GIMA
Geographical Information Management and Applications

Reusability of volunteered geographic information supported by Semantic Web technologies: a case study for environmental applications

by Swarish Vinaash Marapengopi

Master of Science thesis

June 7th 2017

Professor: **Prof. Dr. M.J. (Menno-Jan) Kraak**
Department of Geoinformation Processing
Faculty of Geoinformation Science and Earth Observation
University of Twente

Supervisor: **Dr. Ir. R. L. G. (Rob) Lemmens**
Department of Geoinformation Processing
Faculty of Geoinformation Science and Earth Observation
University of Twente

Utrecht University student number: 4189779

University of Twente student number: S6010016

Universiteit Utrecht
TU Delft
UNIVERSITY OF TWENTE.
WAGENINGEN UR

Acknowledgements

This research would not have been possible without the inspiration and guidance of dr. ir. R. L. G. (Rob) Lemmens. I have to thank him for his help through this process and not giving up on me. I want to thank my parents for always supporting my academic endeavours, without them this would not have been possible. Lastly, I want to acknowledge my fellow (GIMA) students with whom I have had the pleasure of working with. I found it a profound pleasure, exchanging knowledge and collaborating on our various projects.

Abstract

New infrastructures, technologies and standards contribute to an internet that is more complex, dynamic and diverse than ever. It has never been easier to contribute to the growing networks of websites and (social media) platforms. All over the internet there is geographical information; sometimes explicitly, often implicit. To signify this, the term volunteered geographic information (VGI) was popularised in the academic community by Michael Goodchild a decade ago.

The amount of VGI keeps growing, and therefore it is timely to start thinking about how we can maintain the reusability of this data for the future. There are already several techniques in place on the internet that allow the reuse of data (i.e. web APIs, download services, and web scraping). Besides these current technologies, there are so-called Semantic Web technologies that can aid the reusability of VGI. Semantic Web technologies strive to create a web of data rather than a web of documents. It consists of a data standard (RDF), data structures (OWL) and a query language (SPARQL) that enables the development of this web of data.

The goal of this thesis is ***to develop a method in which Semantic Web technologies are used to improve the reusability of VGI.*** This entails the gathering of data from multiple (VGI) sources and creating proofs of concepts on the basis of use cases. These use cases are exemplary cases of how VGI could be reused by means of Semantic Web technologies. The use cases consist of five VGI systems and one authoritative data source in the environmental domain. The metadata and the data from these systems is extracted and reusability is attempted on both levels.

A domain ontology was developed to aid the reusability of VGI. Where possible, existing ontologies are applied, however many features and attributes were not readily available in existing ontologies. This VGI ontology is published online.

Chapter 4 delineates a general method for the reuse of VGI by employing Semantic Web technologies. It consists of four sequential steps, namely:

1. Gather metadata,
2. Gather data,
3. Model the (meta)data in RDF and
4. Upload and query the (meta)data.

This method is applied in Chapter 5 on the selected data sources as a proof of concept. Within the environmental domain, three use cases are developed: trash, weather and air quality.

In conclusion, this thesis has found that combining metadata from multiple sources yielded the most positive results. The Semantic Web technologies provide a structure for previously unstructured metadata which can be used for exploratory queries to discover the intricacies of a system. On the data level, reuse is more difficult because of the data quality of VGI and semantic gaps between the data collection and -processing methods. Semantic Web technologies provide additional structure and information about data however not every detail and interpretation is modellable.

Table of Contents

Acknowledgements ii
Abstract iii
List of figures vi
List of listings vii
List of abbreviations viii
1 Introduction 1
1.1 Background information and research motivation 1
1.2 Semantic Web technologies 2
1.3 Research objectives 4
1.3.1 General research objective 4
1.3.2 Research objectives and corresponding questions 5
1.4 Relevance of research 6
1.4.1 Scientific relevance 6
1.4.2 Societal relevance 7
1.5 Thesis outline 7
2 Related work: Reusability of volunteered geographic information 8
2.1 A definition of VGI 8
2.2 Elements of reusability for VGI 8
2.2.1 Heterogeneity, interoperability and quality of VGI 8
2.2.2 Findability and accessibility of VGI 10
2.2.3 Metadata, provenance and licensing of VGI 11
2.3 Current methods for reusing VGI 12
2.3.1 Application programming interfaces 12
2.3.2 Download services 12
2.3.3 Web scraping 12
2.4 Frameworks and standards 13
2.4.1 The Resource Description Framework 13
2.4.2 SPARQL Protocol and RDF Query Language 14
2.4.3 Linked Open Data Initiatives 14
2.5 Methods and tools to improve syntactic and semantic interoperability 15
3 Methodology: A case study of environmental applications 17
3.1 Research approach 17
3.2 Applications in the use case domain 17
3.3 Use case scenarios 18
3.4 VGI Ontology development 20
3.5 Software and tools 22
4 A method for the improvement of VGI reuse with Semantic Web technologies 23
4.1 Gather metadata 23
4.1.1 Explore the website 23
4.1.2 Identify the licensing situation 24
4.1.3 Identify data access mechanisms 24
4.1.4 Identify the purpose- and methods of data collection 24

