# DATA STORYTELLING: VISUALISING LINKED OPEN DATA OF THE DUTCH KADASTER

Author: B. E. Guliker

Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
Bachelor thesis for Creative Technology

*Supervised by:*
dr. ir. M. van Keulen, Faculty EEMCS
dr. ir. E.J.A Folmer, Faculty BMS, Kadaster

05 July 2019

kadaster

UNIVERSITY OF TWENTE.

# Abstract

The World Wide Web has made it easier than ever to share knowledge with others. Web pages are connected through hyperlinks and together they form a giant linked collection of documents. Guided by the vision of the Semantic Web, linked open data (LOD) connects data sets through URIs which link together and form a giant linked collection of data. Public bodies such as governments and research initiatives already offer many different data sets as linked open data. The Netherlands' Cadastre, Land Registry and Mapping agency - in short Kadaster - has been sharing knowledge on land administration and geospatial information with other countries for decades. These data sets are published on their platform PDOK (i.e. 'Public services on the map'). As part of PDOK, the Kadaster shows the value of their data sets through data stories inside the PDOK Labs environment. This thesis explores and creates a new data story using the linked open data sets of the Kadaster. The process is guided by a literature review on the creation of data stories and the Creative Technology Design Approach. The end result is "The CBS & Kadaster Data Dashboard", a web-based dashboard which allows the user to gain insight into many measures (income, energy usage, demographics etc.) about the municipalities and neighbourhoods in the Netherlands. These measures can be used to gain insight into many societal issues. The report ends with an ethical discussion on data storytelling. The final dashboard will be published on PDOK Labs.

# Contents

# List of Figures

# 1 Introduction

## 1.1 An introduction to linked open data

In today's age a vast majority of our information about the world is digitised as data on the World Wide Web. Government agencies around the world publish data on a wide variety of topics. The true value of data lies in its ability to give new insights. These new insights can be used among others to support policy making and public administration [1]. Data sets are often compared with other data sets to look for relationships or combined to give even more information on a particular subject. For example, when looking at a city there are many different pieces of data, or statistics, available ranging from information about the population, the history of the city or the soil compositions in neighbourhoods. Even though nowadays there is an almost endless supply of data on a broad variety of topics, it is often difficult to combine data from different sources into a single application that retrieves the information straight from the source and is always up-to-date. Linked data aims to solve this problem through so called semantic queries.

## 1.2 The problem

The Netherlands' Cadastre, Land Registry and Mapping agency - in short Kadaster - has been sharing knowledge on land administration and geospatial information with other countries for decades. The Kadaster publishes large data sets, including key registers of the Dutch Government such as the full topography of the Netherlands. Their public data sets are published in the PDOK data catalogue and accessible via an API or as linked data. PDOK stands for 'Publieke Dienstverlening Op de Kaart', i.e., public service on the map. The PDOK platform provides high quality, reliable and most importantly up-to-date spatial data which are used by many businesses and organisations in the Netherlands [2].

However, simply publishing the data does not provide new insight in the data set. To solve this problem, the Kadaster has created the PDOK Labs environment with the intent to show the value of the data. Inside PDOK Labs there are so called data stories, each of which explore some of the data sets the Kadaster has to offer. The data stories offer data visualisations accompanied by descriptive text to highlight interesting insights which can be drawn from a particular data set, thus showing the relevance and value of the data set. Additionally, the underlying SPARQL queries can be viewed. These queries are directly responsible for getting the data from the source, so when something changes in the data set, the visualisation automatically updates according to the latest information. The PDOK Labs environment is constantly in development, there are only data stories for a small subset of all available data sets. The goal of this thesis is to explore and develop a new data story which can be used to show the relevance and value of the linked open data sets of the Kadaster.

## 1.3 Research question

The main research question this thesis aims to answer is: "How to implement a data story which shows the value of the linked open data sets of the Kadaster?"

As a secondary goal, this thesis aims to contribute to the knowledge field of linked data by providing an ethical review on data storytelling and some guidelines which are important for the creation of linked open data stories.

The main research question will be guided by the following three sub-questions which will be further explored in a literature review:

1. Sub Q1: What are existing guidelines/important factors when designing data visualisations?
2. Sub Q2: How can data visualisations be used to tell a narrative?
3. Sub Q3: How can the effectiveness of data visualisations be evaluated?

## 1.4 Outline of this report

The next chapter of this thesis dives further into the technologies behind linked data and aims to answer the three sub-questions by means of a literature review. It concludes with a state-of-the-art research on currently existing tools and visualisations made using open linked data. Chapter three outlines the Creative Technology Design Approach and other methods used during the development of the data story. Chapter four explores the possible data stories that can be made using the data sets of the Kadaster and gives an overview of the structure of the data set. Based on the initial design, chapter five specifies the requirements for the final that is realised in chapter six. Chapter six outlines how the technologies are applied as well as what design choices are made to resolve some of the challenges and problems encountered during development. Finally, chapter seven gives an evaluation of the created to see if it matches the requirements. It also provides an ethical review on the created data story and data storytelling in general. The final chapter contains a conclusion on the end results of this thesis and recommends areas for further research.

# 2 State of the Art on linked data & data visualisation

## 2.1 Introduction to SPARQL & linked data

The World Wide Web has made it easier than ever to share knowledge with others. Web pages are connected through hyperlinks, together they form giant linked collection of documents. Consequently, there is now an abundance of data freely available to us. Public bodies such as governments and research initiatives offer many different data sets on a variety of topics [3]. Despite the abundance of data, interoperability is lacking. It is still a difficult task to combine data from many different sources. Most data bases require distinct ways to access the data and have their data structured according to different standards [4]. The implicit relationship between two data sets cannot be interpreted by machines. By applying the same principles the Web uses to link documents, the concept of linked (open) data aims to solve the problem of separated data and define explicit relationships to make the data.

The concept of linked data was first introduced in 2006 by Tim Berners-Lee [3]. Linked data is part of the Semantic Web, an extension of the World Wide Web through standards defined by the World Wide Web Consortium (W3C). The vision of the Semantic Web has been interpreted in many different ways. According to Berners-Lee, "The first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a Semantic Web - a web of data that can be processed directly or indirectly by machines." [5]. Marshall & Shipman describe three other perspectives related to the Semantic Web which they found shared across literature [6]. Overall, the common goal of the Semantic Web is to create a machine-readable Web and linked data is a means to attain that goal.

### 2.1.1 Linked data - URIs & triples

The philosophy behind linked data is using the technologies behind the Web to link data sources[7]. Separate data sets often describe different properties of the same object. By referring to these objects via a Uniform Resource Identifier (URI), other data sets can connect to the data by referencing these URIs. The importance of the data link ensures an explicit relation between both elements that is clearly defined according to a common standard (RDF, see section 2.1.2). Another advantage of the URI is also that the consumer can plug the URI in their browser to view its references to other URIs.

The next step is to use these URIs to link the data together. Take the city Amsterdam as an example. Amsterdam has a population of 820.000. Linked data is stored in so called triples. A triple consists of three parts: a subject, a predicate and an object. So <Amsterdam (subject)> <has a population of (predicate)> <820.000 (object)> is an example of a triple. The subject "Amsterdam" and the predicate "has a population of can be identified via a URI, the object is in this case a literal (of type integer). The object can also be a different URI, for example: "Amsterdam lies within the province North

Holland". Now the object is North Holland, which has its own URI that other data sets can link to. A linked data set consists of many of these tipples, and are stored in a database called a triplestore. If data sets reference the URIs of other data sets, they become linked together. All in all, Berners-Lee summarises the four principles of linked data, which give a set of best practices for connecting and publishing linked data, as follows:

1. Use uniform resource identifiers (URIs) as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things (often done by the owl:sameAs relationship).

### 2.1.2 Linked data - RDF

Triples are part of the Resource Description Framework (RDF) standard. RDF is made to describe resources, also called a meta data model. Many of the common relationships between objects and subjects are defined in the RDF standard. However, RDF alone is often not sufficient to suit all data structures. There is still the need to define custom data structures and add not yet existing relationships. As such, RDF is commonly extended with other ontology languages, such as the Web Ontology Language (OWL). OWL provides a very helpful relationship "owl:SameAs" which is used to relate two different URIs of different data sets and say they are the same.

There are also different serialisations available for the triples in RDF. The most common is Turtle syntax, however RDF/XML syntax for example is an XML-based syntax which was the first format for serialising. The Turtle syntax is similar to that of the SPARQL Protocol and RDF Query Language, or SPARQL for short. SPARQL will be used later on in this thesis to query the necessary data from a triplestore. All in all, the main purpose of RDF is to provide a structured framework for describing information.

A real life example of two linked data sets are the Kadaster's 'Basisregistratie Adressen en Gebouwen (BAG)', i.e., Key register Addresses and Buildings, and DBpedia. The BAG has among others information on the borders of municipalities, while DBpedia contains all information from Wikipedia and thus has text on the history of a municipality. Now either the BAG could refer to the URI for Rotterdam of DBpedia, or DBpedia could refer to the URI for Rotterdam of the BAG. Both should use the "owl:SameAs" relationship and store it in a triple. This would permanently connect the two data sets making it akin to one big data set.

### 2.1.3 Linked data - 5 star data

In addition to linked data, Berners-Lee defined the quality of data published according to a 5-star scale[5]:

* Make your data available on the Web (whatever format) under an open license

** Make it available as structured data (e.g. Excel instead of an image scan)

*** Make it available in a non-proprietary open format (e.g., CSV instead of Excel)

**** Use URIs to denote things, so that people can point at your stuff

***** Link your data to other data to provide context5

The more stars, the more advantages the data has. Linked open data is seen as the best sort of data and has five stars. The step between every extra star gives numerous new benefits for both the consumers and publishers of the data. The five-star model also helps explain the difference between a traditional RESTful API and linked open data. APIs often do not use URIs to refer to things. This makes it difficult to reference the data in other applications. Furthermore, they lack the clear explicit relationships defined via RDF. Often APIs return a JSON object with for example property called "value". Without context, it is unclear what this "value" means, whereas the relationships in linked data have URIs of their own. All in all, APIs fail to reach the last two points and are thus at most three-star data.

Despite the advantages of linked data, there are also higher costs to publishing better data (higher star rating). It costs more resources to build and maintain a data server than to just upload an image. If the data uses URIs then these URIs also need to be checked for broken or incorrect links that might no longer work when the data changes. In return, other data publishers can link to your data making it easier to discover. The benefits of linked data do outweigh the costs of the initial investment in the long term as five-star data will allow consumers to more easily discover the data and when more data is linked together it strengthens the overall collection of data.

## 2.2 Literature review

Now despite all the data that is being published, the data sets themselves do not necessarily lead to new insights. Long lists of numerical data on their own are not easy to interpret for humans. When public data sets do not have accompanying visualisations to give an impression of what the data is about, they are less likely to be used by others [8]. By adding data visualisations, data sets are easier to explore and analyse.

Data visualisations are important powerful tools to convey a story in a short amount of time. In storytelling, the expression "Show, don't tell" is often used to express that information can be transmitted quicker via visualisations than via verbal communication. The Kadaster themselves has created PDOK Labs where they publish so called data stories. They use data stories to show what insights can be gained from some of their data sets, however the amount of data stories the Kadaster has to over is limited. Even though data stories have proven to be beneficial, many governmental agencies publish their data without visualisations of the data [9]. There exists a lack of insight in many data sets and as such there is need for data stories as a way to show the value of the data sets.

Therefore, the aim of this literature review is to provide an overview on the aspects of creating a good data story through the effective use of data visualisations. Additionally,

this literature review discusses the possible evaluation methods to see if the created data story effectively serves the purpose it was intended for. The main research question is thus as follows: "How to create and evaluate an insightful data story?" To answer this question, three elements of the data story are discussed. First, an overview of the important factors of designing an insightful data story are given. Second, a discussion on how different storytelling methods can be applied on data storytelling. Third, several methods of evaluating data stories are given. In the conclusion, the total outline on how an insightful data story can be created by incorporating the three elements discussed before. Finally, the review finishes with a discussion on the quality of the used literature review and proposes areas for further research on the ethical risks in data storytelling such as framing.

### 2.2.1 Design of insightful data visualisations

There are two aspects to consider when designing a data visualisation, visual encodings and graph types. Firstly, there are multiple types of visual encodings to represent data which each have different strong and weak points. The effectiveness of a data visualisation relies on human cognitive recognition and their ability to convert these visual encoding into information.

Clevelland and McGills [10] rank ten different types of encodings based on accuracy. The ten encodings sorted by accuracy are: 1) position along common scale, 2) position not aligned to scale, 3) length, 4) direction, 5) angle, 6) area, 7) volume, 8) curvature, 9) shading, 10) colour (see also figure 2.1). The most accurate encoding is position along a common scale, while colour is the least accurate. Erik and Ragan [11] analysed the same encodings, and add that despite colour being an inaccurate encoding, it is one of the fastest encodings people notice.

However, Iliinsky warns that the redundant encodings often make a visualisation less comprehensible as they can overload the reader with information [12]. An example is having three lines with different colours which are also marked differently (dashed, dotted, etc.). The visual encodings of Clevelland and McGill are fundamental and often used in research to compare visualisations. Discussion of every encoding is out of the scope of this literature review, instead the focus lies on one of the most discussed visual encodings, colouring.

One of the ways how colouring affects a data visualisation is how it evokes different emotions, moods and enhance memorability. The field of colour psychology looks into how specific colours influences human behaviour. Red is often seen as an proactive, passionate colour, whereas pink is a more feminine colour which shows care [13]. Engelhardt affirms these findings and highlights that certain visualisation feel more truthful based on the colours chosen inside, with blue in general being a more trustworthy colour as it is associated with authority [14].

In spite of colour being an important visual encoding in any visualisation, it is also the most commonly misused encoding. Research has shown that most visualisations do not take colour blindness into account A common colour scale to indicate positive and negative relations are green and red. However, 8% of all males of Northern-European decent is affected by red-green colour blindness They would not see this scale as

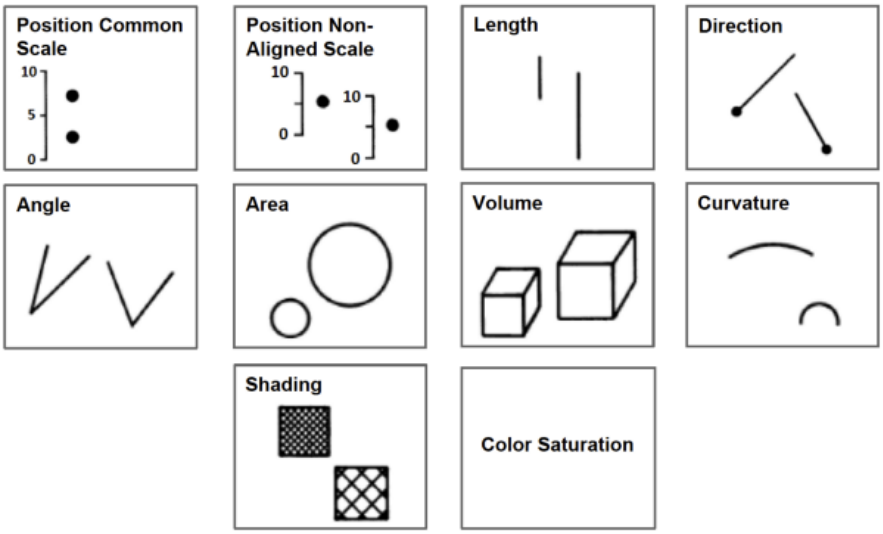



Figure 2.1: The ten elementary encodings by McGill [10]

green-red but instead as yellow-blue. The link between green being a positive colour is then lost and instead feels like an arbitrary choice to them, distracting them from the main message of the visualisation. The second most common type of colour blindness is followed by blue-yellow colour blindness. This causes problems with temperature scales which are often red to blue colour scales. As such these scales are harder to interpret for this type of colour blindness. The use of many different colours in a visualisation has shown to make it less comprehensible. It is recommended to have at most have five colour categories since it is hard for humans to subconsciously remember the meaning of more than five colours. As such it increases the time a user needs to understand a visualisation.

Secondly, another aspect of data visualisation are the graphs which make up the visualisation. Korsa and Moere state the visualisations which are remembered longest are often unique and different than anything the user has ever seen before. visualisations that are unique to the data set are memorable that those that exists only out of common charts such as bar charts and pie charts [15]. Being unique requires the visualisation as a whole to represent the data set.

Despite this, the book by Cleveland and William states that a unique visualisation does impair the time it takes for the reader to understand the visualisation [16]. A unique visualisation should be composed of common charts but integrate them in such a way that they contain familiar elements of existing visualisations. It appears that having unique elements in a visualisation can help people to remember the message of the visualisation, but it does have drawbacks of increased complexity making it harder for users to understand.

For spatial data, the choropleth is a commonly used visualisation which has both advantages and disadvantages compared to other types of visualisations. It maps a certain quantitative scale, such as population density, to a specified colour scale on a map. According to Cockcroft the main advantage of using choropleth is that it is

easily understood since it is a popular visualisation method. [17] They do give a false impression of change around the borders of the defined areas. In figure 2.2 the five main types of spatial visualisations have been shown. Above the figure are the main types of questions each visualisation can answer. Besides the choropleth answering how much something is, the other four types answer the questions, where is it, when did it happen, what is it about, how/why did something happen. The other four graphs are easier to understand, they are less prone to bias, but also less commonly used [18]. In the end, the type of visualisation that is chosen will depend on what question the author wants to answer.


