methodology

This section will provide some insight into the methodology applied to obtaining the desired results. It will first provide an overview of the steps taken, followed by a discussion of the tools. Then a discussion of the methods and procedures followed by an overview of the preparations on the side of the intern before the research could start. Finally, the limitations and subsequent adaptions will be discussed.

[How was the formulated hypotheses tested, justification of the methodological approach, who were involved, what software was used, how were the research / activities carried out, how were data collected etc.].
The following headings may provide a guide

2.1 Method used

In order to achieve the internship goals an approach was set up that consisted of an iterative process. As there is a lacking foundation in literature to compare the results of linking and the efficiency of the linking process to, the method that is used was to do two iterations where the first could be used as a baseline to compare the results of the second attempt to.

The first iteration consisted of attempting to finding a solution to the first three research goals. Finding the semantic overlap between the Kadaster datasets, testing the use of SPARQL construct queries as a linking mechanism and finding the most suitable way to store the results on the platform. Below these steps, how to assess their success and the baseline this provides are discussed:

Goal 1: finding the semantic overlap between Kadaster datasets.

The first goal, finding the semantic overlap between the Kadaster datasets provides an answer to the question whether there is enough semantic overlap to find added value in linking between these datasets.

Output: A model can be created providing an overview of between what classes semantic overlap exists. These join-points are a starting point to for the linking process, where for every created link the questions should be asked does this add any value in the context of the ‘omgevingswet’.

Assessment: Between the different types of join-points domain specific knowledge, similar properties or topological relations, the question can be asked are these types of links useful to create and to what extent do they add value. This can be assessed by the intern himself.

Goal 2: finding a way to query to overlapping entities from the database and establishing links

This goal is focussed on the finding the right approach to establish the links. As the literature review already showed there are two approaches to creating the links. The first being the SPARQL Construct queries and the second being linking tools. During the first iteration the SPARQL Construct queries will be used to create the links in order to provide a base measurement to compare to the results of the linking tools in the second iteration.

Output: The output of this step will be the links as they were defined in the output of the first goal.

Assessment: The process of creating the links can be assessed in multiple ways. First, there is the efficiency of the query writing. This can be used to be compared to the time it takes to set up the linking task in the tools. Secondly, there is execution time, so, how long does it take for the query to run and the links to be created. Finally, there is the completeness. So how many entities are linked and are the links created correct.

Goal 3: determining the most useful way to store the links
This goal is concerned with finding a way to store the links in a fashion that allows for easy access and good findability, while making clear that it concerns links and does not belong to the original key registry.

Output: a plan that describes how the links are stored and how they can be found and identified as links.

Assessment: Discussion session with team members, with expertise in the field of Linked Data as well as insight into the Kadaster strategy concerning key registries. Here the proposed solution can be criticised and possibly be changed or dismissed.

After completing this step and assessing the results a second iteration will follow. The section below describes the goals of the second iteration.

Goal 1: potentially redefining the semantic overlap and add more datasets
The first goal of this iteration depends on the assessment of the first goal of the previous iteration. In the case that there is too little added value to obtain from linking the Kadaster datasets a look will be taken at non-Kadaster datasets that might add value.

Output: A new model describing all datasets and the semantic overlap between them. This again serves as input for the following steps and depicts the links that will be created.

Assessment: Again, this assessment can be done by the intern himself. It will be the result of researching potential different datasets with semantic overlap. If this is the case, this step will be successful.

Goal 2: testing linking tools to see if they provide easier querying of the dataset and creation of links

Output: Overview of the potential linking tools and their suitability. Furthermore, a new set of links for the overlap defined in the output of goal 1.

Assessment: The tools can be assessed on the same criteria as the SPARQL construct queries, and additionally between them usability is a factor.

Goal 3: Determining the most useful way to store the links
As a result of the assessment meeting concerning the previous approach to storing the links a new approach may have to be set up or the previous one may have to be refined.

Output: a plan that describes how the links are stored and how they can be found and identified as links.

Assessment: Another meeting will be planned, again discussing the approach with the appropriate experts from the team. Here some refinement can take place, after which the links will be stored according to the plan

Goal 4: Determining an update strategy
Based on an exploration of how the updates of the datasets are done, an approach can be designed which allows for the links to remain valid in the case of changes to the entities after an update.

Output: An update strategy, also describing suitable tools involved in the updates.

Assessment: The approach can be tested as soon as an update occurs, by checking the links on consistency after the data has been updated.

Goal 5: determining the added value of the constructed links and the process of linking itself

Output: Presentation and report on the results of the research providing a clear conclusion on whether or not linking Linked Data provides added value for the Kadaster in the context of the omgevingswet.

Assessment: Assessment will be done partially by the Kadaster supervisor Erwin Folmer, and partially by the GIMA supervisor Rob Lemmens.

2.2 Context

This section will provide a sketch of the context of this research. First it will provide a short overview of how the supervision during the internship was organised. Then an overview of the used tools will follow and finally a description of the Kadaster Datasets that have been used will follow.

2.2.1 Supervision and support

The supervision from GIMA has been done by Rob Lemmens, who is familiar with Linked Data and therefore suited to provide supervision on both content as process of the internship. From the Kadaster the official supervisor is Erwin Folmer, whose duties on both the IT and the business side of the organisation do not allow him to be present every day. Therefore the day to day supervision was provided by Stanislav Ronzhin, a PHD student who does part of his research at the Kadaster.

2.2.2 Tools

This section describes the tools that have been used during the internship. It describes an RDF editor, which is a text editor with extended capabilities to handle RDF documents. A triple store, which is basically the database of Linked Data. Finally, it includes a linking tool, as described before, a tool that assists in the linking of Linked Data sets.

RDF Editor
The RDF editor that was used for the editing of RDF files is RDF Editor. It can be used to quickly transform the RDF between different serialisations and it offers syntactical support for all RDF formats.

GraphDB
GraphDB is a triple store designed for medium data volumes, up to 100 million triples. It includes an implementation of OpenRefine which allows the user to refine RDF datasets before adding them to the store, but also turning tabular data into RDF. It doesn't naturally allow GeoSparql queries, but a plugin that comes with the download of GraphDB does create this possibility. It also has a highly developed GUI that makes it very easy in its use and is, next to a free version, also available on a commercial license, providing around the clock support. This makes it one of the more commercially developed triple stores, however, little research is available on the performance of GraphDB compared to other triple stores.

SILK
Silk is the linking tool that has been used during the second phase of this internship. It has been developed by the university of Mannheim and they describe it as 'an open source framework for integrating heterogeneous data sources'. It provides an easy to use GUI, with drag and drop functionality to dynamically create the preferred transformations and needed SPARQL queries. Next to linking it also useable for data transformations and in addition to linking between Linked Data sources it supports a range of other formats.

2.2.3 Datasets

In the linking process, several datasets have been used. This section will discuss what each dataset has to offer.

BAG

The Basisregisratie Adressen en Gebouwen (BAG), contains information on all buildings and their addresses in the Netherlands. Its structure and contents are largely defined by law. The Linked Data version created by the Kadaster only became available half way through the internship, however before this a beta version was available, and as the data model is already defined, the differences between the two datasets were limited.

BRT

The Basisregistratie Topografie (BRT) is the digital topographic dataset of the Netherlands. It is offered on different levels of detail, however the Linked Data version only contains the Top10NL, which is the most detailed version, suitable for applications working with a scale between 1:5000 and 1:25000. It contains topographic elements such as roads, buildings, and the administrative borders of municipalities and provinces.

Vergunningen Eindhoven

The Vergunnignen Eindhoven datasets is an experimental dataset containing information on all permits that have been requested in the municipality of Eindhoven. It provides properties such as case numbers, status of the request, issue date, address and a description of the type of permit and what project is planned.

Geonames

Geonames is an open source dataset that contains over 10 million geographical names and has over 9 million unique features spanning 2.8 million populated places and 5.5 million alternate names for places. This is one of the go to datasets for accurate information on geographical locations as it serves over 150 million request per day on their web services. Most of their information is served through their API, however they have part of their dataset available as RDF.

Wikidata

Wikidata is a ‘free and open knowledge base’. It is structured in such a way that it is both human and machine editable. It is the central storage for structured information taken from all the different wiki projects. All data is served as RDF and they provide a SPARQL endpoint from where this can be queried.

DBpedia

DBpedia describes itself as a ‘crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the web’. It is set up to be able to make use of the all information on Wikipedia and be able to ask sophisticated questions which is not possible through the regular Wikipedia. The English DBpedia version alone describes 4,58 million things, of which 4.22 million are classified in their ontology, spanning classes such as people, places, movies, companies, diseases and species.

2.3 Preparation

In order to prepare for the research task following the approach described in section 2.1, some initial preparations where needed. Next to trying to understand the organisation and getting to know the people, there was a need to increase my technical knowledge to be able to perform this internship. These preparations can be categorised in three categories. SPARQL/Linked Data knowledge, understanding the Kadaster data models and finding the right use case.
The initial SPARQL knowledge was quite limited, so the first weeks where taken to properly study the SPARQL standard, focussing on the construct queries and special functions. Of course, this process continued throughout the internship period, however some initial work was needed in order to be able to perform the first iteration of the research approach.

Secondly, in order to be able to properly study the semantic overlap between the different datasets, first an effort was needed to explore all datasets. For every dataset, the data model was studied to gain an understanding of the classes and their relationships. This allowed for a better assessment of what links where logical and would add added value. This also provided insight into possible links that could be created inside datasets, as not all logical relations within datasets where materialised with triples in the RDF representations of the Kadaster data.

Finally, some meetings were planned with people from GIS maatwerk, which is the division within the Kadaster that is responsible for the tailor-made datasets requested and calculations requested by other organisations. They could provide their typical use cases, which could provide a basis for testing the added value of the links and seeing if these use cases could be solved using the new data platform. However, as it turned out, the use cases provided by GIS maatwerk, where all in some way related to the datasets for internal use which have not yet been converted to RDF. Meaning a valid use case could not be found. This means that the added value provided should be assessed in a different manner.

3 Results

This chapter will provide an overview of the obtained results. It consists of three sections. First, there is a description of the applied methodology in practice and in what way it diverted from the planned approach. Secondly, the results will be presented. Finally, an overview will be provided of the limitations of the results, answering questions like, where are the results lacking? What could be done better?