Figure 2.2: Five possible representations of spatial data[1][10]

In conclusion, for a data visualisation to be insightful correct usage of colour is required as it is often misused. The two main pitfalls that have been identified are redundant usage of encodings and not taking into account colourblindness. Furthermore, the colour evokes a mood which can be used to complement the data story or make the visualisation feel more trustworthy. For graph types of spatial data, it depends on what type of question the author wants to answer. However, for an insightful data story, the charts must be incorporated in a way that makes sense to the original story. By furthermore adding unique visualisations the story will be more memorable, but the author must be aware to not go overboard with them and increase the complexity. At its core data visualisations should be as simple as possible and take into account the common pitfalls of dealing with visual encodings such as colour.

### 2.2.2 Storytelling through data

Storytelling has been used throughout a variety of media (books, movies, games etc.) however most stories are built upon the same elements. There are five common elements which can be identified in any story: 1) plot, 2) conflict, 3) character, 4) theme, and 5) setting [19]. These five elements can also be applied to data stories.

The first element is the plot, it means what is happening and why it is happening. When looking at data stories this is the topic of the visualisation. The second element is the conflict, which is the problem or phenomena you wish to highlight. The theme is the central idea of believe of the story. Finally, the setting is the time and place of the story. These five elements can also be applied to data storytelling. The character (or

[1]original image cropped by author to only display relevant information

subject) is the who or what that experiences the conflict and develops throughout the story. For a data visualisation using governmental data sets. In a data story it is who or what the conflict is about. The character of your topic does, but it is important that your target audience cares about the subject. Riche, Isenberg and Carpendale point out that it is important to remember that the audience is likely composed of more than one reader. The variety of backgrounds and levels of interest will differ [20]. As a whole, these five elements can be used as guidelines for creating a good story using data.

### 2.2.3 Evaluating data stories

There are several methods to evaluate different metrics of data visualisations. The first method is usability testing, a form of testing that is most commonly used for prototype testing but has also proved itself to be a useful method for evaluating interactive data visualisations [21]. Usability metrics such as tasks completion time and error rate give an indication whether or not the user can easily interact and complete certain tasks without running into any problems. The main application of usability testing is only for interactive visualisations where the user has clear goal and the user is able to perform actions on the data visualisation such as filtering data or changing the time scale. However, Plaisant recommends usability testing can still be used for static visualisations. It can be done by comparing the amount of time spent to retrieve a certain piece of information from the visualisation [22]. Additionally, Sonderegger and Sauer used usability testing to determine the effects of visual appearance on the perceived attractiveness [23]. As such, usability is not only useful to evaluate the performance of certain tasks, but also determine attractiveness when comparing two different versions.

Another method is based on user and expert interviews. Linton advocates the use of more qualitative data to evaluate, since quantitative data such as completion time and error rate, does not indicate whether the reader got the message of the visualisation [24]. Individual interviews can be expanded to a focus group of a few people who sit together and discuss the things that they like and dislike about the visualisations.

All in all, the evaluation methods all proved to have different strong and weak points. Usability testing is commonly used to gather quantitative results such as time taken and error rate based on specific goals. On the other hand, user and expert interviews can give qualitative results on the effectiveness of the visualisation. The methods together offer a both quantitative and qualitative analysis of the data visualisation and its core message. As such, throughout the design processes multiple evaluation methods can be combined to give a full insight into any confusing elements and other aspects that limit the message of the visualisation.

### 2.2.4 Conclusion

To summarise, this review outlines three elements of creating an insightful data story. The elements that have been discussed are: 1) the creation of insightful data visualisations, 2) the application of different storytelling techniques to data storytelling and 3) the effectiveness of different data story evaluating techniques. First the importance of visual encodings and graph types was discussed. Looking further into visual encodings, colouring was identified as an important encoding that is often misused. The use of

colour scales should be avoided when accurate values need to be displayed. colours have proven to influence the mood of users reading visualisations. Similarly, colours can influence how memorable a story using data visualisations. Despite this colour blindness is often not taken into account and can lead to misinformation. When looking at graph types, unique graphs have shown to increase memorability and capture the attention of the reader. The choropleth is a useful graph type to visualise geospatial data although it must be taken into account that it generalises data to certain areas. All in all, the author themselves has to determine what level of depth they want to offer. Depending on the question the author wants to answer, the right type of visualisation can be chosen.

To answer the question posed in the introduction of this literature review on "How to create and evaluate an insightful data story?". The author must pick the right visual encodings suited for their visualisation and be aware of the up and down sides of the graph types he chooses for the data visualisations. To tell a successful story, the author needs to define a target audience and pick a problem that is relevant to them. The story should start with a brief introduction to the problem and build upon to a final conclusion. The final story can then be evaluated through user testing or user interviews for both quantitative and qualitative results on the effectiveness of the visualisation.

While this literature review is mainly focused on the design part of insightful data stories. An interesting further research area might be truthful stories. During this research, the paper of Brewer and Alan was found which discusses that through unintentional mistakes, a wrong message might be visualised [8]. Since many data visualisations are made public it is important that they present the data in a truthful way, as they can be used by others to support false claims.

## 2.3 Existing tools

### 2.3.1 PDOK viewer

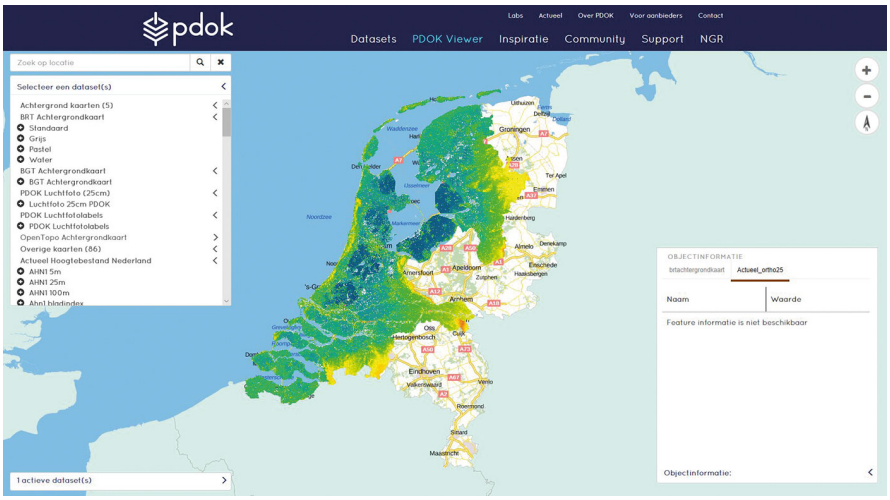



Figure 2.3: The PDOK viewer opened in a browser

The PDOK viewer is made by the Kadaster to quickly explore the numerous of spatial data sets they have to offer. It is a web-based interface which shows a map of the Netherlands on which multiple data sets can be displayed. The user does not require any knowledge about the underlying queries or API calls. By simply clicking on the menu on the left side, the user can select from a drop-down box which data set he or she wants to view. In the bottom left corner there is a legend and, in the bottom right corner, the user can find additional information when he/she clicks on an element on the map. The viewer is a nice example of a dashboard that allows inexperienced users to explore the data. A downside of the viewer is that the only types of visualisation seem to be choropleths, static points, or regions. It is not possible to for example compare to regions side by side or view a different kind of chart such as for example a bar chart which sorts all the regions by value. This would make it easier to spot trends or outliers.

### 2.3.2 Sparklis

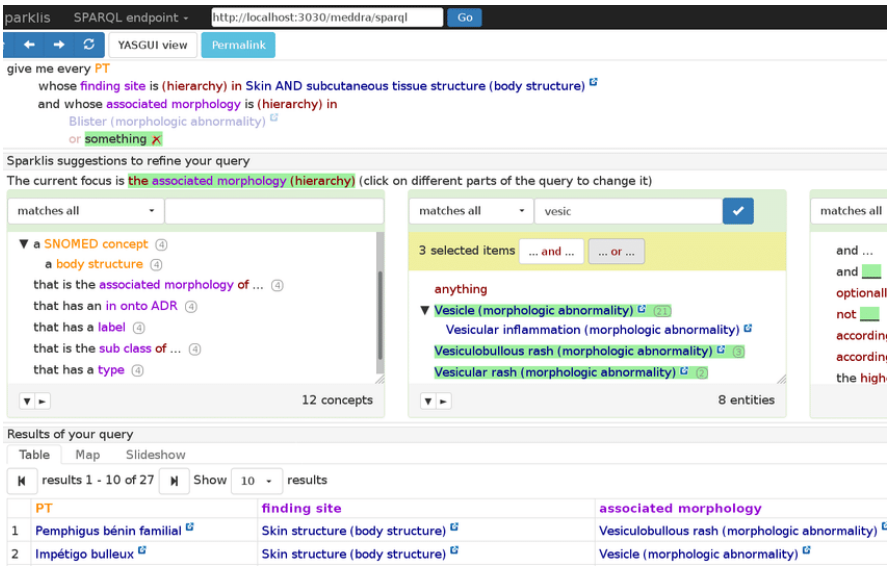



Figure 2.4: Sparklis Web GUI

The Sparklis web GUI, created by Sebastien Ferre, allows users to build SPARQL queries by selecting elements of a sentence instead of coding. It can query sixteen different SPARQL end points, including popular data sets such as DB-pedia and also the data sets from the Kadaster. By picking types and relations from a list of examples from the data set, the user can build a sentence which describes the type of query he wants to run. For example: "| give me every building | that is a house | and exists in Amsterdam | and is built before 2000|". Every part of the sentence is selected via the interface. When the user is satisfied with his or her query, he can switch to the SPARQL view and see the underlying SPARQL query.

### 2.3.3 Facet browser

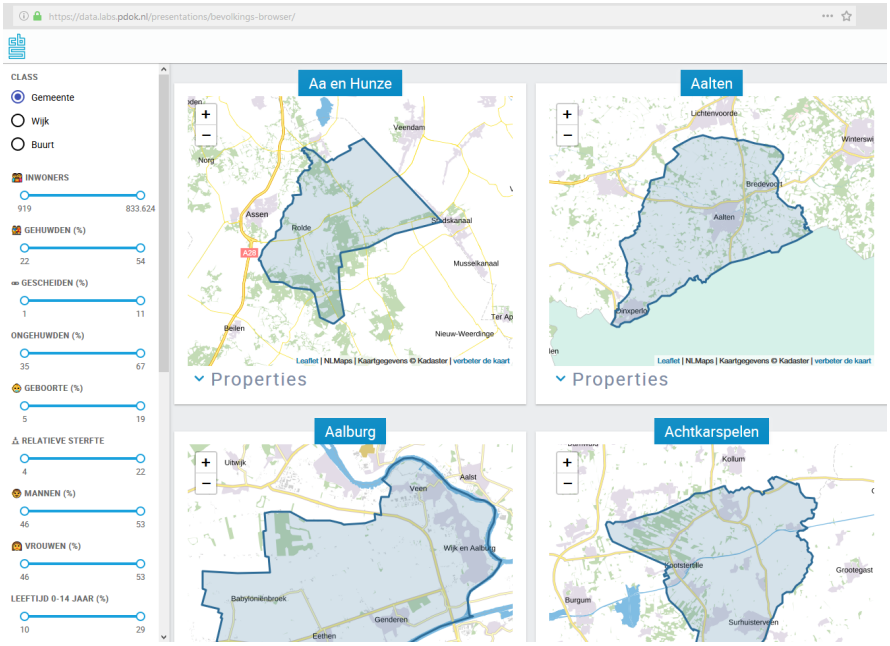



Figure 2.5: The "bevolking" (e.g. population) facet browser

The Kadaster has multiple facet browsers which allow you to filter regions based on certain parameters. The population facet browser for example has all kinds of filters related to demographics. For example, you can view all the regions with a female population between 30% and 50% or the regions with an average number of residents per household bigger than five. Also by clicking on the region, you get the details on all the values that are recorded for that region. This tool is great for quickly finding outliers or focusing on regions that satisfy specific criteria. At the same time you can also compare the properties of regions and see if there are any similarities between certain properties.

### 2.3.4 YASGUI

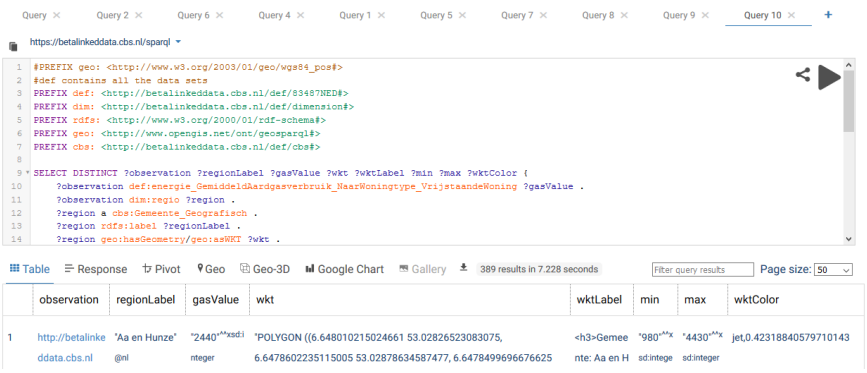



Figure 2.6: Query being inputted into YASGUI (Yet Another Sparql GUI)

The YASGUI, which stands for Yet Another SPARQL GUI, is a text box in which the user can code SPARQL queries and run them on a desired SPARQL-endpoint. The result can be displayed in tabular form but also a variety of other visualisations such as plotting the records on a map. It also supports Google charts, so it can create a basic variety of graphs. One of the downsides is that the GUI does not have many data transform capabilities so the query has to be exactly right and in one single query because combining results is not an option. But that is mostly because it is mostly for data exploration.

# 3 | Methodology

This chapter mentions the methods that have been used in this thesis. The purpose of each method is briefly explained including how it will be applied within the project.

## 3.1 The Creative Technology Design Process

In this bachelor thesis a systematic approach will be used to develop the final solution to the research question. The approach that has been chosen is the Creative Technology Design Approach [25]. It is part of the bachelor Creative Technology and consists of four main phases: ideation, specification, realisation and evaluation, as can be seen in figure 3.1.

The approach focuses on iterative design and applies the divergence-convergence model. First a long list of possible solutions is created during brainstorm sessions, no matter how unrealistic the solutions might seem at first. These solutions are used as inspiration for other solutions and throughout the design process and the less suitable options are eliminated. Each phase also focuses on spiral models, meaning that the processes within a phase affect each other. As such, each of the steps within a phase is gone through multiple times. The four phases each serve as chapters in the structure of this report. What is discussed in each phase is explained below.

### 3.1.1 Ideation

The ideation phase serves as a starting point for finding a good solution to the design question. In this bachelor thesis that means a good solution to show the value of the Kadaster's linked open data sets. This requires analysing the available data sets for possible data stories. The target audience and stakeholders are also taken into consideration. This phase concludes with a low-level prototype, which will be a mock-up of the dashboard and visualisations of the data story.

### 3.1.2 Specification

In the specification phase the preliminary idea are extended into a fully-fledged list of requirements. These requirements will be used as guidelines during the realisation phase. Finally, in the evaluation phase it will be check if the actual solution aligns with the earlier posed requirements.

### 3.1.3 Realisation

The realisation phase will work out the requirements of the specification phase into a complete end product. This includes the selection of the required technologies to build the end product. The end product is then decomposed down into components. The main part of this phase will be about the realisation and integration of these components. Each of these components is also evaluated and changed accordingly if it does not align

with the requirements. When the end product is finished, the design process moves on to the final evaluation phase.

### 3.1.4 Evaluation

The evaluation phase checks if all the requirements from the specification phase have been met. Additionally, user tests are performed to test the usability of the end product. If the requirements are not met or big errors are discovered during user testing, the product goes back into the realisation phase and a new interaction of the product will be made. Once all user tests are passed on a satisfactory level and the requirements have been met, the Creative Technology Design Process will be complete.

## 3.2 Requirements Elicitation

The requirements of a product must be clear before its developed. The requirements are what determine if the final product is good enough or needs to be revised.

### 3.2.1 Functional and Non-Functional Requirements

There are two types of requirements: functional and non-functional requirements. Functional requirements specify functionalities that need to be part of the final solution. Non-functional requirements are based on the design and looks of the final solution. They mostly influence the quality and the environment of the final solution.

### 3.2.2 MoSCoW Method

The MoSCoW method is used to prioritise the requirements into four different categories:

1. **Must have**
   These requirements are vital for the success of the project. When these requirements are not met, the project should immediately be revised or else it should be considered a failure.

2. **Should have**
   These requirements are still very beneficial for the final data story, but will not result in a failure of the project. If they can not be implemented in time, it is advised that a temporary workaround is used to still complete the requirement.

3. **Could have**
   These requirements are nice additions to the system that are wanted by the stake holders or target audience. They do not have a high impact on the overall functionality but mostly enhance it and improve user-friendliness. These requirements can be left out when the deadline for the project is at risk.

4. **Won't have**
   These requirements are explicitly mentioned to not be included in the final result. However, these requirements can still be helpful for later work after the project is finished.

## 3.3 Usability testing & user interviews

The final data story will be evaluated through usability testing. Usability testing was discussed in the literature review in section 2.2.3. For usability testing, users are asked to complete a series of tasks. An observer measures key metrics such as completion time, success rate and number of encountered errors. The metrics are used to evaluate the user-friendliness of the design and identify actions the user has trouble with. Furthermore, the metrics are quantitative data that can be analysed by comparing results of two different designs and seeing which one is more effective.

The usability tests are followed up with a user interview and short survey to gauge if the user things the dashboard has an added value. During the interview the user will be asked for the thing he enjoyed most about the dashboard and if anything was confusing or felt that was missing. The user interview and survey both provide quantitative data. All in all, the combination of usability testing and user interviews will result in a clear idea on whether or not the dashboard was well understood by the user and easy to use.