3.1 Employed methodology

The approach for this internship consisted of two iterations of linking the Kadaster data. Partly to compare the results of two linking approaches, partly to have a midterm assessment of some of the strategies that were devised concerning the storing of the data, the use of linking predicates and the added value created. During the first iteration SPARQL construct queries were used to create the links. The links that were created were only between different Kadaster datasets. A model was created that provided an overview of the assumed semantic overlap and based on that model SPARQL queries were written to create links to materialise this semantic overlap. During this linking process, several design choices had to be made. First, the two datasets must be linked using a predicate. This predicate can be taken from an existing ontology, of if none are available, it can be created for this specific purpose. Furthermore, a decision should be made on where the links can be stored, again different logical possibilities arise. The links can be stored in the same graph as one of the existing datasets or they can be stored in a separate location. For both these important choices a strategy has been devised, which will be presented in the next chapter. After this first iteration, the results have been assessed in a meeting with different team members, discussing the suitability of the approach and discussing in what way it could be improved. This served as input for the second iteration.

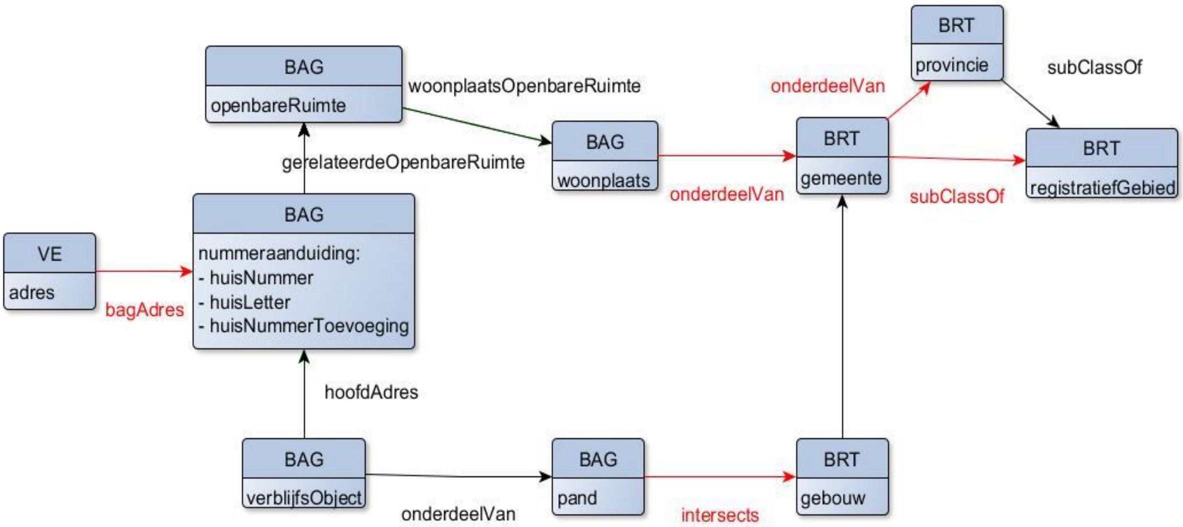
The second iteration started with input for devising a different strategy for the linking predicates as well as storing the data. Furthermore, the results of the first iteration together with the start of a new project period which saw an additional requirement to link to external data source, the second iteration also included a look at prominent external datasets. The linking the second iteration was done using linking tools instead of SPARQL construct queries to see if these tools are more efficient or easier to use than SPARQL. This time the results of have been discussed in an in-depth technical meeting with the relevant team members and during a more general presentation to the entire team.

3.2 Findings / Results

The results of the two iterations will be provided in this section. For both a model showing the links and linking predicates will be provided first, followed by an explanation of the used approach for the linking predicates as well as the way to store the created links. For a more detailed analysis of the results, see the following chapter.

First iteration

Figure 3.1 shows the discovered Semantic Overlap between the classes in the internal datasets and the links created. The arrows show the direction of the links with a label providing the used link predicate. The black arrows represent the links that are already present within the datasets, the red arrows show the links that have been created. All links have been created using SPARQL construct queries. Several different types of links have been created, using different strategies.



The approach that has been selected for the linking predicates is based on the idea that the links would be stored as a new property of the existing entities in the different classes. By storing the links here this would allow them to serve as if they were part of the datasets to begin with, providing extra references to outside sources. Because the links would be stored this way, the linking predicates that were used where also created within the namespace of the source dataset. So in the case of the links between Vergunningen Eindhoven and BAG, the linking predicate would be VE:bagAdres.

The links

Vergunningen Eindhoven – BAG

The first link that was created was between Vergunningen Eindhoven (VE) and BAG. The VE dataset has a class called *Vergunningen Eindhoven* and entities in this class have a property address. The key registry on addresses, the BAG, also contains addresses, however, these are not as straightforward as they are in VE. The BAG class that contains addresses is *Nummeraanduiding*, this class contains information on the addresses in the form of 4 properties; *postcode, huisnummer, huisletter, huisnummertoevoeging*. These together provide a full address; however, the name of the street is not in there but rather in a related class *OpenbareRuimte*. The streetname is however required to create the links, as it can only be done by string comparison. To create the links it was necessary to recreate the BAG addresses in the same format as they are presented in VE. To achieve this the SPARQL concat function has been used, to collect the separate address elements from the two classes and concatenate them into an address in the same format. Using a filter function the addresses have been compared and the matching addresses where linked using the VE:bagAdres predicate.