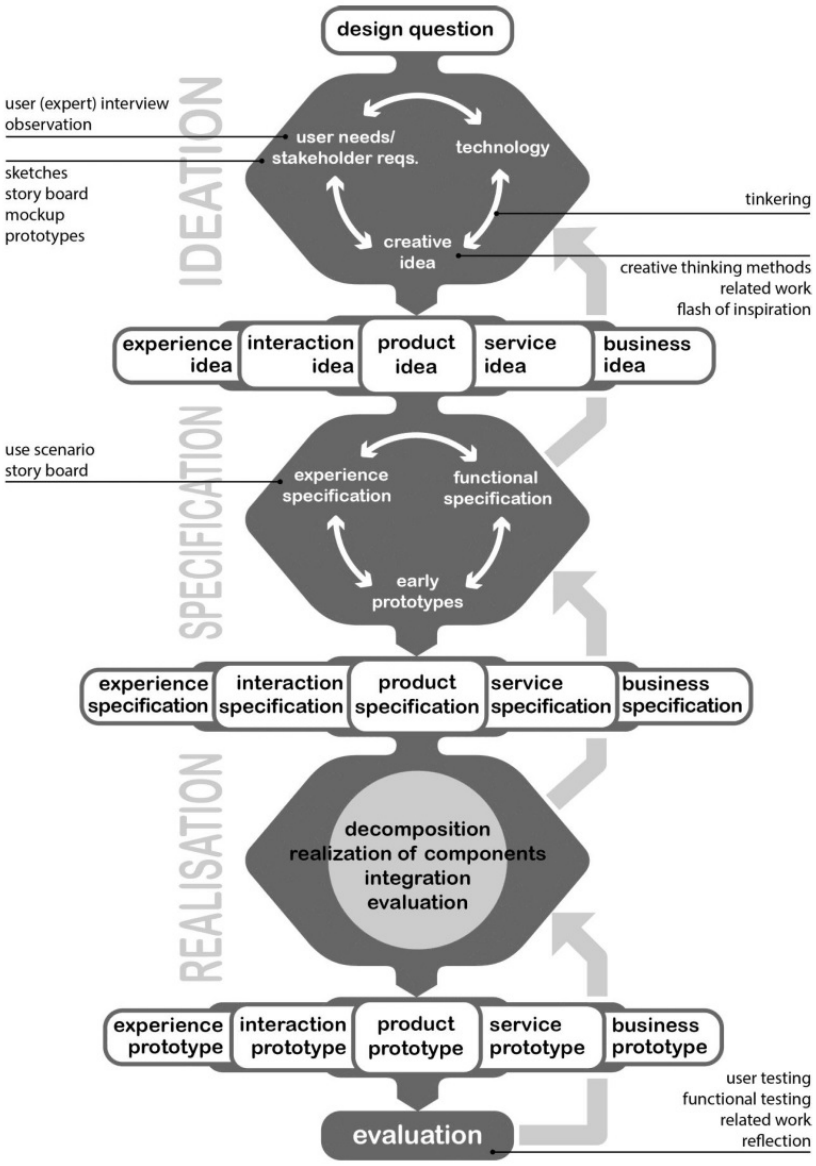



Figure 3.1: Overview of the Creative Technology Design Process[25]

# 4 | Ideation

The ideation phase focuses on the design process of the initial idea. First the available linked open data sets had to be analysed to decide on which one is picked for the final data story. Based on this data set multiple data stories are developed. This is followed by an elaboration on the actual target audience and use case for the visualisations. Finally, the target audience, use case and idea for the visualisation are combined into an initial visualisation prototype in the form of a sketch. The section concludes with an overview of the structure of the data set. This will help querying the data during the next realisation phase.

## 4.1 Exploration of linked data sets - Kadaster

As mentioned in the introduction of this thesis, the Kadaster publishes their open data sets on PDOK. Almost all these data sets are geospatial data sets which are disclosed as Web Map Service (WMS) or Web Feature Service or other well-known geo-services. Not all of these data sets are available as linked open data, in fact there are just three large data sets of the Kadaster. Below is a list which contains an overview of all the data sets which the Kadaster discloses as linked data.

1. The "Basisregistratie Adressen en Gebouwen" (BAG): it contains all official addresses. An address has been assigned a specific naming by a municipality assigned to a public space (ex. street name with a house number and zip code). Each address
2. The "Basisregistratie Kadaster" (BRK): it contains all cadastral registrations (ownership) of real estate including a map of all cadastral plots of land.
3. The "Basisregistratie Topografie" (BRT): it contains many topographic elements of the Netherlands (ex. speed & traffic signs, free standing trees, mussel plantation). One of the popular subsets of the BRT is the TOP10NL, which contains ten important common types of topographic elements such as buildings, infrastructure, terrain and lakes/rivers.

In addition to these three data sets, there is a linked data set available called the "Kerncijfers: wijken en buurten (2016)" i.e. key figures: districts and neighbourhoods. This data set is originally published by the Central Bureau of Statistics (CBS). It contains measures on a big variety of topics ranging from demographic information to energy consumption of different building types. A measure is a property that is measured in a specific region. For example, total population or average income. In the data set there are observations four regional levels available: (1) national; a single observation for the entirety of the Netherlands, (2) municipality (in Dutch: gemeente), (3) district (wijk) and (4) neighbourhood (buurt).

In a collaboration between the CBS and the Kadaster, the data set has been published as linked open data for a single year (2016). The CBS and Kadaster often collaborate as the CBS requires information from the BAG, BRK and BRT on for example region

borders. The goal of making this data set publicly available as linked data is to find out which possibilities arise when publishing CBS data as linked open data.

In the end, the main the key figure data set has been chosen as the main data set for the visualisations. It has been chosen based on the fact that it can be used to give insight into many societal issues such as population growth/shrinkage, energy consumption, crime rates and income distributions. On the other hand, the main data sets of the Kadaster have a more technical orientation. They of course perform a crucial role in both law (ex. plot ownership) and as a source of geospatial information for the Netherlands. However, they already have existing visualisations inside the BAG viewer and PDOK viewer which are already actively being used within municipalities by using tools developed by the Kadaster.

Furthermore, the key figure data set has the advantage that it is not only interesting for municipalities for policy making, but also regular citizens of the Netherlands that are interested in knowing a bit more about their neighbourhood. In addition to this, the only visualisations that exist of the key figure data set is published on StatLine, a service offered by the CBS. StatLine however is limited to line and bar charts and does not offer a way to compare regions or look at the development of individual municipalities online. The CBS only gives entire year progressions about measures of the entire Netherlands. There lies a lot of potential in looking at the developments of municipalities, or at an even lower level, the developments within neighbourhoods over the years. Additionally, it can offer a way to compare developments of similar regions and see if there is a trend among specific kinds of regions. The next section will look into possible data stories that can be told using the measures of this data set.

## 4.2 Possible data stories for CBS Kerncijfers wijken en buurten.

The key figure data set of the CBS contains a total of 153 measures. These measures can be organised in different categories, of which the biggest four are: (1) population, (2) Housing, (3) Energy, (4) Income. By combining the data from multiple measures, it is possible to answer complex questions about the region. The end result is shown in the large table of figure 4.1. Appendix .1 contains a complete list of all the available measures in the data set.

## 4.3 The idea: The CBS & Kadaster Data Dashboard

Since there are so many data stories that can be created using the key figure data set, the final idea is to make a tool to fully explore and visualise the key figure data set. The user is then able to select the measures and regions relevant to what they want to know. The end product will be a web application called the CBS & Kadaster data dashboard. This solution is a trade of between having one specific story that only applies to a single region that is really detailed and having a more global data story that applies to all regions in the Netherlands. In conclusion, the final dashboard facilitates the full exploration of the key figure data set and allows for insight into measures about all different regions in the Netherlands.

### 4.3.1 Target audience & use case

The main target audience of the dashboard consists of both the governments of the municipalities and the government of the Netherlands, or more specifically governmental organisations dealing with managing all the municipalities in the entire country such as the Kadaster. For municipalities this dashboard is useful to gain more insight into the developments of their districts and neighbourhoods. The government can use it for naming and shaming of municipalities that are under performing. For example, if the municipalities are asked to implement measures to reduce overall electricity usage, this dashboard can provide insight in which municipalities have improved the most and those who did not improve.

Additionally there are two secondary target audiences: researchers and citizens of the Netherlands. For researchers this tool can be used to quickly gather data on relevant data and explore relationships on measures about the entire Netherlands. For regular citizens of the Netherlands, this tool can be used to learn more about their own neighbourhood and municipality. The dashboard can show the more busy districts in the city or provide information on the amount of immigrants inside a certain neighbourhood. These measures can be of importance when buying a new house and picking a suitable neighbourhood in a new city for example.

## 4.4 Brainstorming and feedback session

The initial design was developed during multiple brain storming sessions as well as specifying the requirements that are discussed in chapter 3. During the brain storming sessions a low-level prototype was created. This prototype is an initial sketch of the main page of the dashboard, as can be seen in figure 4.2. Additionally, a visualisation was made using a SPARQL query to retrieve the values of gas usage of all municipalities in the Netherlands, as can be seen in 4.3. The initial idea for the dashboard was presented at the Kadaster together with the design sketch and an initial visualisation build with YASGUI.

The presentation was a success and people responded eagerly with response on what measures would be interesting to explore. Some of the required functionalities where discussed, such as interactive filters, the option to export the queried data to a CSV file (Commma-separated values) and a filter to only highlight specific regions of interest. This discussion is later turned into a full set of requirements. In the end, this presentation was a green light to continue with the development of the CBS & Kadaster data dashboard.

## 4.5 Data layout of Kerncijfers: wijken en buurten

Before starting with the actual development of the data stories, it is important to understand the ontology of the data set. Ontology refers to the structure of the data set; meaning the underlying relationships between the concepts and their properties. Understanding the structure will makes it easier to query the data later on. Since the data set is available as linked open data, all of the concepts, measures and records of the data

set can be accessed via URIs. The main link for the linked open data of the CBS is https://betalinkeddata.cbs.nl/.

Currently, this page is exclusively for the key figures data set, since it is the only linked open data set of the CBS. The link can be opened in any browser to view the main page of the data set. The web page can be used to explore the data and its underlying concepts, also called classes. Each of these classes has a unique URI for each of its instances and can have relationships to other classes. When a URI is entered in the browser, it displays the information of that object and the relationships it has to other classes. This is in accordance with the four principles of linked data of Berners-Lee mentioned in section 2.1.1. Below is a list of all the unique classes and their relevant relationships to other classes. The URIs can also be used to explore their relationships to other classes.

**URI of the main data set**

The key figures data set can be found using the following URI:
http://betalinkeddata.cbs.nl/id/dataset/83487NED.
If the CBS added more data sets they would be accessed via a different unique ID at the end of the URI: http://betalinkeddata.cbs.nl/id/dataset/[unique_ID].
*Relationships:*

identifier: the unique ID of this particular data set.
periode: a time period to which the measurements of this data set apply to.
beschrijving: a summary of what is in this data set.

**URI of data set slices**

The main data set consists of slices for each separate measure. A slice is a data set which is a subset of another data set. The slices can be retrieved using the following generic URI: http://betalinkeddata.cbs.nl/83487NED/id/slice/[measure_name]
An example: http://betalinkeddata.cbs.nl/83487NED/id/slice/bevolking_Geslacht_Vrouwen
*Relationships:*

inDataset: the data set of which this data slice is a subset of
unit: references a unit of the measure of this data slice.
observation: links to a observation part of this data slice.

**URI of observations**

Each slice consists of observations about their respective measure. Each observations has a region code and an associated value. Each observation can be retrieved using the following generic URI: http://betalinkeddata.cbs.nl/83487NED/doc/observation/{measure_name}_{regioncode}
An example: https://betalinkeddata.cbs.nl/83487NED/doc/observation/SterfteTotaal_26_BU16800700
*Relationships:*

inObservationGroup: the data slice this observation is part of
unit: the unit of this observation (always equal to the unit of the data group)
region: the region to which this observation applies.

[measure definition]: this variable relationship is a type of measure definition. This connects the observation to a specific value. The relationship measure defines what the value expresses. So a relationship of "Number of inhabitants" means the connected value expresses the number of inhabitants. The next paragraph lists the URIs of the measure definitions.

**URI of measure definitions**

The relationship between an object and subject is often also an URI. The CBS extends the RDF Data Cube with relationships for each measure. Every observation has a defined measure relationship to an object which contains the value of the measure. These definitions can be accessed by the URI: `https://betalinkeddata.cbs.nl/def/83487NED#{measure-name}`

An example: `http://betalinkeddata.cbs.nl/def/83487NED#bevolking_AantalInwoners`

*Relationships:*

unit: the unit of this measure definition

description: a textual description of the measure which makes clear what the measure entails. For example, for gas usage, it specifies that it is only for private gas usage (so the gas usage of companies has not been account for).

**URI of measure units**

Each measure has an associated unit, the different units can be accessed using this URI. Each unit has two properties: a full name, and a symbol as shorter notation.

An example: `https://betalinkeddata.cbs.nl/cbs/doc/unit/VoertuigPerHuishouden`

*Relationships:*

full name: the name of the unit (ex. "persons per $km^2$" or "kilowatt per hour")

symbol: the symbol notation of the unit (ex. "persons/$km^2$" or "kWh")

In addition to this textual description of the data set, there are also tools that can visualise the classes of SPARQL endpoints. The tool LD-VOWL visualises extracted information from these endpoints using the Visual Notation for OWL Ontologies, or abbreviated as VOWL[26]. The SPARQL endpoint of the key figure data set has been visualised in figure 4.5. The web tool LD-VOWL provides an interactive interface where the user can move the classes around and click on relationships or classes to view more details. As can be seen, the result is too chaotic to make sense of by just looking at the static image itself. As such, a handmade diagram of the layout has been created as seen in figure 4.6.

The hand made diagram only lists the essential relationships which are necessary for querying the necessary values for the dashboard. The diagram is based on the LD-VOWL visualisation and the explanation above. Certain common relationships such as "rdf:label" have been left out since every classes has a label which is a textual description of on instance of that class. The blue circles represent classes, the white boxes represent relationships and the yellow boxes represent literal values. The hierarchy "data set -> slice -> observation" can now also clearly be seen. This diagram will be used again in the realisation phase to build the necessary queries.

## 4.6 Conclusion

The final idea is a dashboard where the user is able to query, explore and gain insight into the CBS Key figure data set. The main target audiences of the dashboard are the governments of municipalities that want to gain insight in the developments of their neighbourhoods or districts. The tool is also helpful to identify improving or worsening municipalities based on certain measures the user is interested in. In addition to governmental organisations, it is also an interesting tool for researchers to look for relationships between measures or developments across larger regions of the Netherlands. The tool will also provide the ability for citizens to gain insight into their neighbourhoods and their developments overtime.

The final idea for the dashboard consists of at least two pages: (1) Explore the Netherlands: where multiple regions can be viewed at the same time, and (2) Explore your municipality: which provides insight into multiple districts or neighbourhoods and how certain measures of these regions evolved over time. The design of the first page is outlined inside an initial design sketch which is also the low-level prototype of the dashboard. Finally, this chapter gave an overview on the structure of the key figure data set that is used for the dashboard. The next section will continue with the development of the initial idea.

| Message for data story:<br>**Which regions/cities…** | Related measures. | Issue / topic. |
|---|---|---|
| … are growing the fastest?<br>… are shrinking the most?<br>… have an increasing old-age dependency ratio? (Often called 'grey pressure' in Dutch) | Population age groups.<br>Population gender.<br>Population size. | Urbanization.<br>Ageing of the population. |
| … have an increase in the number of foreigners? | Immigrants (western).<br>Immigrants (non-western). | Immigration crisis of Europe (2015). |
| … have increasing amounts of people feeling lonely? | Single person households.<br>Average household size. | Loneliness. |
| … have a shortage of affordable houses for starters? | Houses for rent.<br>Houses of private owners. | Housing market. |
| … have an increase in the number of electric cars per household?<br>… have implemented climate change measures that actually helped decrease energy consumption?<br>… are expanding their renewable energy sources? | Electricity and gas usage per housing type.<br>Solar energy generation.<br>Cars per household. | Climate change. |
| … have the most even income distribution?<br>… have a growing wage gap between the rich and the poor? | Percentage of households with income of lowest 20%.<br>Percentage of households with income of highest 20%. | Unfair income distribution. |
| … have a lot of poor families?<br>… have a lot of house wives?<br>… have decreasing incomes? | Average income.<br>Average income from secondary sources.<br>Average working persons per household. | Poverty. |
| … have the highest amount of theft cases per inhabitant?<br>… have increasing amounts cases of assault?<br>… are the most dangerous to live in? | Theft.<br>Vandalism.<br>(Sexual) Assault. | Crime. |

Figure 4.1: Table with possible data stories for the key figure data set after multiple brainstorm sessions.