BAG – BRT

Between BAG and BRT two sets of links have been created. Both are based on topological relations between the two entities. First, there is the link between the classes *Pand* and *Gebouw*. These two classes both describe structures in the Netherlands, however the way these structures are represented differs, as do the properties to describe them. An example of this is that where a BAG Pand would represent a house in a row of houses as a separate Pand, the BRT Gebouw class would represent the entire row of houses as one Gebouw. However, where the BAG structures are described based on their function (living, office, education), the BRT provides more specific classifications (Hospital, Prison, Lighthouse etc.). So, there is some semantic overlap between the entities as they both describe buildings, but also clear differences. However, because of the additional information they provide about the buildings, there is still added value in creating the links. To do this, the geometries can be compared using GeoSPARQL. The intersects function provides all Pand and Gebouw entities that have intersecting geometries. As a linking predicate BAG:intersect has been used.

The second area of semantic overlap between the BAG and BRT datasets is between the classes *Woonplaats (Place)* and *Gemeente* (*Municipality*). Every gemeente is made up of one or more woonplaatsen. The relationship between the classes is obvious and the added value of linking the two classes is that all BAG data can then be queried on a municipal level (e.g. the total number of buildings with a living function within a certain municipality). The process of linking has first been attempted using GeoSPARQL, the first results quickly showed that the geometries of BRT and BAG did not line up, meaning the links created were incorrect. However, in the download of the BAG in its original format a *Woonplaatscodetabel* is provided, specifying the relationship between these two classes by defining the matching codes. These codes are provided in the linked data version of both datasets as a property of these classes. So, the linking process required the conversion of the *Woonplaatscodetabel* to RDF and using this RDF dataset as a temporary extra property for the BAG woonplaats class. This property could then be used to find the correct municipality to link to and the links were saved using the BAG:onderdeelVan property.

BRT – BRT

When looking at the second link created between BAG and BRT, between woonplaats and gemeente, a similar relationship can be found between the classes gemeente and provincie (Province) within BRT. Even though the classes are within the same dataset, creating an additional link within the dataset can also provide added value. The added value is similar to the added value of the woonplaats - gemeente link in that it not allows to query BAG data on a provincial level if the previous links are also created. The same predicate has been used, however now stored within the BRT dataset, so BRT:onderdeelVan. Because the classes are within the same dataset, meaning the geometries matched, and because of the limited amount of entities, GeoSPARQL could be used to establish the links.

Limitations

The selected approach to the linking process showed some limitations. First, there is the use of GeoSPARQL of which the current implementation in GraphDB is not efficient enough to be used in this setting. For example, the link between Pand and Gebouw, which was only created for the municipality of Eindhoven, took 46 hours to execute. This is not fast enough since the approach should also be able to perform when daily updates of the data occur. Furthermore, creating all linking predicates within the namespace of the source dataset proved to be problematic as predicates that are used multiple times are more useful if they can be reused. This means that it is better to define the linking predicates in a neutral ontology, not linking it to a namespace of one of the existing Kadaster datasets.

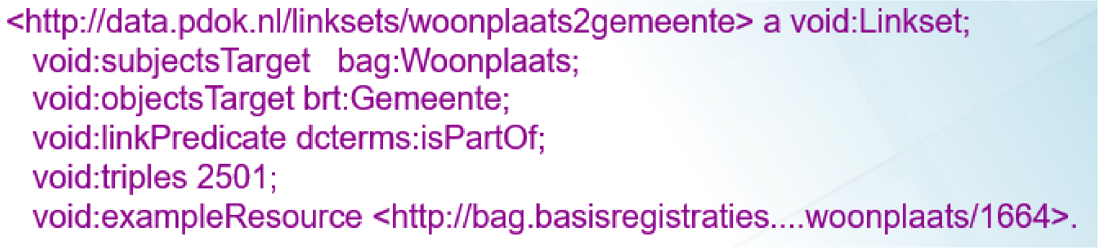
The assessment

During the assessment of the first results, two issues came to light. First, there is the storing of the links as additional properties in the source dataset. Because it concerns key registries which are devised in such a way that they store specific information in the most concrete way possible, based on only storing data that is necessary, adding additional properties is against the core idea of these registries. Especially when the number of links starts growing the selected approach was deemed unsustainable. Also, the approach concerning the linking predicates received some criticism. The ontologies describing the data are based on the original data models, and have been devised in close cooperation with the Kadaster team currently responsible for these datasets. In case of the BAG, the data stored in the dataset is even defined by law. Therefore, it is not desirable to create extra properties within this namespace.

Second iteration

Figure 3.3 shows the semantic overlap and the created links during the second iteration of the linking process. This diagram shows all links that have been created during the linking process. During this iteration, the first step was selecting the linking tool that would be used. In order to determine which tools should be tested an overview of tools was taken from (Siorpaes, Simperl, Thaler, & Hofer, 2011). From these tools, all aimed at different types of linking and varying in the further possibilities, for example data transformation before linking, LIMES and SILK seemed the most promising for the task at hand. However, due to technical problems and issues with the version of Java used for other tools, installing limes turned out to be complicated. Based on further literature research, where limes and silk where tested against each other for different linking tasks, SILK proved to be the more efficient and more promising tool (Rajabi, Sicilia, & Sanchez-Alonso, 2014). Therefore, no further effort was made to find a solution to the technical issues and SILK was used as linking tool for this internship.

Before starting the second iteration of linking, first a new approach has been devised concerning the linking predicates and the way to store the data. Based on the assessment of the previous strategy, the new strategy should aim at storing the links as separate datasets and either reusing existing predicates or creating the new predicates in a new namespace. When links are stored as separate datasets another thing that should be considered is the findability of the datasets. The Vocabulary of Interlinked Datasets (VoID) is the solution for this. It is a vocabulary that allows the description of links stored as separate datasets, so called linksets. In order to describe and find these linksets VoID offers a distinct classname: *Linkset*, and a set of properties allowing the description of the source and target of a linkset, number of triples, example resources and the link predicate. These properties can then be used to find the linksets based on all the specified criteria. Figure 3.2 shows an example VoID description of the linkset between BAG woonplaats and BRT gemeente. The subjectTarget predicate refers to the source class and the objectsTarget predicate refers to the target class.



```
<http://data.pdok.nl/linksets/woonplaats2gemeente> a void:Linkset;
   void:subjectsTarget   bag:Woonplaats;
   void:objectsTarget brt:Gemeente;
   void:linkPredicate dcterms:isPartOf;
   void:triples 2501;
   void:exampleResource <http://bag.basisregistraties....woonplaats/1664>.
```

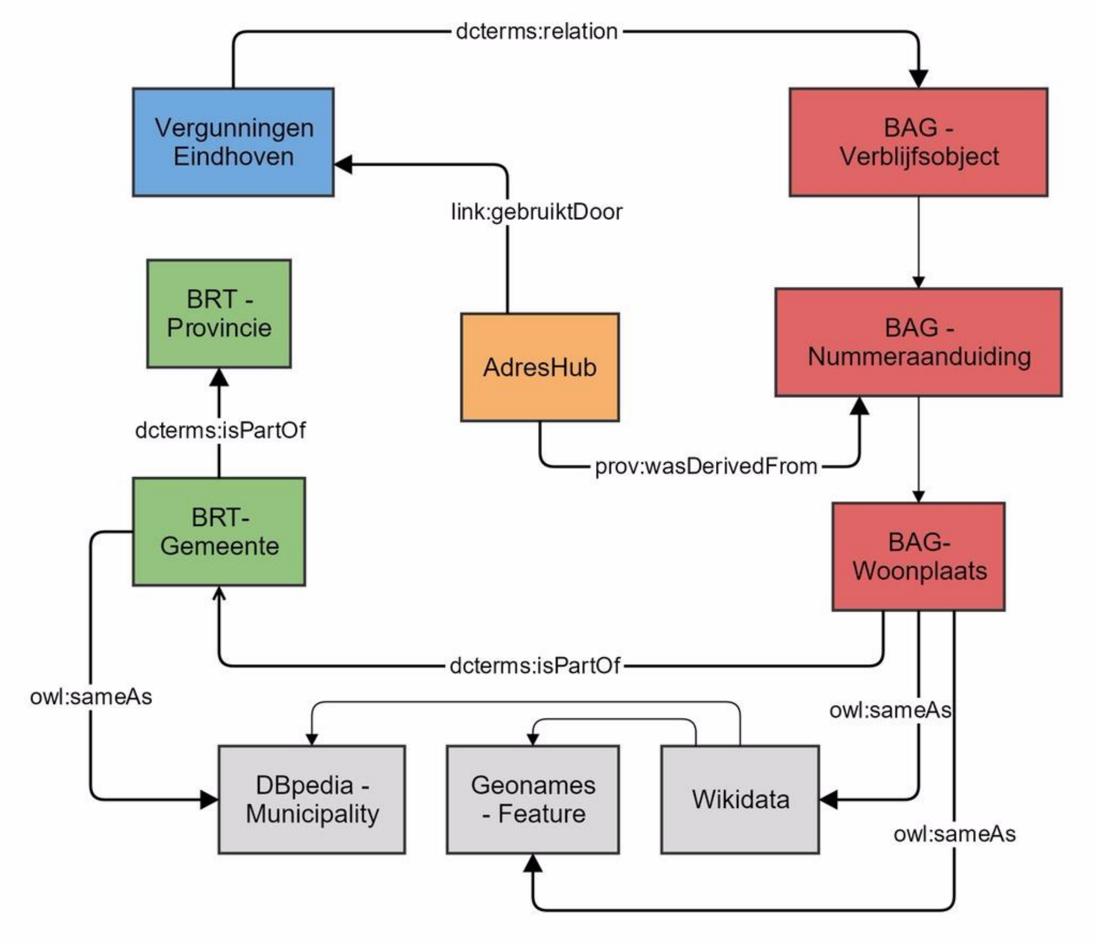
Further research into possible existing ontologies that could be used to provide the predicates that describe the relationship between the linked entities showed that there are many possibilities in ontologies such as DCterms, SKOS, and OWL ontologies. Because of this useable range of possible predicates and because devising an own ontology and linking predicates comes with the responsibility for the Kadaster to maintain this ontology for the second iteration the approach that was selected for the predicate selection is making use of existing ontologies where possible. The selected predicates will be discussed in the following section.

The links

During the second iteration of the linking process the SILK interlinking framework was used to create the linksets. Silk is an easy to use tool, that allows you to define a dataset to be used in the linking by specifying the SPARQL endpoint it should access and the class name. From here it pulls in all the properties and lists them in the drag and drop linking interface. In this interface, you simply drag in the properties needed and connect them using an arrow to either one of the transformation tools or one of the comparison tools. The comparison tools take two inputs of either one of the properties or the output of one of the transformation tools. Below the linking interface the link predicate can be specified that should be used in the comparison tool finds a positive match. Instead of generating a SPARQL construct query which does all the transformations the linking tool sends a regular select query to the endpoint and then does the transformation and comparison locally using its own algorithms.