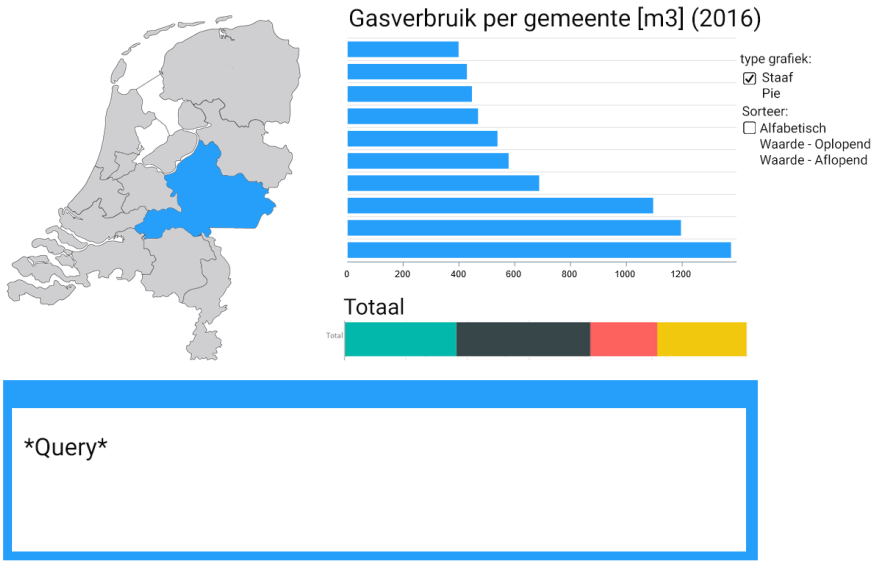



Figure 4.2: The final prototype: a design sketch of the first dashboard page

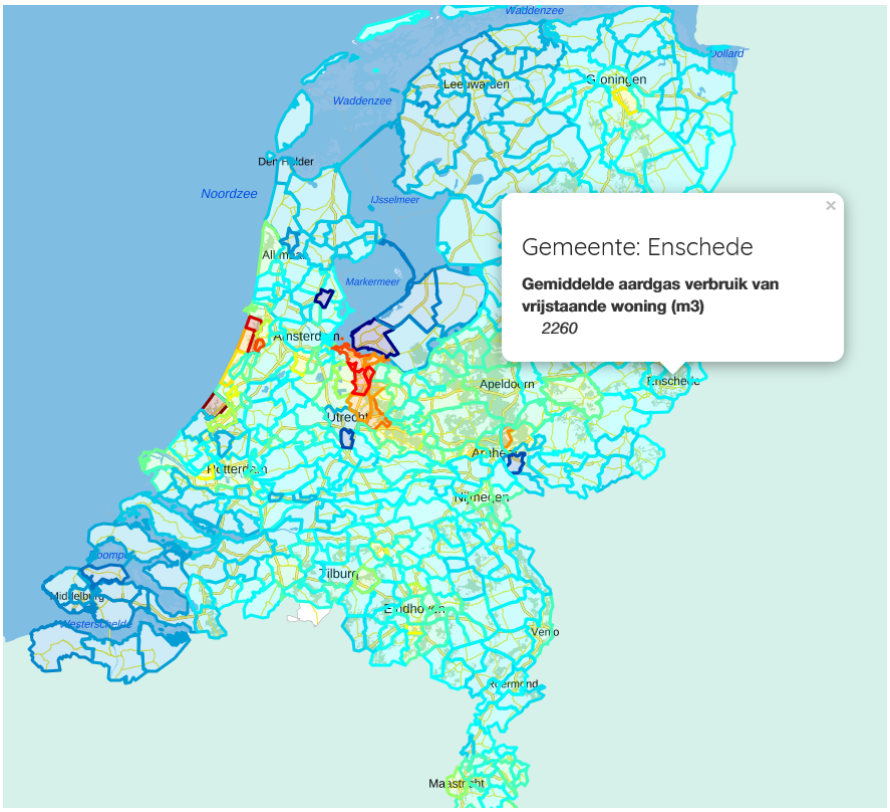



Figure 4.3: An initial visualisation made with YASGUI, plotting a region for each municipality with the average gas usage per free standing house.

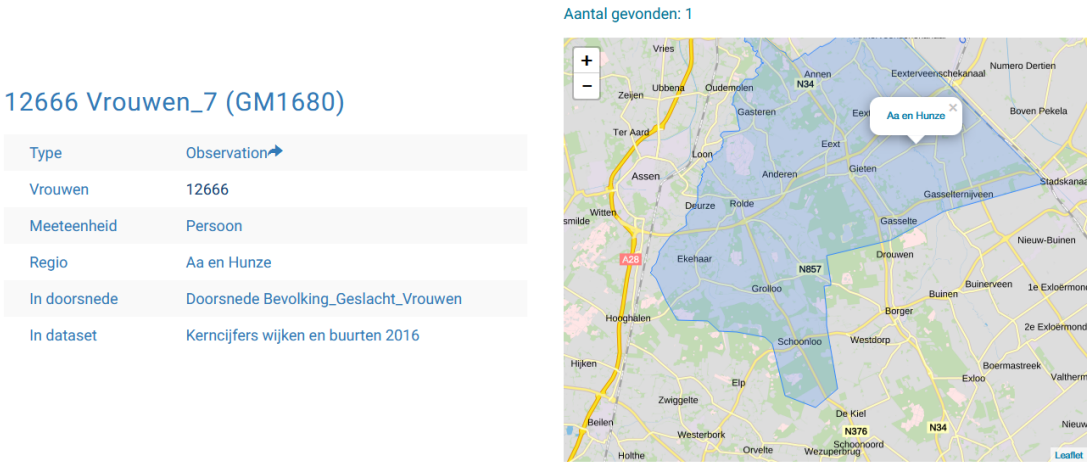



Figure 4.4: URI of the province Utrecht entered in the browser

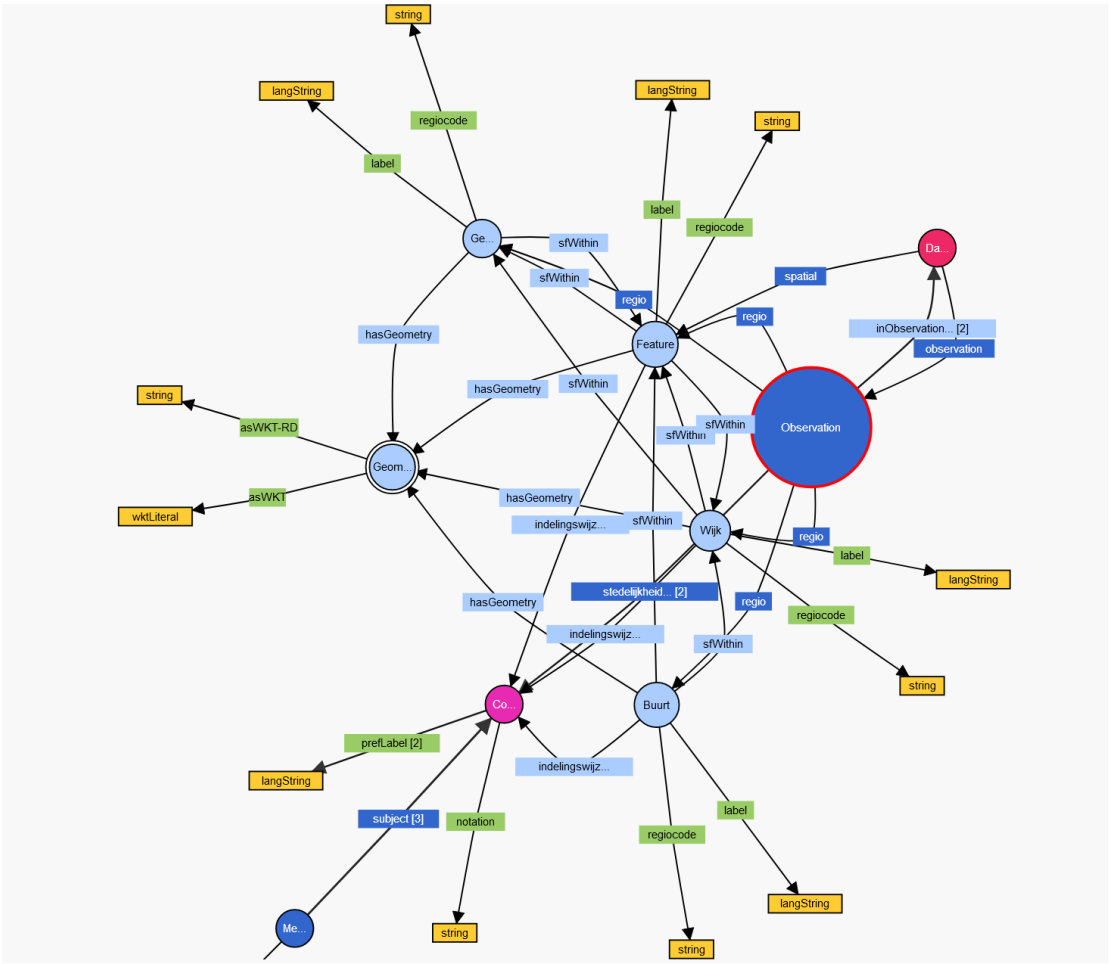



Figure 4.5: Visualisation of the linked data classes by LD-VOWL

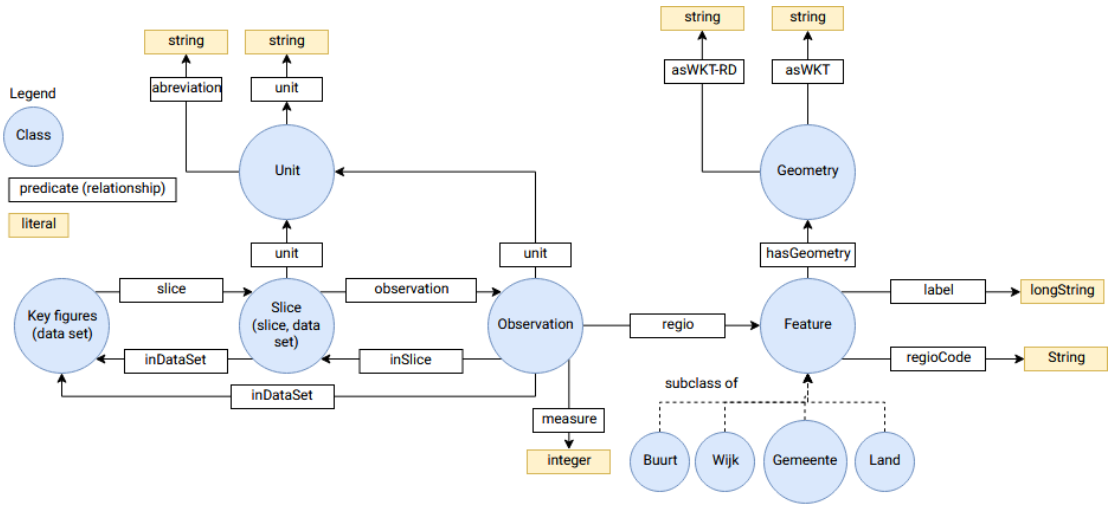



Figure 4.6: Custom made diagram representing the layout of the data set.

# 5 | Specification

In this chapter, the requirements which guide the creation the dashboard are specified. This will be done according to the specification phase from the Creative Technology Design process as described in section 3.1.

## 5.1 Requirements

In this section, the requirements for the dashboard and visualisation project are listed. As mentioned in the methodology, there are two types of requirements functional and non-functional requirements. Finally, the requirements are split using the MoSCoW method and have their rationale explained via a brief comment highlighted in italics.

### 5.1.1 Functional requirements

**Must have**

FR1 The dashboard can retrieve all available measures / variables.

FR2 The dashboard can be sort values alphabetically and numerically (ascending/descending).
*This will allow the user to more quickly identify outliers and spot*

FR3 The data is all retrieved from the source through SPARQL (not stored locally).
*As part of the motto of the Kadaster: "data at the source", the data will be retrieved dynamically each time the page is loaded to highlight the linked data aspect*

FR4 The regions all have there boundaries plotted on a map.
*This allows the user to identify certain trends between large regions of the Netherlands*

FR5 The URIs of the objects can be accessed in the visualisation.
*This strengthens the message of linked open data.*

FR6 The visualisation can be sorted alphabetically and numerically (ascending/descending).

**Should have**

FR7 The visualisation can display the same variable multiple years.*Allows for the insight in the development of regions over time*

FR8 The dashboard can limit query results.
*Necessary to make sure large queries do not crash the browser*

FR9 The dashboard has an option to view the queries.
*This also strengthens and lays the connection between the dashboard and linked open data*

FR10 The dashboard has an option to view raw query results.
*idem as above, also the user to see the URIs*

FR11 The dashboard has an option to export results to CSV.
*Much requested feature by Kadaster employees, this allows users to work with the data using their own tools.*

FR12 The dashboard must work in the latest version of Chrome, Firefox and Safari.

**Could have**

FR14 The dashboard has an option to change the chart type of the visualisation.

FR15 Parse all region names correctly and attempt to catch any spellings mistakes.

FR16 The dashboard should have a filter that has auto-complete for the region values

FR17 Regions can be filtered on the map using a lasso selection tool.

FR18 The visualisation shows a percentage increase over a selected time period.

**Won't have**

FR19 The user can run its own query on the data visualisations.

FR20 The user can make region selections on the map by dragging its mouse (box or lasso selection etc.).

### 5.1.2 Non-functional requirements

**Must have**

NFR1 The dashboard language is available in Dutch. *Since dutch is the main language used on the website of the Kadaster.*

NFR2 The dashboard's filters should be easy to use.

NFR3 The dashboard should be displayed even though the data has not loaded yet.

NFR4 The visualisation should match the house style of the Kadaster.
*Because the final dashboard will be published on the PDOK Labs environment of the Kadaster.*

NFR5 The visualisation highlights the correct chart element when user selects a chart element, it must also high.

NFR6 The user receives feedback (an error) if the query fails.
*Feedback is important so the user knows if the action was successful or not.*

NFR7 The visualisation should use different colours for different regions (if multiple are plotted).

**Could have**

NFR8 The dashboard language is available in English.

NFR9 The dashboard has a progress bar when data is being loaded.

NFR10 The visualisations should provide feedback when the data is being loaded.

NFR11 The dashboard and visualisations are responsive and load correctly on mobile devices.

NFR12 The dashboard has a colour blind mode.

**Won't have**

NFR13 The dashboard can change ability to change the layout (move elements around); *while a custom dashboard would essentially allow inexperienced users to create data stories dashboards, it lies out of the time scope of this project.*

NFR14 The dashboard can change styles (different colour, night mode).

# 6 | Realisation

The realisation phase will give a detailed overview of the development of the dashboard. First, a brief explanation is given on the technologies that are used in this project. This is followed by an overview of all the pages and their components which need to be realised. Each of these sections elaborates upon how the technologies are used to realise the components. Finally, each section provides insight in the design choices that have been made during the building process.

## 6.1 Technologies related to the project

This section explains the different technologies and software libraries used during the development of the final dashboard. The final dashboard is a web based application and as such most of the coding is done in HTML, CSS and JavaScript. No JavaScript framework was chosen as this did not fit the small scope of the dashboard. Instead, a combination of pure JavaScript, HTML and CSS is used with the addition of software libraries.

### 6.1.1 SPARQL - Linked data queries

The SPARQL Protocol and RDF Query Language is used to query data from linked open data sets. The SPARQL endpoint of key figures data set is `https://betalinkeddata.cbs.nl/sparql`. The SPARQL version used in this thesis is V1.1, which is currently the latest version. A basic query can been seen in figure 6.1. Each query consists of a `SELECT` command followed by the variables that need to be part of the results table. In SPARQL, a variable always starts with a question mark. The asterisk is a wildcard character which means return all variables as part of the result.

The `SELECT` command is always followed by the `WHERE` command which specifies a condition for the information is retrieved from the data set. This condition is most commonly a triple consisting of one or more variables. In the example, a triple of three variables `?sub`, `?pred`, `?obj` is given. The variables correspond to the subject, predicate and object of a triple. Since all three parts of the triple are set as a variable, it will return all triples in the data set. As such, the `LIMIT 10` command is used to limit the results of the data set to a maximum of ten results. When an URI is used instead of a variable, the query will only return the triples containing that specific URI.

The order of the arguments is also important, as the first argument refers to the subject, the second argument to the predicate and the third argument the object. For example, by swapping the second argument `?pred` with `rdf:type`, the query will return the triples where the predicate is a `rdf:type`. Here, `rdf:type` uses what is called a prefix. SPARQL uses the `PREFIX` command to define prefixes of URIs, so the full URI name does not have to be written out in the query. This is helpful when multiple URIs in the query have the same base URI.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT * WHERE {
  ?sub ?pred ?obj .
} LIMIT 10
```

Figure 6.1: Simple SPARQL query returning 10 triples

### 6.1.2 YASGUI - YASQE (a SPARQL Query Editor) and YASR (a SPARQL Resultset Visualizer)

YASGUI was already briefly touched upon in section 2.3.4. It is a JavaScript library that consists of two parts YASQE, a SPARQL Query Editor, and YASR, a SPARQL Resultset Visualizer. The query editor of YASGUI (YASQE) is mainly used in this project to show what queries are running to retrieve the data for the dashboard. It offers among others syntax highlighting and auto complete in a web-based SPARQL editor. The SPARQL query can then also be executed on a SPARQL endpoint.

YASR can then be used to visualise the results inside of a table, on a map using Leaflet or in a Google Chart. YASGUI was chosen because it is already a well-established SPARQL query editor and it is also what the Kadaster uses inside of PDOK Labs.

### 6.1.3 D3 - Combining HTML, JavaScript and SVG for visualisations

The D3 JavaScript library is used to create data-driven HTML-pages, most often through scalable vector graphic (SVG) or direct manipulation of other HTML-elements [27]. In this project it is used to create graphs and tables which update automatically when the underlying data changes. D3 is chosen because it is the most extensive charting library with a large community which also offers manipulation of the document object model (DOM). The document object model is how every modern browser represents a web page as an object which can be manipulate via JavaScript. D3 has built-in functions that help manipulating the web page and creating graphs that are created via SVG elements.

Other charting libraries such as Vega, Chart.js and Google Charts do not offer the same level of flexibility that D3 has to offer. The charting libraries can only create visualisations and not manipulate elements on the HTML-page. The downside is that D3 is overall considered to have a steeper learning curve. This thesis however will provide the necessary technical explanation to understand how the visualisations work, without prior knowledge of D3.

### 6.1.4 Leaflet - Mapping library

The Leaflet JavaScript mapping library is used to plot the geometries of regions on a map of the Netherlands [28]. The Kadaster offers the NLmaps library which also uses Leaflet and includes four types of maps (standard, pastel, grey and satellite photo) and multiple feature layers. The dashboard however will not use the NLmaps library but instead Leaflet will be manually configured. This is because NLmaps adds functionality which is not necessary for the dashboard. By implementing Leaflet ourselves, it keeps

the source code easier to understand and reduces the amount of libraries. Besides this, the Kadaster themselves also provides documentation on how to setup using their WMS & WFS geo-services with Leaflet natively.

### 6.1.5 jQuery & jQuery UI elements

The jQuery library provides functions for commonly required JavaScript functionalities [29]. Instead of writing these functions themselves, developers can save time by using this library. The same goes for jQuery UI elements, which provides a library for UI elements such as sliders. Since maximum performance is not required for this project and time is of the essence, these two libraries are used to speed up the development process.

### 6.1.6 Bootstrap Material Design and Data Tables

The Boostrap Material Design CSS library is used to create the user interface of the dashboard [30]. The library is based on Twitter's Bootstrap framework and styled according to Google's modern Material Design guidelines. The main advantage is Bootstrap's Scaffolding layout based on a grid consisting of 12 columns that makes creating responsive page layouts easier. The colour scheme that is chosen for the Dashboard is blue, as it matches the colours of the Kadaster. Additionally, the Data Tables jQuery plugin is used to created styled HTML tables which can be sorted and filtered [31].

## 6.2 The layout of the dashboard and its components

Based on the final idea and design sketch from the ideation phase in section 4, the dashboard will consist of at least two pages: "explore the Netherlands" and "explore your neighbourhood". During the iterative design process two more pages where added: "explore relationships" and "run a query". The next sections will each cover the realisation of one page and their respective components. Each page starts with a summary of its required functionality. The section of the first page is the longest, as slightly modified versions of similar components are used on the other pages and as such require only little explanation.

## 6.3 Page 1: Explore the Netherlands

The goal of this page is to allow the user to explore the linked open data set on a large scale by comparing multiple regions at the same time. This can be done by selecting a region type and specifying the regions that the user wants to see. By default, if no region type is specified, it will give the user all available results for that region type. The user is able to view the data in a bar chart and this bar chart is linked to a map on which the regions are plotted. Both the map and the bar chart are interactive. The results can be used by users to gain insight in the measures on a national scale for all the municipalities.

### 6.3.1 Querying the data

Firstly, before any of the components of the page are build, the data needs to be queried from the SPARQL endpoint. The query needs to return all observations for a specific measure and only the observations of a specified region type or region name. In addition to this, the query needs to retrieve some textual information about the measure unit and region names which can be used in the graph. Lastly, the geometry of the regions needs to be retrieved so Leaflet can plot the region on the map. All these requirements are in a single query which is shown in figure 6.2. The final query in this figure will help to explain which steps are taken to get the necessary data.

```
1  prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2  prefix geosparql: <http://www.opengis.net/ont/geosparql#>
3  prefix def: <http://betalinkeddata.cbs.nl/def/83487NED#>
4  prefix cbs: <http://betalinkeddata.cbs.nl/def/cbs#>
5  prefix dim: <http://betalinkeddata.cbs.nl/def/dimension#>
6  select ?label ?regioLabel ?region ?value ?ligtIn ?ligtInLabel ?wkt {
7      values ?measure { def:energie_GemiddeldAardgasverbruik_GemiddeldAardgasverbruikTotaal }
8      ?measure rdfs:label ?label .
9      ?observation ?measure ?value .
10     ?observation dim:regio ?region .
11     #Possible regional levels: Land_Geografisch, Gemeente_Geografisch, Wijk, Buurt
12     ?region a cbs:Gemeente_Geografisch .
13     ?region geosparql:hasGeometry ?geometry .
14     ?geometry geosparql:asWKT ?wkt .
15     #Next part adds ?ligtIn, which is the region in which ?region lies (needed for wijken)
16     ?region rdfs:label ?regioLabel .
17     ?region ?pre ?ligtIn .
18     ?ligtIn rdfs:label ?ligtInLabel .
19     FILTER(?pre = geosparql:sfWithin).
20 } limit 30
```

Figure 6.2: Main query for retrieving all observations of all municipalities for the measure gas usage

Section 4.5 explored the layout of the key figures data set. Figure 4.6 shows that all observations have an associated measure definition. This measure definition can be used as a predicate to filter the observations on a specific measure. Line 7, the first line of the `WHERE` command, defines the variable `?measure`, as the measure that the

user is interested in (in this case gas usage). The `VALUES` command assigns multiple URIs to a single variable. This command will make it easier to query information about two or more measures in the future. Also note, the actual `WHERE` command is missing in this query as it is not required to use the command in SPARQL 1.1. It will implicitly be added after the variables of the `SELECT` command. Line 8 retrieves the name label of the measure via the `rdfs:label` predicate, and sets it to the variable `?label`. The next line finds the observations (`?observation`) and their associated values (`?value`) that have this measure as a predicate. For each found `?observation`, it checks for the predicate `dim:regio` which finds the associated region. The result of this first part is all observations and their regions for a specific measure, including the label of the measure.

As the first comment on line 11 suggests, the next section filters the regions based on a region type. Line 12 uses the "a" keyword, which is a shorthand notation for the predicate `rdf:type`. This filters the regions to contain only municipalities (`cbs:Gemeente_Geografisch`). This URI can later be replaced with any of the other three region types. Next, the regions `?geometry` is found via the `geosparql:hasGeometry` predicate. Each geometry has a predicate `geosparql:asWKT` which returns the geometry as well-known text (WKT) for the Web Mercator map projection (EPSG:3857/WGS84), which is also the default map projection Leaflet uses. There also is a predicate `geosparql:asWKT-RD` which returns the geometry for a the "Rijksdriehoeks" map projection (EPSG:28992).

The final section first finds the label of the region, so the region name. A small problem arises when using the predicate `geosparql:sfWithin`. It does not return any results despite the relationship existing for each observation. However, when first all the predicates of a region are queried, and the result is filtered using the `FILTER` command, it does return the correct URI for the region it lies in. The cause to this problem could not be found even when discussed with the Developers of the Kadaster and asking on community forums online. The variable `?ligtIn` holds the region, in which the variable `?region` lies. This is especially useful for districts and neighbourhoods. For a municipality, this is always the URI of the Netherlands. The Netherlands itself does not have the predicate `geosparql:sfWithin` as part of the data set.

Finally, the query includes a limit to make sure not too many records are returned, as there are over 500 municipalities in the Netherlands, which means even more districts and neighbourhoods. Retrieving all neighbourhoods is possible, and the query actually takes less than a minute. However, some browser crashes occurred when the data was being visualised with D3 and Leaflet. This concludes an explanation on the first and biggest query that is able to retrieve the value of any measure, from any region type including the region geometry.

The user needs to be able to select a measure definition. This is done via a drop down menu. The values of this drop down menu are generated dynamically by querying all the possible measure definitions, see the query in figure 6.3. All measure definitions have a `rdf:type` which is equal to the `purl:MeasureProperty` predicate.

Finally, the d3-sparql SPARQL library is used to execute the queries on the SPARQL endpoint. The result is a single array of JSON objects with the variables for each result. Each of the JSON objects contain all the variables that are part of the result. This format allows it to be used with D3 in the next section.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX purl: <http://purl.org/linked-data/cube#>
SELECT ?measureProp WHERE {
    ?measureProp rdf:type purl:MeasureProperty
}
```

Figure 6.3: URI of the province Utrecht entered in the browser

### 6.3.2 D3 Bar chart: visualising the data

The queried data is used to create a visualisation with D3. The basis of D3 is that it create HTML or SVG elements based on records in a data set. This data set must be an array of values or JSON objects. For example, it can create five <rect> elements based on an array with five values. D3 can then assign width properties to these rectangle elements to create a bar chart. When the underlying data changes, the graph needs to be updated. D3 uses a `exit()`, `enter()`, `update()` pattern for this. If the data array is still of the same length, only all existing elements need to be updated with the `update()` function. If there are less records in the data set available, some elements need to be removed with the `exit()` function. Finally, if there are more records in the new data set, the `update()` function adds the necessary extra elements. This also happens the very first time when adding the data set, since there are no existing elements yet created.

This design pattern allows for animating the elements which are removed differently from the new elements that are added. In addition to this the scales are also updated by calling the axis functions. In the source code, the complete function which updates the data is named `d3update()`. The final bar chart can be seen in figure 6.4. One requirement was also that the dashboard was responsive and view able on mobile. This is done via a separate function which recalculates the size and updates all d3 elements. This solution is based on an online example from Brendan Sudol [32]. When the web page is re scaled, a event is fired that will trigger the bar chart to update itself. All updates of the bar chart are animated by transitioning between the old and new state as this helps improve the user experience. All in all, the bar chart is the first component to be finished for the final dashboard but some changes were made throughout the design process.

The first design of the bar chart had vertical bars for each of the regions. This causes problems when many regions are plotted in the graph at the same time, since many of the region labels started overlapping. So instead, a horizontal bar chart was created where the labels are on the y-Axis. The labels are hidden when more than 30 regions are plotted in the same graph. Additionally, it was hard to read the exact values from the axis. This problem is solved by adding a tool tip to each of the elements. When the mouse of the user hovers over a bar element, it highlights this element and gives information about what region this bar applies to and what the value of the measure is.

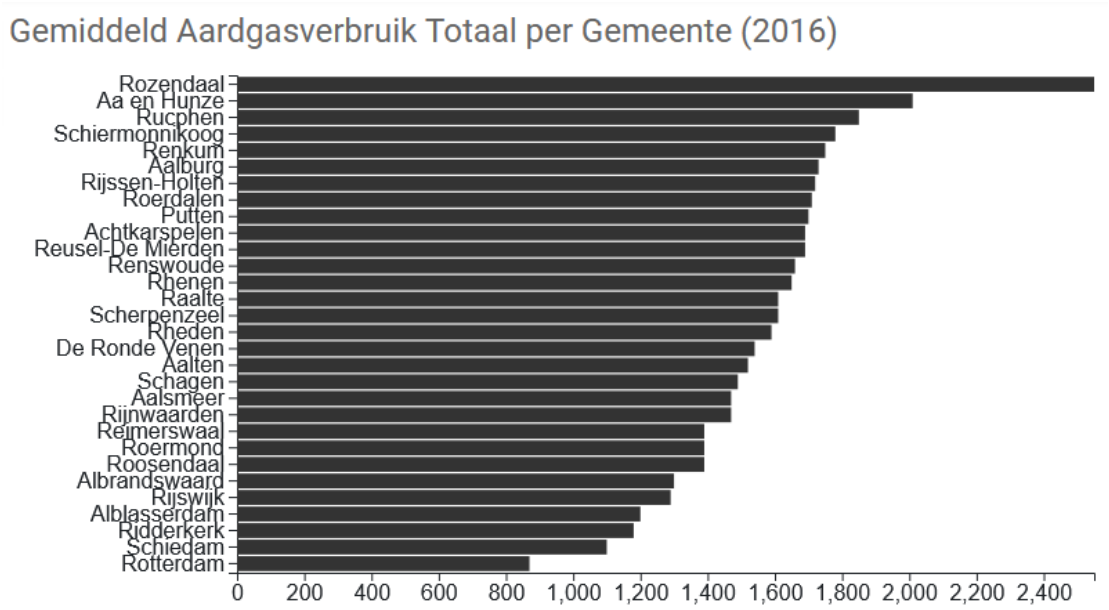



Figure 6.4: Custom D3 Bar chart plotting the query results

### 6.3.3 Side bar with filter options

The user needs filters which allow them to select the measures and regions they are interested in. These filters are part of the sidebar of the dashboard. The first element is a button where the user can start the query with the currently selected filters. The first filter is a drop down menu which contains The drop down is dynamically generated by jQuery using the results from the query in figure 6.3. At first, the label of the measures would be used for the names in the drop down menu. However these labels turned out to not be descriptive enough. Instead, the URI of the measure is used since it contains more information such as the category group followed by a slightly longer name describing the measure.

Below the measure drop down is an input field which allows the user to specify which regions to query. It filters all region names containing that string entered by the user. As the project progressed, an auto-complete drop down was added to help the user pick the region. When a string is entered it gives a maximum of 10 suggestions of municipalities. Multiple regions can be queried by separating them with a comma (ex. "Amsterdam, Rotterdam"). The application automatically parses the string when the user enters a letter and adds the auto-complete box. The drop down menu is currently only active for municipalities due to time constraints of the project. For the different region types it is still possible to filter by name, however no suggestions are given. This functionality can still be added in the future.

Two option sections allow the user to select the region type and sorting function of the visualisation. The region selection consists of four radio buttons, each button corresponding to a different regional level. The user has to press the query button again to update the data set with the new regions. The sorting selection has three

radio buttons which either sort the visualisation alphabetically on the region name, on ascending values or descending values. Changing the sorting selection automatically triggers D3 to sort the bar chart. The sorting process is also animated so the user gets immediate feedback and knows that the graph is updated.

Finally, there is a number input box at the bottom which limits the amount of results. This is necessary for when the user does not specify a region type and the query returns all regions of that type. The largest number the user can pick is 1000. This is enough to at least return every municipality of the Netherlands. The D3 visualisation is able to handle even larger data sets but the wait time of the query also becomes larger with these huge data sets. The limit specified in the box is automatically added to the end of the query. All in all, with this limit in place it is tested that with a good internet connection, the query result is visualised in under 30 seconds. The complete side bar with its filters can be seen in figure 6.5.

### 6.3.4 Leaflet map: plotting region geometries

A Leaflet map is used to plot the geometries of regions on a map of the Netherlands. The map itself is retrieved dynamically from `https://geodata.nationaalgeoregister.nl/tiles/service/wmts/brtachtergrondkaart/EPSG:3857/{z}/{x}/{y}.png`. This URL is a so called Web Map Tile Service (WMTS) endpoint. Leaflet requests the necessary map tiles needed for a specific zoom level. In Leaflet, a new feature layer is generated which contains the polygon geometries of the regions. As Leaflet does not support the WKT format directly, the WKT is converted to Leaflet objects using a plugin called Wicket [33]. When a query is executed, the main JavaScript file checks if the variable `?wkt` is present. If it is, it parses the values and adds them to the Leaflet feature layer which is then displayed on the map.

An eventlistener is assigned to each feature, when a feature gets clicked it displays the relevant region name and value in a tool tip. Each feature is linked to an element in the D3 visualisation via there index in the data array of the query result. This ID is used to highlight both the region and the respective bar chart at the same time when the user browsers of a bar of a region on the map. Finally, to add a distinction between the regions on the map, each map gets a different fill colour depending on its value. The values are mapped to a D3 colour scale which goes from white, for the lowest value, to dark blue for the highest value.

### 6.3.5 Query editor: showing the linked data aspect

At the second feedback session of the Kadaster the initial dashboard prototype was presented. It contained all of the functionality mentioned above: a bar chart, a map and filters. One of the main feedback points was that the linked data aspect needed to be more pronounced. To solve this a live query text box is added. For this the library YASGUI is used which gives us a complete SPARQL query editor out-of-the-box. This Kadaster themselves uses YASGUI to show which queries are run in order to generate the visualisation inside of PDOK Labs.

For the dashboard, YASGUI only needs to display the query. Since the visualisations are already handled by D3, only YASQE, the SPARQL query editor of YASGUI is used.

Some elements of YASQE, such as allowing the user to execute the query are removed. The user can only change and execute the query via the sidebar and not type his own queries. Since YASQE can also execute SPARQL queries, the d3-sparql library became redundant and instead YASQE is used when the user presses the "Execute Query" button. The user can define his own fully custom queries on a separate page, see section 6.6 later on in this thesis.

Since YASQE on its own does not support multiple queries, a navigation bar was added with tabs for each of the queries. When the user wants to view for example the measures query, he can switch tabs by clicking one of the buttons and then the query text box automatically displays the correct query. Besides showing the user how the data is retrieved, the main goal of the query box is to show the user how changing certain filters affects the queries. This can also help understand how SPARQL queries work. The final query box can be seen in figure 6.7. In conclusion, the complete first page can be viewed in figure 6.8.