Vergunningen Eindhoven – BAG

The same link as in the initial iteration has been recreated with one exception. The link this time does not link to the class *Nummeraanduiding* but to the class *Verblijfsobject*. This has been done because the nummeraanduiding describes the address, but is not the object that the permit is issued for, the permit is issued for a specific home. The address is merely a related class describing this home. Therefore, semantically the relationship between vergunningen and verblijsobject is more logical. Silk has been used to create these links, using the levensthein distance comparator the addresses on the permit and the addresses related to the verblijfsobject could be compared. As a linking predicate dcterms: relation has been used. This predicate is described in the dcterms ontology as: 'a related resource' which is less specific than the previously used predicate, but solves the issues surrounding the previous predicate strategy.



<u>BAG – BRT</u>

During this iteration, the link between BAG woonplaats and BRT Gemeente has been recreated. The advantage of using SILK for this is that it is possible to add RDF dumps as a dataset in the linking process. This means that it is no longer necessary to add the *woonplaatscodetabel* used for the matching of the operations as a property of the woonplaats class, but can just be used on its own. However, conversion to RDF is still needed to use the table. To compare the codes and create the links again the Levenshtein distance comparator has been used. As a linking predicate dcterms:isPartOf has been used. This predicate describes 'a related resource in which the described resource is physically or logically included'. This is a bit broader than the previous predicate, but is semantically similar, and the Semantic Web promotes the reuse of ontologies, which makes it a better solution.

<u>BRT – BRT</u>

The link between BRT Gemeente and BRT Provincie could not be recreated using SILK. Even though SILK does have a series of geographic comparators, the tool kept crashing when it tried to import the geometries to be used in the comparison. Therefore the links have been recreated using a different linking predicate, dcterms:isPartOf, however this has been done in the same way as the first linkset was created.

BRT – Dbpedia

During this iteration, on top of the internal links also external links have been created. The first link that has been created is between BRT and DBpedia. The DBpedia dataset provides a lot of information on municipalities that is not present in the BRT dataset (e.g. the mayor, picture of the flag, number of inhabitants). Because of a direct match between the properties describing the name of the municipalities (dbo:naam and brt:naamOfficieel) it was easy to use silk to load both datasets and link them using the levenshtein distance comparator. The link predicate that was used is the owl:sameAs, this predicate description is: 'The property that determines that two given individuals are equal'. Even though there is some discussion surrounding this predicate, whether two entities in different datasets can really be described as the same it is still an often used predicate that can be interpreted as 'every property describing the entity in that dataset is valid for the entity in this dataset'.

BAG – Wikidata

Wikidata, like DBpedia, provides a lot of information that is not available in the Kadaster data (e.g. list of monuments, identifiers for other datasets, elevation above sea level). The link between BAG and wikidata was a rather easy one to establish, because of their list of identifiers in other datasets. One of these datasets is the BAG. This made it very easy to set up silk to use the levenshtein distance comparator to compare the BAG IDs and create a link using the predicate owl:sameAs.

BAG – Geonames

The link between the BAG and Geonames was a free one, in that wikidata already provides a link to Geonames. So through here the Geonames URIs could be retrieved and the link could be established using owl:sameAs. As Geonames is the largest RDF collection of placenames and alternative names it is a very useful dataset to link to.

AdresHub

Figure 3.3 showed, among the already discussed datasets, the AdresHub as one of the datasets. To improve the findability of the linksets, an attempt has been made to create a system which, based on reusable notions could provide easier access to and understanding of the Kadaster data and what could be found here. The idea behind this is to take reusable notions such as Adres or Gebouw, which occur in multiple datasets and create an additional dataset which provides URIs for all entities within this dataset. The example here is using the notion of Address, which is easier to understand than BAG:nummeraanduiding. Using a SPARQL construct query all the relevant properties from BAG were extracted and used to form one property providing a full address. These were then linked to the original nummeraanduiding using the predicate prov:wasDerivedFrom which is described as: '*A derivation is a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.* '. Then links were created to other datasets which use addresses, such as VE. This time the link predicate has been created within the *link* namespace as this approach already defined new classes and properties to create the dataset, it would no longer be an issue to define own link predicates. The predicate used was: link:gebruiktDoor, specifying that a dataset uses this address. If this is used to link to multiple datasets then one single points of access would provide all information the Kadaster has about a specific address. However, as too little time was available during the internship period to further explore this idea, it is worth documenting for possible further research.

3.3 Limitations of the results

The results of the internship are close to the expected results; however, this does not mean that there are no limitations. The most important limitation of these results is that due to the inability to find a proper use case in the context of the omgevingswet, it was not possible to see if the linksets provided a better approach to solving issues faced by the Kadaster than the current way of working. Furthermore, the linkset that has been created that links to Geonames is relevant and at the same time has limited usability, this is because Geonames does not have a SPARQL endpoint, therefore finding the right Geonames URI in the linkset, does not allow the user to find the related information in the Geonames database, however this URI can be used on the Geonames API to find additional information, but this requires an extra effort. Finally, the AdresHub proved to be harder to accomplish than expected. Initially it was intended as a LinkHub that provided multiple reusable notions, however finding the semantic overlap between for example Pand and Gebouw and finding one artificial entity that would cover both and allow the linking to additional datasets proved very time consuming. Therefore, the idea was beyond the scope of this internship.