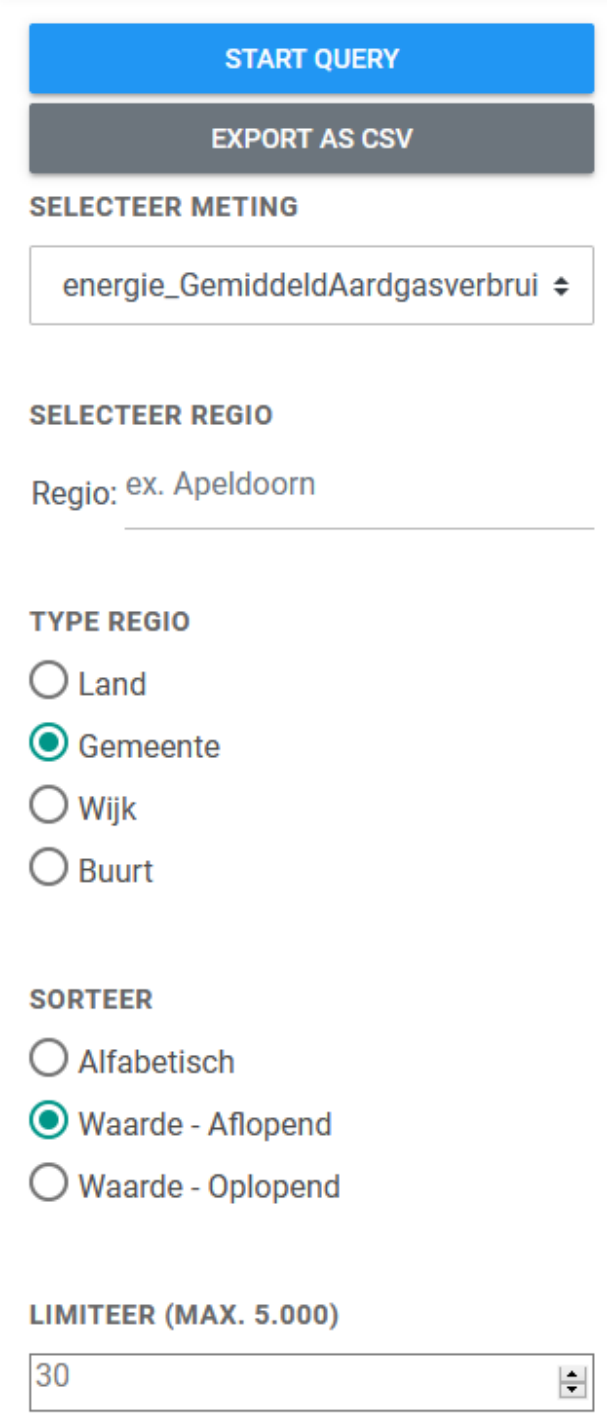



Figure 6.5: Sidebar consisting of filters for the data set


Figure 6.6: Leaflet map plotting municipality regions with a tool tip


Figure 6.7: YASQE query highlighting with multiple tabs for each query

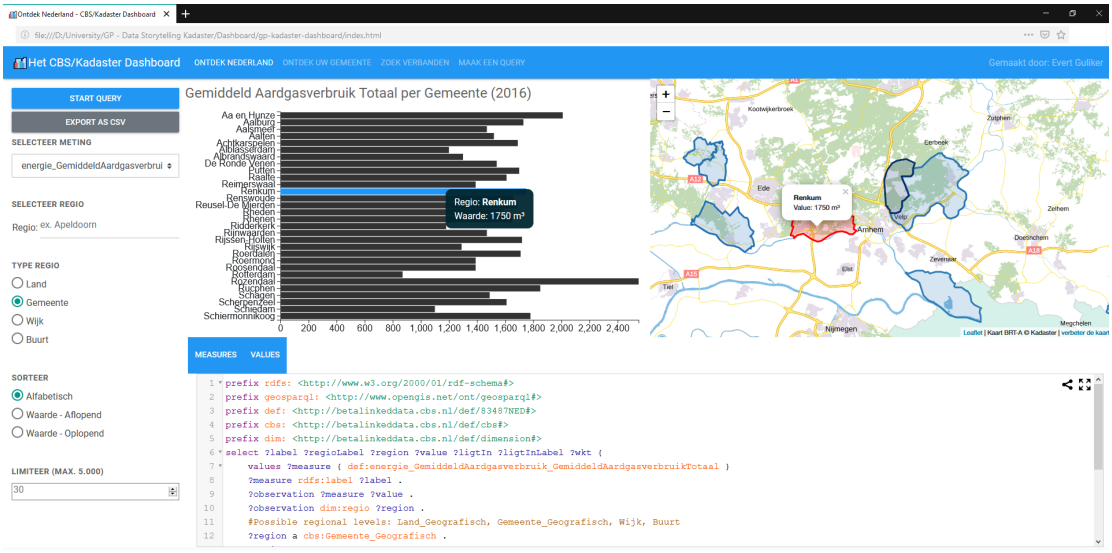



Figure 6.8: Page 1: "Explore the Netherlands"

## 6.4 Page 2: Explore your municipality

This page aims to provide a more in depth look at a specific municipality and its districts and neighbourhoods. For a specific measure, it shows multiple regions in a line chart and how they developed over time. It uses the CBS key figures not only from 2016, but also other years. These data sets are retrieved through the API of StatLine from the CBS. The formatting of these data sets was changed after the year 2012, so for the dashboard only the period of 2013 up until 2018 is retrieved. This saves time since as it would otherwise be necessary to manually map all the measures to their correct definitions for the new and old years of the data set.

### 6.4.1 Data cleaning: district names

Another problem was that after the year 2013 the names of many districts (wijken) changed. Multiple districts where named `"District XX:"`, where "XX" is a random number followed by the region name. Unfortunately, there exist no consistent region codes that can be used as an identifier across multiple years. The best option is to use the names. To make matters worse, there exist multiple variations of these prefixes such as `"WijkO1"` `"Wijk O1:"` or even one with a spelling mistake: `"Wijke 1"`. To remove these prefixes, the following RegEx (Regular Expressions) filter was used: `/^Wijk(e)?(?!aan | bij | en) ?[0-9]?[0-9]? ?:?-? ?/` This filter removes any of the aforementioned prefixes while also leaving the word "Wijk" as part of districts where it is correct and part of the actual name of the region. Examples of such regions are: "Wijk bij Duurstede" and "Wijk aan Zee".

### 6.4.2 D3 Line chart: showing progressions over time

The graph for this page needs to show a progression of a certain measure over time. This can be for either all districts of a specific municipality, or for all neighbourhoods in a district. These progressions over time can be used to find out if certain policies have had their desired effect, for example: "Has gas usage really decreased after the Dutch government implemented a subsidy for induction furnaces?" Many of the data stories mentioned in section 4.2 look into increases or decreases of certain measures. Before the data can be plotted, first the data from different years has to be queried and there are a few other problems too that need to be solved such as changing region names and changing region borders.

#### Merging the linked data with other years

The data from the StatLine API returns a JSON object which can be merged with the results from the linked data of 2016. Unfortunately it is not possible to request an individual measure of the key figures data sets so instead the whole .CSV file is downloaded every time a request is fired. This slows down the query, so instead the files are stored locally on the server. Currently, whenever a new user requests the data, the script checks if the data is up to date. If the data is older than one day, it requests new data from the server. This means that the first user of the day has to wait longer for the

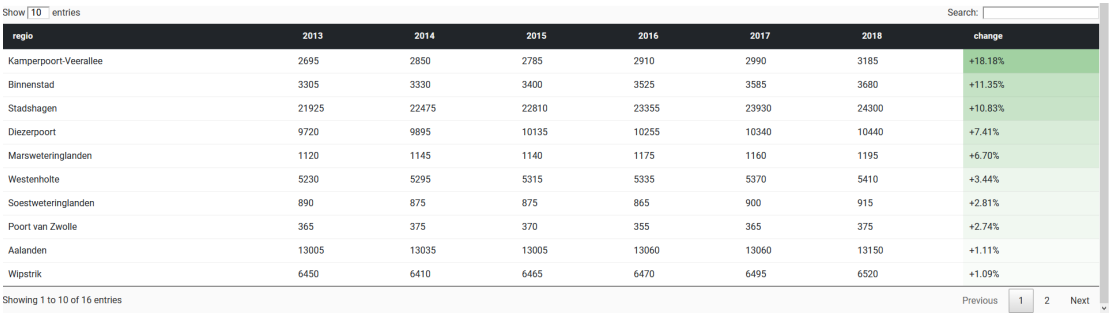
visualisation to load. Ideally, when the dashboard is running on a server, it request new data at midnight. The CBS key figure data set is not live data and as such it does not change frequently. Thus, it is fair to set the update limit to once a day. With the data in place, a proper line chart is created with D3 using the year as x-coordinate and the measure values as y-coordinate.

#### A solution for changing region borders

The CBS themselves are reluctant when it comes to comparing different points in time for their regional data sets. This is because the borders of regions are not static. Neighbourhoods expand or shrink and sometimes even new ones are created or existing ones merged with others. For this reason, it would wrong to compare the development between two years when the actual region has grown or shrunk. However, there is a solution to this problem. Every year, the CBS includes a measure on whether or not the region compared to the previous year has changed. As such it is possible to visualise a time series and indicate which regions have stayed the same and which have changed with a simple coloured dot can indicate. This dot is green if the region stayed the same and red if it changed. It will also interrupt the line so it is clear the region ended.

### 6.4.3 Table with relative change

The D3 line chart was extended with a table that contains the same data as the line chart as well as a column that calculates the relative change between the first and last year. The result of this table can be exported as CSV so the user can work with the data themselves. Exporting to CSV was a much requested feature from people at the Kadaster since many municipalities like to work with the data in their own ways. The relative change is also mapped to a D3 colour scale. No increase (0%) has a white coloured cell, whereas a decrease is coloured red and an increase is coloured green. Finally, the entire table is styled using the data tables plugin for jQuery. Now the region names can be filtered and each of the columns can be sorted. This completes the final element of the second page. In conclusion, the complete second page can be viewed in figure 6.10.


Show 10 entries Search:

| regio | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | change |
|---|---|---|---|---|---|---|---|
| Kamperpoort-Veerallee | 2695 | 2850 | 2785 | 2910 | 2990 | 3185 | +18.18% |
| Binnenstad | 3305 | 3330 | 3400 | 3525 | 3585 | 3680 | +11.35% |
| Stadshagen | 21925 | 22475 | 22810 | 23355 | 23930 | 24300 | +10.83% |
| Diezerpoort | 9720 | 9895 | 10135 | 10255 | 10340 | 10440 | +7.41% |
| Marsweteringlanden | 1120 | 1145 | 1140 | 1175 | 1160 | 1195 | +6.70% |
| Westenholte | 5230 | 5295 | 5315 | 5335 | 5370 | 5410 | +3.44% |
| Soestweteringlanden | 890 | 875 | 875 | 865 | 900 | 915 | +2.81% |
| Poort van Zwolle | 365 | 375 | 370 | 355 | 365 | 375 | +2.74% |
| Aalanden | 13005 | 13035 | 13005 | 13060 | 13060 | 13150 | +1.11% |
| Wipstrik | 6450 | 6410 | 6465 | 6470 | 6495 | 6520 | +1.09% |

Showing 1 to 10 of 16 entries Previous 1 2 Next

Figure 6.9: A table showing the progression of a measure over the years, including the relative change between the first and last year


Figure 6.10: Page 2: "Explore your municipality"

## 6.5 Page 3: Find relationships

The third page was added after the second feedback presentation at the Kadaster. Many of the people present where interested in seeing if any measures had a relationship between them, for example average income versus average gas usage. The idea of this third page is to provide a tool for mainly researchers to find relationships between two measures of the key figure data set. The only new component on this page is a scatter plot with a trend line.

### 6.5.1 Scatter plot with a trend line

The scatter plot is created similarly to the line chart from page 2. Instead, both axis now use two different measure values. This is where the `VALUES` command from the query in figure 6.2 will really help. The values of a second measure are simply retrieved by adding another definition to this. The side bar is updated to now include two measure drop down menus, one for each axis.

When looking for relationships, regression or trend lines are often used. An example on how to calculate the function of a trend line is built into the dashboard to automatically calculate a linear regression and add it to the graph. This function was later replaced by the D3 library, d3-regression. This library does not only generate a linear regression for any data set, it also includes other regressions such as: quadratic, logarithmic and local polynomial regression. An option is added which allows the user to select the type of regression from the sidebar. Not every type of regression will fit the current data set and then the line cannot be drawn. Most often, a linear regression is all that can be applied to the key figure data sets. In conclusion, the complete third page can be viewed in figure 6.11.

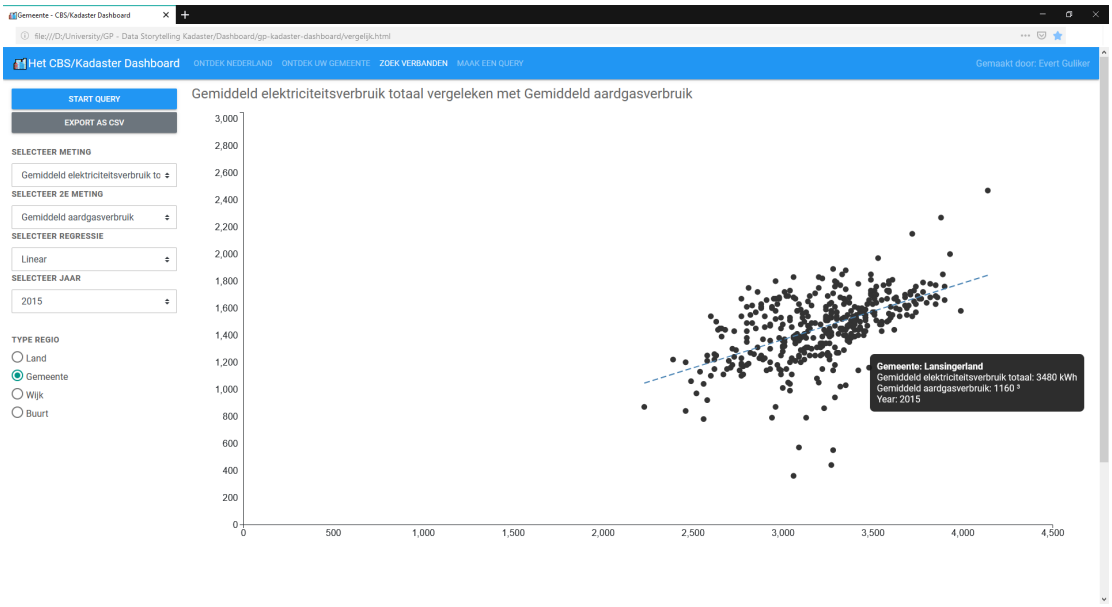



Figure 6.11: Page 3: "Discover relationships"

## 6.6 Page 4: Create a Query

This final page was added so that experienced users can query the key figure data set themselves. It also allows the user to try out the queries used on the first page and see what the results look like. The main and only component of this page is the YASGUI library. The user can enter queries in the YASQE query editor and visualise the results with YASR. By default it will generate a table of the selected variables, but it can also has some options for graphs. All in all, YASGUI is the only component on the page so the user can execute queries themselves.

## 6.7 Publishing the dashboard

Linked open data emphasises transparency. As such, the dashboard also has to be transparent on how the data is retrieved. After all, it is the Kadaster's own motto to retrieve data at the source. The dashboard is transparent by showing the user what query is being run below the visualisation. To further emphasise the transparency, the source code of the entire dashboard will be published as open source on the git.snt environment of the University of Twente under a Creative Commons Attribution NonCommercial ShareAlike 4.0 International license [34]. This allows anyone to have insight into the source code of the dashboard and contribute to the code if they feel like it.

## 6.8 Conclusion

The final CBS & Kadaster data dashboard consists of a total of four pages. The overall goal is to allow the user to gain insight in the measures that are relevant to them. The creation of the dashboard was driven by the motto of the Kadaster: "Data at the source", as such all data is retrieved dynamically using SPARQL queries or in case of the second page, through the API of the StatLine.

To summarise the final product, the first page: "explore the Netherlands" allows the user to explore the linked open data set on a large scale by comparing multiple regions at the same time. The user can freely select a region and a region type which automatically updates the linked data query below the resulting bar chart. Besides the values being plotted on a bar chart they are being displayed on a linked map as can be seen in figure 6.8.

The second page allows for a more in depth look at a specific municipality and its districts/neighbourhoods or all municipalities in the Netherlands. For a specific measure, it shows multiple regions in a line chart and how they developed over time. Below the graph is a table which also contains the relative change between the first and last year. This can be used to name and shame municipalities that have not been improving certain measures such as electricity consumption.

Finally, the third page is mainly focused on researchers and allows two measures to be compared in a scatter plot. The scatter plot automatically generates a linear regression line that tries to fits the data. Using this tool it can quickly be seen if there appears to be a relationship between two measures. The fourth page is a small additional page where queries can be directly queried on. It servers as a testing ground for experienced users to try some of the queries from the first page. This concludes the complete realisation of the CBS & Kadaster data dashboard.

# 7 | Evaluation

The final phase evaluates the dashboard created in the realisation phase. First, usability testing is used to check if actual users are able to find the data and create the correct visualisations with the dashboard. The usability tests end with an informal user interview followed by a short survey to find strong and weak points of the dashboard. After processing the feedback of the user tests, the requirements created in section 5.1 are evaluated. Finally, the chapter concludes with an ethical reflection on data storytelling and this dashboard.

## 7.1 User testing

### 7.1.1 Testing procedure

The user tests are conducted by giving the test participants specific tasks to complete inside the dashboard. They will need to find certain values of measures for specific regions using the filters as part of the side bar. The user test only tests the two largest pages: page 1 "Explore the Netherlands" and page 2 "Explore your municipality". Before the start of the user test, the participant will first be briefly introduced to the concept of linked data and the goal of the dashboard. This goal will be explained the same way for every participant as follows "The Dashboard allows the user to pick measures about the region that they are interested in". This is only a brief summary. The main goal of the introduction is to provide the user with some context to understand the purpose of the dashboard. Finally, the user is asked to provide written permission that there data is anonymously used in this thesis, if no permission is given the test cannot continue.

In the next five minutes following the explanation, the user is able to freely explore the first page for up to five minutes. During the user tests, the observer will note down two metrics: completion time and error rate. The completion time is how long the user takes before completing the task. The error rate is how many errors the user experiences when performing the tasks. These errors can be rated minor, moderate or severe. The error will also be noted down separately so it can later be resolved. Now both the observer and user are ready to start the user test. First the user is given the following three tasks for the first page:

1. Task 1: Show the average gas usages of all municipalities and sort the bar graph with ascending values.
2. Task 2: Find only the total population of the municipality that you live in and click on it on the map.
3. Task 3: Find the average income of the entire Netherlands.

After these tasks have been completed, the user gets up to five minutes to explore the second page. Then he is asked to retrieve the following pieces of information from the second page:

1. Task 4: Show the average income of all districts of Enschede.
2. Task 5: For the districts of task four, give the name of the district that has the largest relative increase in average income and specify how much this relative increase is.

At the end of the usability there is an informal interview with the user where some questions are asked on how the tasks went according to the user. In addition to this, the user is asked to fill in a short reflection. This form can be seen in appendix .3. The form consists of two open questions and 10 statements that have to be rated on a Likert scale from "strongly agree" to "strongly disagree". After the user hands in their form, the user test has finished. The user is then thanked and rewarded with some food and drinks.