4 Discussion of results

The previous chapter presented the results as they are, with some reference and analysis of individual results, however, this chapter will provide a more general assessment of the results, related to the overall goal of the internship. It will first summarise the results presented in the previous chapter. Then an analysis will follow of the overall result. Section 4.3 will assess the significance of the results in a broader perspective. The next section reiterates the research goals and provides an assessment of whether the goals have been reached. The final section discusses what is next in finding the best way to link Linked Data for the Kadaster.

4.1 Summary of the results

This internship has increased the number of linksets at the Kadaster from zero to seven. It has tested two approaches to storing the links and finding the right link predicate to describe the relationship between two entities, and has finally come up with a solution that seems like a sustainable approach for both elements. The first iteration attempted to store the links and create the predicates within the namespace of the original dataset, however, discussion with the team provided too much valid criticism on this approach to continue this way. Therefore, a new strategy has been devised using the Vocabulary of Interlinked Datasets to store the linksets in a separate graph, but still making them findable. As the first iteration proved that linking between only Kadaster datasets using the currently available datasets was of limited value, during the second iteration a look was taken at linking to external datasets, and useful links were created to WikiData and DBpedia. Furthermore, during the second iteration the SILK interlinking framework was successfully tested and proved in most cases to be more efficient than using SPARQL construct queries for the linking process.

4.2 Analysis of the results

An assessment of the overall results leads to two important conclusions. The first is that linking does provide added value for the Kadaster, and the second being that using linking tools provides a significant advantage during this process. This section will go into these two results and provide a further explanation.
The actual result of the internship are of course the linksets that have been created. However, more important is the process that was gone through during the creation of the linksets. An important remark here is that the approach to storing the data and the predicate selection are more relevant results for the Kadaster than the created linksets. This is because with the selection of the right tools and a linking strategy in place, the creating of new linksets should be easier, this is therefore more relevant than the linksets themselves.

When it comes to looking at the added value of creating links the initial plan was to compare the efficiency of specific queries to the way answers to these questions are found in the current setup. However, as a relevant use case was not found, added value must be assessed in a different way. As already explained in the previous chapter presenting the results, added value is there in many cases when it comes to the type of questions you can ask. The combining of two types of building representations to add more specific types to query on, or the aggregation of BAG statistics on a municipal level. In many cases linking also allows you to materialise tacit knowledge, where a regular user, not experienced with the Kadaster datasets would look at a class like Nummeraanduiding it would not be clear what this entails. However, by linking there from the adres in the permit class it materialises the knowledge present in the Kadaster for everyone to see. All these are exapmles of semantic enrichment of the data. Another aspect that has highlighted the added value of linking the data is quality control. During the linking process there is an automated check on completeness in and consistency between datasets. When the links between the BRT and DBpedia were established initially only 379 of the 390 municipalities were matched to each other based on the names. It turned out that DBpedia used the Dutch names or all Frisian municipality as official name, where the BRT used the Frisian names. It also turned out that two municipalities in DBpedia were not provided with the dbo:naam predicate, an where therefore not found. This in one linking task provided insight into inconsistency and incompleteness in the DBpedi datasets. This demonstrates that even without the ability to compare efficiency in solving issues using linked Linked Data, there is added value in generating linksets.

The second conclusion that can be drawn is that the use of linking tools is of added value in the linking process. First, the locally performed comparison algorithms are far more efficient than the filter function provided by SPARQL. In the creation of the addresses from BAG and comparing them to the addresses in the VE dataset, the initial SPARQL query took about 11 hours, where SILK performed this task in 11 minutes. There is a side note however that in cases where small sets of data have to be compared using identical identifiers (e.g. Wikidata – BAG) then a SPARQL query can save time as the time to set up the linking tasks in SILK is longer than writing a SPARQL query, when proficient in both. Next to the efficiency benefits, there are more reasons to use SILK over the initial approach. The most important one being easy reuse of the linking tasks. Once set up the tasks can be edited easily and rerun to tweak the results. Furthermore, after a linking task it provides a result overview with a measure of similarity for matched entities, showing the values for the matched entities. Then it allows the user to easily add or remove matches or results based on expert knowledge. Furthermore, the number of comparison operators is far larger than in the SPARQL standard. Finally, SILK provides an API which allows the reuse of the set up linking tools, the insertion of data and the extraction of the results, all from distance. This makes it very easy to insert the linking of datasets in the dataset creation process with just a few lines of code.

4.3 Significance of results

Where the previous sections discussed what should be taken away from the results as they were presented in the previous chapter, this section looks at the significance of the results. Are they to be taken as absolute facts or are there some side notes that need to be considered, and what limitations are there. A first important aspect in this case is the fact that the quality assessment of the resulting linksets is hard to do. As explained in the previous section, in some cases the faults are obvious (379 links for 390 municipalities) however, in a lot of cases this is less clear. When matching millions of buildings in two different datasets, it is impossible to just from the numbers and with a quick look through the data assess to what extent the linking task was successful. A method needs to be devised to be able to provide some measure of correctness. Therefore, the linksets generated during this internship can serve as an example of the way linking works, and the way in which it adds value, however the current strategy is not ready for production level linksets.

A second aspects is that the it has proven very hard to solve the issue of linking based on topological relations. Both GeoSPARQL which performs very slowly and the geo comparators in SILK do not provide the right solution, however the testing of SILK on this aspect has been limited, it might be a memory issues, which can easily be solved, or the issue might lie in the way the geometries are served by the Kadaster endpoint. Where the previous chapter stated that the geo comparators were not working properly, crashing the software, this needs to be taken as a very limited test of these capabilities.

4.4 Discussion of research goals

Now that the results have been discussed, a review of the research goals is needed to assess the success of the internship. The goals are provided below, and are discussed per goal.

Providing a clear overview of the semantic overlap that exists between the Linked Datasets

The semantic overlap that has been discovered during this research has been presented in the two models in chapter 2. The goal has been reached during the first iteration, where the second iteration also provides an overview of semantic overlap with outside data sources. Even though a clear overview of the semantic overlap between datasets has been provided this does not mean that there is no further overlap to be found between the datasets.

Find a way to query the entities with semantic overlap from the database and establishing links between them

Two methods to approach this goal have been tested, the SPARQL construct query and the use of the SILK interlinking framework. It can be concluded that SILK is the more efficient and easy to use solution. However, it takes a bigger initial effort to use than SPARQL does. In exchange for this effort a larger selection of comparison tools is provided and a far more efficient matching process. The goal was finding a method, and two were found, meaning that the goal was reached.

Determine the most useful way to store the generated links

The storing of the links, including in this case the strategy concerning the link predicates used has been developed over the whole internship period. An initial plan was provided and discussed with the relevant team members, after which a switch was made to a new approach. The results being that the links should be stored as separate linksets according to the standard set in the VoID vocabulary. This strategy seems to be a sustainable solution, reusing existing link predicates where possible.

Determine an update strategy for the links
The determining of an update strategy is the only goal that has not been reached. During the internship period it turned out that in the current stage of the development of the platform the datasets that were linked do not yet receive updates. Therefore, the development of a strategy to deal with these updates did not have any significance as it could not be tested. However, the API that is provided in SILK, triggering linking tasks from a script based on the receiving of an update should be possible. This is the recommended approach as soon as testing becomes a real possibility. However, as this is rather a suggestion than a thought-out strategy the goal has not been reached.

Determine the added value of the created links
The determination of added value of the linksets, could be determined in the way that was initially planned due to the lacking of a proper use case. However, the empirical process provided a lot of insight into what other ways the linking process adds value for an organisation such as the Kadaster. Among other things it provides a wider range of properties that entities can be selected on, and it provides an automated way of quality assessment of the datasets that are being linked. Therefore, it can be concluded that at least part of the added value has been determined.

4.5 Remaining gaps / needs for further research

As discussed in the introduction, before the start of this internship the linking datasets effort at the Kadaster had not yet started. This internship research provided an first exploration of the possibilities but in a lot of directions more research is needed in order to devise a fully successful and efficient linking strategy.

The first and maybe most important area that needs research is a way to assess the quality of the produced linksets. Even though silk provides a nice overview to evaluate the resulting links, this is still a manual task, which cannot be done in the case of large datasets being linked. Some form of automated quality checks have to be devised. Furthermore, as discussed in the previous section, the current possibilities to link based on topological relations is subpar. This while the data the Kadaster serves is largely geo-data meaning that a lot of the relationship between entities in datasets are topological in nature. SILK might still prove to be the solution for this problem, however, other approaches should be tested as well.

Another aspect that needs further research or at least further testing is working with large datasets. Most of the current linksets are based on a small number of entities. The largest being the links between BAG and VE, but even these are only the permits for Eindhoven, with around 18,000 entities. Especially when updates start occurring daily the efficiency of the linking strategy will be an important aspect. On top of efficiency testing there is a need for research into the possibilities of updating the linksets in accordance with the updates of the regular datasets. Here again SILK might be able to play a role through the use of the API in combination with the reusable linking tasks.

5 Conclusion

This chapter will provide an answer to the research question stated in the first chapter: ‘To what extent does the linking of the datasets hosted by the Kadaster as Linked Data provide added value in the context of the omgevingswet’. The answer to this is clear, there is a lot of added value to be had. The idea behind the omgevingswet includes the combining of data from a lot of different datasets. These datasets will have a lot of semantic overlap as they describe related issues. To turn this into an understandable and clear collection of data, linking them using linksets shows some clear advantages. First, the discovered semantic overlap, knowledge that might be based or experience or specific domain knowledge can be materialised so that it can be used by anyone working with the data. Furthermore, the linking provides a far more efficient way of querying the data, and allows collection of data from one dataset based on properties defined in another. Also, if the issues with querying on geographic relationships remains in issue, linksets can serve as storage of these relationships so that the query only has to be performed on updates, rather than for every query that is run. Finally, the added value shows in the fact that the linking process provides an automated form of data quality checks. When a lot of dataset from different providers will have to be combined in order to come to an answer to omgevingswet questions then an initial check of the consistency and completeness of all the data that occurs during the linking process provides a lot of added value.

Bibliography

Parundekar, R., Knoblock, C. A., & Ambite, L. (2010). Linking and Building Ontologies of Linked Data | ISWC 2010, 598–614. Retrieved from http://iswc2010.semanticweb.org/accepted-papers/334

Rajabi, E., Sicilia, M.-A., & Sanchez-Alonso, S. (2014). An empirical study on the evaluation of interlinking tools on the Web of Data. *Journal of Information Science*, *40*(5), 637–648. http://doi.org/10.1177/0165551514538151

Siorpaes, K., Simperl, E., Thaler, S., & Hofer, C. (2011). A survey on Data Interlinking Methods.