### 7.1.2 Participants

The best scenario are the people working in the administration of municipalities as they are the primary target audience of the dashboard. Unfortunately, it was not a possibility to come into contact with anyone related to this occupation. Instead, the usability was evaluated with a group of students and adults, as citizens are the secondary target audience who will mainly be the ones to use it in the PDOK Labs environment.

The first group of participants consists of mostly bachelor students of the University of Twente. Their ages range between the 20-25 years old. The bachelor students follow the following programs: Creative Technology (2), Bio-medical Engineering (2), Applied Mathematics (1) and Applied Physics (1). In total six students partake in the usability test.

Similarly, a group of six adults was recruited via close contacts and parents of friends. Their ages range between 40-55 years old, with the average age lying around 46 years old. It will be interesting to see if there are any difference between the results of the two user groups.

### 7.1.3 Results and conclusion

The results of the usability testing can be seen in appendix .2. The table contains the time in seconds it took each user to find the required information. The green records are the students, whereas the blue records are the adults. Additionally, some basic measures such as standard deviation and average completion time for both groups is given.

The results of the survey can be seen in appendix .4. The results of the statements have been mapped to a point scale from one (strongly disagree) till five (strongly agree). The results of the open questions have been left out but are summarised in this section.

As expected, the usability tests show that the students are a bit faster than the adults on average. However, this is only by a small margin which suggests that the dashboard is about equally well understood. The decisive factor might be that the students have more experience digital interfaces or are more quickly to adapt to a new interface. On average all tasks where completed in under a minute. This can be

considered as an acceptable time for the dashboard. An interesting observation is that the tasks on the first page where completed faster than the tasks on the second page.

One critical error that was observed during the user tests was that the districts and neighbourhood buttons where not in the right order. The list would logically go down from the highest level (national) down to the lowest regional level (neighbourhood). However, the district region type was at the bottom not the neighbourhood. This confuses some users as it is counter intuitive for this list to not be sorted.

For the survey there were multiple statements that had a counter-statement to double check if the user was filling in the questionnaire truthfully. Overall, the design was intuitive and clear for most users, receiving a 4.25 on average. It also did not take the users long to find the data they wanted according to the survey. Some users remarked that finding the correct measure in the list took up the most time. Perhaps in the next iteration, the measure groups from the CBS can be used to group the measures together. Another drop down menu would then be used to first select the measure group and then the actual measure itself.

Most users were convinced that the tool would be useful to municipalities. One statement "I was able to learn more about Linked Open Data" was rated neutrally. This is not bad since the goal of the dashboard is not necessarily about learning more about Linked Open Data. It can still be a helpful addition for the future to provide more background information on linked data inside the dashboard for less experienced users.

In conclusion, the dashboard was well received by the participants, receiving a 4.25 on a scale of 5 for the intuitiveness of the design. Many users complimented the looks of the dashboard and found it a nice opportunity to learn something new about their own neighbourhood. The users did not learn much new about Linked Open Data through the dashboard. No critical error where encountered during the usability testing, they were mostly small mistakes and oversights. Small changes that have been made include restructuring of some filter elements. A recommendation for future design iterations is shrinking down the measure lists to make it easier to quickly find the right measure.

## 7.2 Requirements evaluation

This section evaluates the functional and non-functional requirements created in section 5.1 to see whether or not the functionality of the dashboard is a success.

### 7.2.1 Functional requirements

Almost all of the functional requirements are met with the final prototype: 6/6 "must have", 6/6 "should have" and 3/5 "could have" requirements are completed. The requirements that could not be completed are:
NF14: "The dashboard has an option to change the chart type op the visualisation".
NF17: "Regions can be filtered on the map using a lasso selection tool"

### 7.2.2 Non-functional requirements

For the non-functional requirements the total requirements met is as follows: 7/7 "must have" and 2/5 "could have". There were no should have requirements as they were all deemed important enough to be a must have requirement. The requirements that could not be completed are:
NFR8: "The dashboard is available in English"
NFR9: "The dashboard has a progress bar when the data is loaded" - *The only feedback given is that the chart title changes to "data loading"*
NFR12: "The dashboard has a colour blind mode"

All in all, the requirements that could not be met where not realised because of time constraints. Other functionalities took priority over them. One requirement that is under active development is the English translation of the dashboard. For now it should not be a problem since the main target audience is Dutch citizens. Since all must have and should have requirements both functional and non-functional have been met, the dashboard's functionality can be considered a success.

## 7.3 Ethical reflection on data stories

Visualisations have the power to influence others and as such there is a need to consider good ethics in our approach of creating data stories. Ethics referring to a set of principles (often moral principles) that guide the behaviour of a person. Ethical conduct is an important objective in practice. When dealing with data visualisations there are many opportunities to twist the truth and deceive the reader with untruthful visualisations. This section discusses the views of two authors and highlights when a visualisation can be considered an ethical success or a failure and how failure can be avoided.

On the responsibility of data scientists, Drew Skau, a PhD Computer Science visualisation student at NC Charlotte says "I shall not use visualisation to intentionally hide or confuse the truth which it is intended to portray. I will respect the great power visualisation has in garnering wisdom and misleading the uninformed. I accept this responsibility will fully and without reservation, and promise to defend this oath against all enemies, both domestic and foreign." (page 3)[35] This quote is from Skau his article about "A Code of Ethics for Data visualisation Professionals". It list a number of ethical guidelines for making data visualisations. The four principles he considers for a good ethical are: honest data collection, truthful data analysis (not manipulating results), a clear well understood design, resonating message for the target audience.

On the other hand, data visualisations can be seen as a combination of journalism and engineering according to Alberto Cairo. Engineering refers to designing a visualisation to efficiently get its point across to the target audience. He combines this with the "ethos" of only reporting the truth from journalism. Which is comparable to miller. Alberto states that the intuitive sense is useful. In summary, creating a straightforward visualisation which clearly shows what it wants to tell you while keeping in mind the helpful and harmful aspects of this message.

Considering both of these authors' perspectives on data visualisation, the visualisation

is considered a failure when it does not show the message it was intended for. A data visualisation should therefore not be left open to interpretation, unless the author just wants to display data. However, this thesis concerns itself more with data stories, in which case the author wants to tell something to the reader. A wrong interpretation of this message would be catastrophic, as readers can use this wrong information themselves to formulate opinions or spread the misinformation to others. Even after knowing the visualisation was wrong, they might still be left with a bias towards a certain topic due to their previous knowledge.

To summarise, the moral purpose of these ethical guidelines is to improve the reader's well-being through understanding. A graphic that is confusing or misleading is unethical, regardless of its intent. It creates misunderstanding for the audience. The success of the visualisation can be defined as the opposite of what was discussed in the paragraph above. It should be clear and have a single clear message that can only be interpreted a single way and not be biased towards certain individuals.

### 7.3.1 Privacy

Before even making the visualisation, the first aspect to consider is the privacy of the person or the entity the data is about. A lot of data can be found through the internet. Sometimes it is clear that the data is free to use and other times the data is gathered by the data scientist himself. On Facebook, Twitter and other social media networks there is a lot of publicly available data that is related to persons. By posting a tweet or post, the person publishes his information to all of the world to see. However, a user might now be aware of the fact that this information could also possibly be used against himself.

A clear case of privacy sensitive data is personal details. The Kadater themselves already considers this aspect before publishing the data publicly. An example of semi-private data which the Kadaster does store, are the "WOZ-waardes" which is a valuation of real estate. This data can be requested via their website but it is limited to five addresses per day per web session. Automatic retrieval of this sensitive data is punishable by law, as well as collecting the data as means to build your own data set of WOZ-waardes. When using this personal information it might be good or bad, depending on who you ask, to exposes millionaires who own a lot of expensive properties in a city. In the end privacy, remains something to keep in mind when using personal data.

### 7.3.2 Validity & Deceptive data

Like any piece of information, the validity of a data visualisation is often unsure unless the source is checked. Many people take something at face value and if they are shown a graph with a source, chances are they will not check the original publication. This allows for people to manipulate data and publish false information. This is a big problem of the twenty-first century since everybody has a readily available method to spread (false) information: the internet. As such, checking sources has become essential, but despite this people often too busy to check. It is an ethical responsibility of a researcher to ensure that the information is correct.

But what if you manipulate the data to promote good behaviour? Then it can be seen as, the end justifies the means. However, in and of its own, the act is still wrong because the truth is being manipulated. Taking only a specific sub set of the data is also a way of manipulating the overall message of the visualisation.

Deceptive data is a different kind of manipulation of data in which the values of the data visualisation might be correct, but it is displayed in such a manner that purposely tries to trick somebody to think something else. This can still be seen as wrong and should be avoided at all costs, even when it is done on accident. All in all, the researcher should not have to lie to get his point across and should instead use different methods or data sets to convince the audience.

### 7.3.3 Causation vs. Correlation

Causation and correlation are two different things. A graph which shows that when ice cream sales increase, the amount of homicides increases too, does not mean that buying ice cream turns you into a killer. However, these erroneous relationships can be used to wrongly convince people that there are factors which influence one another. It can also be a matter of coincidence and this example is often labelled as confirmation bias when we use to pre-existing views to extrapolate a wrong new view.

### 7.3.4 Data leak

The last ethical risk is a data leak. This is usually only a problem when dealing with privacy sensitive information in interactive visualisations which request data from a server. Hackers might target this visualisation in hopes of uncovering the data themselves. Especially with visualisations with private information it is crucial that the personal information is protected.

Summary of possible solutions:

1. **Privacy:** written consent from parties involved or notify parties about their data being used as far as possible.
2. **Validity:** reference to sources or even better, showing underlying queries used in retrieving the data from a data base.
3. **Deceptive data:** intuitive design taking into account wired truths (such as green is positive, red is negative)
4. **Causation vs correlation:** only use correlations if they are related, a direct cause of one another and it's absolutely certain they influence each other to a certain extent.
5. **Data leak:** obfuscate private information or try to avoid using sensitive data in the first place

## 7.4 Ethical reflection regarding the dashboard

The most prominent ethical problem in the dashboard, is the risk of people drawing the wrong conclusions from the data. As was discussed in the previous section on causation versus correlation, a correlation does not necessarily imply causation. In addition, if a trend line can be fitted to the data set, this trend line is often an average and does not imply a relationship between the two variables.

This all depends on a number of statistical measures such as variance, standard deviation or the coefficient of determination (R-squared). Users can draw wrong conclusions from the dashboard, especially the fourth page where two variables can be compared and are fitted with a trend line. Users can quickly spread these false conclusions through the internet. The spread of this misinformation needs to be prevented and the user must be reminded to not take the trend line at face value.

The final dashboard is updated with a warning that not every relationship implies a connection between the two variables and that statistical tools should be used to analyse if the relationship is significant. For now, the R-Squared value will also be provided for each trend line. In the future more explanation on what is and what is not a relationship can prevent users from drawing the wrong conclusions. Since the dashboard will be published on the PDOK Labs environment it is most likely that the users will have a statistical background, as such the small solution will be enough for now until an new iteration of the dashboard is published.

## 7.5 Conclusion

The evaluation phase shows that the final dashboard meets all of the most crucial requirements formulated during the specification phase. Some of the "could have" requirements were not met in the end because they required too much effort to be realised in the time span of this thesis. During usability testing some small improvements were found for the user interface, however the functionality of the dashboard was all working correctly. Many users found the design intuitive to use and were able to find the required data in under a minute. This can possibly be further brought down by grouping the measures list as users were spending most of their time scrolling through the list looking for the correct value.

The survey showed positive results on both the design and the amount of useful information they could retrieve from the dashboard. Users where convinced that it would be a useful tool for the intended audience, the municipalities of the Netherlands. Despite the users clearly seeing the linked data aspect in the dashboard, they did not learn much about linked data from the dashboard. This is fine, as the main goal is to give the user insight in the data and not teach about linked open data.

The ethical reflection outlines five common ethical pitfalls that can be identified in data visualisation projects: privacy, validity, deceptive data, causation vs correlation and data leaks. The causation vs correlation problem turned out to be one of the ethical problems of the data dashboard. The fourth page that was added last, automatically generates a trend line when two measures are compared. This led some of the users to believe that there might be a correlation between the measures, even though in reality

there is not and its either pure coincidence or the trend line has a high variance and thus a weak correlation. This problem was solved by including a warning on the page, stating that it should be assumed that there is no statistical causation between the two measures. It is only a tool to find possibilities for relationships. For the researchers, the R-squared value was also added to make it explicit for researchers if there was a strong or weak relationship.

# 8 | Conclusion

The research question defined at the start of this thesis is:
***"How to implement a data story which shows the value of the linked open data sets of the Kadaster?"***

The answer to this question is guided by the Creative Technology Design Approach. First in the ideation phase, a suitable linked open data set of the Kadaster is picked. For this data set potential data stories are created which resulted in the idea to create a dashboard that provides insight in the key figures: districts and neighbourhoods data set published by the CBS in collaboration with the Kadaster. This data set contains many important measures about different regional levels of the Netherlands.

The end result of the ideation phase is "The CBS & Kadaster Data Dashboard", a web-based dashboard which allows the user to gain insight into many measures (income, energy usage, demographics etc.) about the municipalities and neighbourhoods in the Netherlands. The dashboard essentially allows the user to generate a data story that is relevant to them.

The main target audiences of the dashboard are the governments of municipalities that want to gain insight in the developments of their neighbourhoods or districts. The tool is also helpful to identify improving or worsening municipalities based on certain measures the user is interested in. In addition to governmental organisations, it is also a tool for researchers to look for relationships between measures or developments across larger regions of the Netherlands. Furthermore, the tool will provide the ability for citizens to gain insight into their neighbourhoods and their developments overtime.

The realised dashboard was evaluated using usability testing and a user survey. The results were positive and showed that users liked the intuitive design and were able to quickly retrieve the information in under a minute. The users also commented on the large amount of information they could retrieve from the dashboard. Users were convinced that it would be a useful tool for the intended audience, the municipalities of the Netherlands.

The secondary goal of this thesis is to contribute to the knowledge field of linked data by providing an ethical review on data storytelling and a literature review on guidelines for the creation of good data visualisations and linked open data stories. The ethical review outlines five common ethical risks when designing data visualisations and data stories: privacy, validity, deceptive data, causation vs correlation and data leaks.

This ethical reflection is also applied to the dashboard itself. The causation vs correlation problem turned out to be one of the ethical problems of the data dashboard. The problem existed on the fourth page were a trend line was automatically generated when two measures are compared. This led users to believe that there is a correlation between the measures, even though in reality there is none. This problem was solved by including a warning on the page, stating that it should be assumed that there is no statistical causation between the two measures. It is only a tool to find possibilities for relationships. For the researchers, the R-squared value is included to identify strong and weak relationships.

Additionally, the literature review conducted during the initial phase of the project outlines that the author must pick the right visual encodings suited for their visualisation and be aware of the up and down sides of the graph types he chooses for the data visualisations. To tell a successful story, the author needs to define a target audience and pick a problem that is relevant to them. The story should start with a brief introduction to the problem and build upon to a final conclusion. When dealing with linked data stories, it is important to not forget to show the aspect of linked open data. This can be done by adding the queries that are used to generate the visualisations and allow to find the URI of the data elements by clicking on the labels of the visualisation.

In conclusion, the implementation of a data story to show the value of the linked open data sets of the Kadaster was a success. The final dashboard is currently available on the author's personal website at `www.evertguliker.nl`. The source code is also published online at: `https://git.snt.utwente.nl/s1789449/gp-kadaster-dashboard`. Lastly, at the moment there are plans to also include the dashboard in the PDOK Labs environment at `https://data.labs.pdok.nl/stories/`. The dashboard will be publicly available here for anyone to try it out.

## 8.1 Recommendations for future research

For future research I discussed with Wouter Beek, the creator of YASGUI, the possibilities of visualisation components for linked data. During the development of the dashboard, employees of the Kadaster asked if they themselves could be able to create custom visualisations using the linked open data. This idea, of facilitating so called components is an interesting avenue. This will not only make it easier for less experienced users to create data story dashboards using linked data themselves, it will also increase the variety of different dashboards that can be made.

The visualisations in this projects are made specifically for the queries so that less experienced users can experiment with applying specific filters. It should be possible to rewrite these components such that also the underlying visualisation type can be changed. This would require a common data format for all visualisations. All in all, for future research, it is recommended to explore the possibilities of visualisation components that can be used to build dashboards backed by linked open data.

## 8.2 Acknowledgements

I would like to start off by thanking both my supervisors; Maurice van Keulen and Erwin Folmer for their guidance during the development of this thesis. Erwin provided me with opportunities to present my progress at the Kadaster and receive crucial feedback from other employees. Additionally, I would like to thank Wouter Beek, for his enthusiastic response on my dashboard and allowing me to visit Tripply and together discuss some of the projects they are working on with linked open data. Finally, a big thank you to all the volunteers who participated in user testing and provided feedback on the dashboard. The final CBS & Kadaster data dashboard would not be complete without them.

# Bibliography


[1] N. Shadbolt, K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser, W. Hall, *et al.*, "Linked open government data: Lessons from data. gov. uk," *IEEE Intelligent Systems*, vol. 27, no. 3, pp. 16–24, 2012.

[2] E. Folmer, W. Beek, L. Rietveld, S. Ronzhin, R. Geerling, and D. den Haan, "Enhancing the usefulness of open governmental data with linked data viewing techniques," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[3] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: The story so far," in *Semantic services, interoperability and web applications: emerging concepts*, IGI Global, 2011, pp. 205–227.

[4] C. Bizer, "The emerging web of linked data," *IEEE intelligent systems*, vol. 24, no. 5, pp. 87–92, 2009.

[5] T. Berners-Lee, J. Hendler, O. Lassila, *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[6] C. C. Marshall and F. M. Shipman, "Which semantic web?" In *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, ACM, 2003, pp. 57–66.

[7] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis lectures on the semantic web: theory and technology*, vol. 1, no. 1, pp. 1–136, 2011.

[8] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo, "Vistrails: Visualization meets data management," in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '06, Chicago, IL, USA: ACM, 2006, pp. 745–747, ISBN: 1-59593-434-0. DOI: 10.1145/1142473.1142574.

[9] A. Graves and J. Hendler, "Visualization tools for open government data," in *Proceedings of the 14th Annual International Conference on Digital Government Research*, ser. dg.o '13, Quebec, Canada: ACM, 2013, pp. 136–145, ISBN: 978-1-4503-2057-3. DOI: 10.1145/2479724.2479746.

[10] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984. DOI: 10.1080/01621459.1984.10478080.

[11] B. Saket, A. Srinivasan, E. D. Ragan, and A. Endert, "Evaluating interactive graphical encodings for data visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 3, pp. 1316–1330, Mar. 2018, ISSN: 1077-2626. DOI: 10.1109/TVCG.2017.2680452.

[12] N. Iliinsky and J. Steele, *Designing data visualizations: Representing informational Relationships*. " O'Reilly Media, Inc.", May 2011. DOI: 10.1327/838-2-234-23126-3_2.

[13] I. Spence and S. Lewandowsky, "Displaying proportions and percentages," *Applied Cognitive Psychology*, vol. 5, no. 1, pp. 61–77, 1991. DOI: 10.1002/acp.2350050106.

[14] Y. Engelhardt and C. Richards, "A framework for analyzing and designing diagrams and graphics," English, in *Diagrams 2018: Diagrammatic Representation and Inference*, P. Chapman, A. Moktefi, S. Perez-Kriz, and F. Bellucci, Eds., ser. Lecture Notes in Computer Science. Germany: Springer, May 2018, pp. 201–209, ISBN: 978-3-319-91375-9. DOI: 10.1007/978-3-319-91376-6_20.

[15] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister, "What makes a visualization memorable?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2306–2315, Dec. 2013, ISSN: 1077-2626. DOI: 10.1109/TVCG.2013.234.

[16] W. S. Cleveland, *Visualizing Data*. Hobart Press, 1993, ISBN: 0963488406.

[17] S. Cockcroft, "A taxonomy of spatial data integrity constraints," *GeoInformatica*, vol. 1, no. 4, pp. 327–343, Dec. 1997, ISSN: 1573-7624. DOI: 10.1023/A:1009754327059.

[18] R. Kosara and J. Mackinlay, "Storytelling: The next step for visualization," *Computer*, vol. 46, no. 5, pp. 44–50, May 2013, ISSN: 0018-9162. DOI: 10.1109/MC.2013.36.

[19] N. L. Stein, "The definition of a story," *Journal of Pragmatics*, vol. 6, no. 5, pp. 487–507, 1982, ISSN: 0378-2166. DOI: 10.1016/0378-2166(82)90022-4.

[20] B. Lee, N. H. Riche, P. Isenberg, and S. Carpendale, "More than telling a story: Transforming data into visually shared stories," *IEEE computer graphics and applications*, vol. 35, no. 5, pp. 84–90, 2015. DOI: 10.1109/TVCG.2013.126.

[21] C. North, "Toward measuring visualization insight," *IEEE Computer Graphics and Applications*, vol. 26, no. 3, pp. 6–9, May 2006, ISSN: 0272-1716. DOI: 10.1109/MCG.2006.70.

[22] C. Plaisant, "The challenge of information visualization evaluation," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ser. AVI '04, Gallipoli, Italy: ACM, 2004, pp. 109–116, ISBN: 1-58113-867-9. DOI: 10.1145/989863.989880.

[23] A. Sonderegger and J. Sauer, "The influence of design aesthetics in usability testing: Effects on user performance and perceived usability," *Applied Ergonomics*, vol. 41, no. 3, pp. 403–410, 2010, Special Section: Recycling centres and waste handling - a workplace for employees and users., ISSN: 0003-6870. DOI: 10.1016/j.apergo.2009.09.002.

[24] M. A. Linton, J. M. Vlissides, and P. R. Calder, "Composing user interfaces with interviews," *Computer*, vol. 22, no. 2, pp. 8–22, Feb. 1989, ISSN: 0018-9162. DOI: 10.1109/2.19829.

[25] A. Mader and W. Eggink, "A design process for creative technology," E. Bohemia, A. Eger, W. Eggink, A. Kovacevic, B. Parkinson, and W. Wits, Eds., pp. 568–573, Sep. 2014.

[26] M. Weise, S. Lohmann, and F. Haag, "LD-VOWL: extracting and visualizing schema information for linked data endpoints," in *Proceedings of the 2nd International Workshop on Visualization and Interaction for Ontologies and Linked Data (VOILA 2016)*, ser. CEUR-WS, vol. 1704, CEUR-WS.org, 2016, pp. 120–127. [Online]. Available: http://ceur-ws.org/Vol-1704/paper11.pdf.

[27] M. Bostock. (2011). D3.js - data-driven documents v5.9.7, [Online]. Available: https://d3js.org (visited on 03/29/2019).

[28] V. Agafonkin. (2015). Leaflet.js v1.4.0, [Online]. Available: https://leafletjs.com/ (visited on 03/29/2019).

[29] J. Resig. (2006). Jquery and jquery ui javascript library v3.4.1, [Online]. Available: https://jquery.com/ (visited on 03/29/2019).

[30] FezVrasta. (2015). Material design for bootstrap v4.1.1, [Online]. Available: https://fezvrasta.github.io/bootstrap-material-design/ (visited on 04/12/2019).

[31] SpryMedia. (2009). Datatables plugin for jquery v1.10.19, [Online]. Available: https://datatables.net/ (visited on 06/15/2019).

[32] B. Sudol. (2014). Responsive d3 charts, [Online]. Available: https://brendansudol.com/writing/responsive-d3 (visited on 03/29/2019).

[33] A. Endsley. (2012). Wicket - wkt data in leaflet, [Online]. Available: https://arthur-e.github.io/Wicket/ (visited on 04/19/2019).

[34] (2010). Creative commons 4.0 international licence, [Online]. Available: http://creativecommons.org/licenses/by-nc-sa/4.0/ (visited on 07/01/2019).

[35] S. C. Lewis and O. Westlund, "Big data and journalism: Epistemology, expertise, economics, and ethics," *Digital journalism*, vol. 3, no. 3, pp. 447–466, 2015.

# Appendices

# .1 CBS Wijken en Buurten - Measures list (Dutch)

1

| *Code - naam van variable* |
|---|
| **Algemeen** |
| regio: Regioaanduiding |
| gm_naam: Gemeentenaam |
| recs: Soort regio |
| gwb_code: Codering |
| ind_wbi: Indelingswijziging wijken en buurten |
| **Bevolking** |
| a_inw: Aantal inwoners |
| a_man: Mannen |
| a_vrouw: Vrouwen |
| a_00_14: 0 tot 15 jaar |
| a_15_24: 15 tot 25 jaar |
| a_25_44: 25 tot 45 jaar |
| a_45_64: 45 tot 65 jaar |
| a_65_oo: 65 jaar of ouder |
| a_ongeh: Ongehuwd |
| a_gehuwd: Gehuwd |
| a_gesch: Gescheiden |
| a_verwed: Verweduwd |
| a_w_all: Westers totaal |
| a_nw_all: Niet-westers totaal |
| a_marok: Marokko |
| a_antaru: Nederlandse Antillen en Aruba |
| a_suri: Suriname |
| a_tur: Turkije |
| a_ov_nw: Overig niet-westers |
| a_geb: Geboorte totaal |
| p_geb: Geboorte relatief |
| a_ste: Sterfte totaal |
| p_ste: Sterfte relatief |
| a_hh: Huishoudens totaal |
| a_1p_hh: Eenpersoonshuishoudens |
| a_hh_z_k: Huishoudens zonder kinderen |
| a_hh_m_k: Huishoudens met kinderen |
| g_hhgro: Gemiddelde huishoudensgrootte |
| bev_dich: Bevolkingsdichtheid |

[1]Source: CBS Wijken en Buurten, Toelichting Variablen, Retrieved from: `https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/wijk-en-buurtstatistieken/kerncijfers-wijken-en-buurten-2004-2018` on July 15, 2019

| Wonen |
| --- |
| a_woning: Woningvoorraad |
| g_woz: Gemiddelde woningwaarde |
| p_1gezw: Percentage eengezinswoning |
| p_mgezw: Percentage meergezinswoning |
| p_bewndw: Percentage bewoond |
| p_leegsw: Percentage onbewoond |
| p_koopw: Koopwoningen |
| p_huurw: Huurwoningen totaal |
| p_wcorpw: In bezit woningcorporatie |
| p_ov_hw: In bezit overige verhuurders |
| p_e_o_w: Eigendom onbekend |
| p_bjj2k: Bouwjaar voor 2000 |
| p_bjo2k: Bouwjaar vanaf 2000 |
| **Energie** |
| g_ele: Gemiddeld elektriciteitsverbruik totaal |
| g_ele_ap: Appartement |
| g_ele_tw: Tussenwoning |
| g_ele_hw: Hoekwoning |
| g_ele_2w: Twee-onder-één-kap-woning |
| g_ele_vw: Vrijstaande woning |
| g_ele_hu: Huurwoning |
| g_ele_ko: Eigen woning |
| g_gas: Gemiddeld aardgasverbruik totaal |
| g_gas_ap: Appartement |
| g_gas_tw: Tussenwoning |
| g_gas_hw: Hoekwoning |
| g_gas_2w: Twee-onder-één-kap-woning |
| g_gas_vw: Vrijstaande woning |
| g_gas_hu: Huurwoning |
| g_gas_ko: Eigen woning |
| p_stadsv: Percentage woningen met stadsverwarming |
| **Inkomen** |
| a_inkont: Aantal inkomensontvangers |
| g_ink_po: Gemiddeld inkomen per inkomensontvanger |
| g_ink_pi: Gemiddeld inkomen per inwoner |
| p_ink_li: 40% personen met laagste inkomen |
| p_ink_hi: 20% personen met hoogste inkomen |
| p_n_act: Actieven 15-75 jaar |
| p_hh_li: 40% huishoudens met laagste inkomen |
| p_hh_hi: 20% huishoudens met hoogste inkomen |
| p_hh_lkk: Huishoudens met een laag inkomen |
| p_hh_osm: Huish. onder of rond sociaal minimum |

| Sociale zekerheid |
| --- |
| a_soz_wb: Personen per soort uitkering; Bijstand |
| a_soz_ao: Personen per soort uitkering; AO |
| a_soz_ww: Personen per soort uitkering; WW |
| a_soz_ow: Personen per soort uitkering; AOW |
| |
| **Criminaliteit** |
| g_wodief: Totaal diefstal uit woning/schuur/e.d. |
| g_vernoo: Vernieling, misdrijf tegen openbare orde |
| g_gewsek: Gewelds- en seksuele misdrijven |
| **Bedrijfsvestigingen per sector** |
| a_bedv: Bedrijfsvestigingen totaal |
| a_bed_a: A Landbouw, bosbouw en visserij |
| a_bed_bf: B-F Nijverheid en energie |
| a_bed_gi: G+I Handel en horeca |
| a_bed_hj: H+J Vervoer, informatie en communicatie |
| a_bed_kl: K-L Financiële diensten, onroerend goed |
| a_bed_mn: M-N Zakelijke dienstverlening |
| a_bed_ru: R-U Cultuur, recreatie, overige diensten |
| **Motorvoertuigen** |
| a_pau: Personenauto's totaal |
| a_bst_b: Personenauto's; brandstof benzine |
| a_bst_nb: Personenauto's; overige brandstof |
| g_pau_hh: Personenauto's per huishouden |
| g_pau_km: Personenauto's naar oppervlakte |
| a_m2w: Motorfietsen |
| **Voorzieningen** |
| g_afs_hp: Afstand tot huisartsenpraktijk; afstand |
| g_afs_gs: Afstand tot grote supermarkt |
| g_afs_kv: Afstand tot kinderdagverblijf |
| g_afs_sc: Afstand tot school |
| g_3km_sc: Scholen binnen 3 km |
| **Oppervlakte** |
| a_opp_ha: Oppervlakte totaal |
| a_lan_ha: Oppervlakte land |
| a_wat_ha: Oppervlakte water |
| **Postcode** |
| pst_mvp: Meest voorkomende postcode |
| pst_dekp: Dekkingspercentage |
| **Stedelijkheid** |
| ste_mvs: Mate van stedelijkheid |
| ste_oad: Omgevingsadressendichtheid |

## .2 Usability testing results

| task1 | task2 | task3 | task4 | task5 |
|---|---|---|---|---|
| 19 | 34 | 28 | 49 | 58 |
| 23 | 33 | 29 | 50 | 45 |
| 32 | 38 | 43 | 59 | 53 |
| 45 | 54 | 34 | 48 | 45 |
| 18 | 28 | 35 | 59 | 60 |
| 27 | 42 | 34 | 45 | 50 |
| 34 | 43 | 46 | 61 | 50 |
| 53 | 49 | 54 | 64 | 58 |
| 22 | 32 | 45 | 53 | 49 |
| 34 | 50 | 55 | 49 | 47 |
| 23 | 53 | 34 | 48 | 54 |
| 39 | 48 | 31 | 70 | 55 |
| **AVERAGE** | | | | |
| 30,8 | 42,0 | 39,0 | 54,6 | 52,0 |
| **AVERAGE STUDENTS ONLY** | | | | |
| 27,3 | 38,2 | 33,8 | 51,7 | 51,8 |
| **AVERAGE ADULTS ONLY** | | | | |
| 34,8 | 42,3 | 41,0 | 56,0 | 52,7 |
| **AVERAGE DIFFERNCE (ADULTS MINUS STUDENTS)** | | | | |
| 7,5 | 4,2 | 7,2 | 4,3 | 0,8 |
| **STD DEV** | | | | |
| 10,4 | 8,5 | 8,9 | 7,5 | 4,9 |

# .3 Survey form

**The CBS & Kadaster Dashboard**

By filling in this survey, you agree to the following: The results of this survey will be anonymously included in a summary in the final report. If you do not want your results to be included, please indicate so or do not fill in this survey.

The survey is for the bachelor thesis of Evert Guliker for Creative Technology at the University of Twente. The topic is on data story telling using linked open data. A final dashboard has been created which can be used to explore one of the linked open data sets: the key figures: districts and neighborhoods of the CBS.

Thank you very much for taking your time to fill out this survey.

**General questions:**

Your age: ……

Your gender: male / female / prefer not to answer / other: ……

Are you a Creative Technology student: yes / no

Have you received an explanation before hand about the dashboard: yes / no

**Please cross a single box with how much you agree with the following statements:**

| **Statement:** | **Strongly Disagree** | **Disagree** | **Neutral** | **Agree** | **Strongly Agree** |
|---|---|---|---|---|---|
| The dashboard was intuitive to use. | | | | | |
| I was able to quickly find the information I was asked for. | | | | | |
| The dashboard was too cluttered and provided to many options. | | | | | |
| I was able to learn more about Linked Open Data through the dashboard | | | | | |
| It took me a long time to understand the graphs. | | | | | |
| I was unable to learn any new information from the data set | | | | | |
| I do not think this tool is useful for any municipalities. | | | | | |
| I now know a bit more about my own neighborhood. | | | | | |
| I understood where the data was coming from / the source of the data | | | | | |
| I would use this tool myself sometime in the near future. | | | | | |

What where some negative points of the dashboard? (i.e. things that were missing or out of place)

What where your favorite features of this dashboard?

# .4 Survey results

| The dashboard was intuitive to use. | I was able to quickly find the information I was asked for. | The dashboard was too cluttered and provided to many options. | I was able to learn more about Linked Open Data through the dashboard | It took me a long time to understand the graphs. |
|---|---|---|---|---|
| 4 | 5 | 2 | 3 | 1 |
| 5 | 4 | 1 | 1 | 1 |
| 4 | 5 | 2 | 4 | 2 |
| 4 | 4 | 2 | 4 | 2 |
| 4 | 5 | 1 | 2 | 1 |
| 5 | 5 | 2 | 4 | 1 |
| 4 | 4 | 3 | 3 | 2 |
| 4 | 5 | 3 | 2 | 1 |
| 5 | 5 | 3 | 4 | 2 |
| 4 | 5 | 2 | 3 | 1 |
| 3 | 3 | 1 | 2 | 1 |
| 5 | 5 | 2 | 3 | 1 |
| **AVERAGE** | | | | |
| 4,25 | 4,58 | 2,00 | 2,92 | 1,33 |
| **STD DEV** | | | | |
| 0,60 | 0,64 | 0,71 | 0,95 | 0,47 |

| I was unable to learn any new information from the data set | I do not think this tool is useful for any municipalities. | I now know a bit more about my own neighborhood. | I understood where the data was coming from / the source of the data | I would use this tool myself sometime in the near future. |
|---|---|---|---|---|
| 1 | 1 | 4 | 4 | 4 |
| 1 | 2 | 5 | 3 | 4 |
| 2 | 1 | 4 | 4 | 3 |
| 3 | 2 | 5 | 5 | 4 |
| 1 | 1 | 4 | 5 | 3 |
| 2 | 2 | 4 | 4 | 3 |
| 2 | 3 | 4 | 4 | 3 |
| 3 | 2 | 4 | 5 | 4 |
| 1 | 2 | 4 | 3 | 5 |
| 1 | 1 | 4 | 5 | 4 |
| 2 | 1 | 5 | 5 | 2 |
| 2 | 2 | 4 | 4 | 3 |
| **AVERAGE** | | | | |
| 1,75 | 1,67 | 4,25 | 4,25 | 3,50 |
| **STD DEV** | | | | |
| 0,72 | 0,62 | 0,43 | 0,72 | 0,76 |