4.1.5 Identify spatial and temporal coverages 24
4.2 Gather data 24
4.2.1 Query web APIs, use download services or apply web scraping 25
4.2.2 Determine the relative quality of the data 25
4.3 Model the (meta)data in RDF 25
4.3.1 Apply DCAT, DQV and VGI ontologies to metadata 25
4.3.2 Apply PROV, GEO, XSD and VGI ontologies to data 26
4.4 Upload and query the (meta)data 26
4.4.1 Upload the (meta)data into a triplestore 26
4.4.2 Visualizing and exploring the results 27
4.4.3 Query the (meta)data using SPARQL 28
5 Proof of concept 29
5.1 Use case 1: TrashHunters & MORA 29
5.1.1 Metadata level 29
5.1.2 Data level 31
5.2 Use case 2: hetweeractueel.nl & KNMI 32
5.2.1 Metadata level 32
5.2.2 Data level 34
5.3 Use case 3: AiREAS & enviroCar 35
5.3.1 Metadata level 35
5.3.2 Data level 37
6 Discussion, conclusions and recommendations 38
6.1 Discussion of proposed method 38
6.2 Discussion of the proof of concept results 39
6.2.1 General discussion on proof of concept 39
6.2.2 Use case 1: Trash 39
6.2.3 Use case 2: Weather 40
6.2.4 Use case 3: Air quality 40
6.3 Conclusions 41
6.3.1 Main conclusions 41
6.3.2 Answering research questions 42
6.4 Recommendations 45
6.4.1 For VGI system developers 45
6.4.2 For standardization organizations 46
6.5 Further research 46
7 References 47
8 Appendix A: Web scraping script 52
9 Appendix B: Overview of RDF Ontologies, namespaces, publishers and use cases 53
10 Appendix C: Data structures of sampled projects. 54
10.1 Use case 1: TrashHunters and MORA. 54
10.2 Use case 2: hetweeractueel.nl & KNMI 55
10.3 Use case 3: AiREAS & enviroCar 56

List of figures

Figure 1. The original Semantic Web layer cake by Tim Berners-Lee (2000) on the left and a more recent adaptation by Domingue, Fensel and Hendler (2011) on the right. 4
Figure 2. Excerpt of visual representation of the developed VGI ontology. 21
Figure 3. Example of query results in a triplestore. 27
Figure 4. Results of the query in Listing 1, displaying metadata from MORA, TrashHunters and Amsterdam. 30
Figure 5. Results of the query in Listing 2, displaying VGI contributions by volunteers and their submission date. 31
Figure 6. Results of the query in Listing 3, displaying the available parameters and units of measurement for hetweeractueel.nl and KNMI. 33
Figure 7. Results of the query in Listing 4, displaying the data for Hoek van Holland on May 5th 2016 (edited for legibility) 35
Figure 8. Results of the query in Listing 5, displaying the units of measurements and data quality indicators for AiREAS and enviroCar (edited for legibility) 36
Figure 9. Excerpt from TrashHunters in JSON format (edited). 54
Figure 10. Report made in MORA in JSON format. 54
Figure 11. Result of web scraping hetweeractueel.nl on May 5th 2016. 55
Figure 12. Climatological data from the KNMI on May 5th 2016. 56
Figure 13. Excerpt of CSV data from AiREAS on 22nd of May 2016. 57
Figure 14. Excerpt of JSON data from enviroCar on 22nd of May 2016. 58

List of listings

Listing 1. Use case 1: TrashHunters & MORA metadata query. 30
Listing 2. Use case 1: TrashHunters & MORA data query. 31
Listing 3. Use case 2: hetweeractueel.nl & KNMI metadata query. 33
Listing 4. Use case 2: hetweeractueel.nl & KNMI data query. 34
Listing 5. Use case 3: AiREAS & enviroCar metadata query. 36

List of abbreviations

API	Application Programming Interface
CSV	Comma Separated Values
DCAT	Data Catalog Vocabulary
DERI	Digital Enterprise Research Institute
DQV	Data Quality Vocabulary
Exif	Exchangeable image file format
FOAF	Friend of a Friend
FOI	Freedom of Information
GI	Geographic Information
GPS	Global Positioning System
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
KNMI	Royal Netherlands Meteorological Institute
LOD	Linked Open Data
MORA	Melding Openbare Ruimte Amsterdam
OGC	Open Geospatial Consortium
OSM	OpenStreetMap
OWL	Web Ontology Language
RDB	Relational Database
RDF	Resource Description Framework
RDFa	Resource Description Framework in Attributes
RDFS	Resource Description Framework Schema
SDI	Spatial Data Infrastructure
SPARQL	SPARQL Protocol and RDF Query Language
SSN	Semantic Sensor Network
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VGI	Volunteered Geographic Information
W3C	World Wide Web Consortium
XML	Extensible Markup Language

1 Introduction

1.1 Background information and research motivation

The internet has been growing in volume, variety and velocity since its conception. More people are than ever are connected to it and using it (International Telecommunication Union, 2015). Some data is published publicly on the web while others are used internally within organizations. Many websites require credentials and subscriptions in order to use them. Overall, the internet is becoming more embedded in the lives of people (at least in the industrialised world).

Data that is publicly available on the internet, sometimes contains geospatial characteristics (e.g. place names, coordinates) or refers to a certain spatial phenomenon (e.g. urban development, landmarks). As a result of the advent of the so-called Web 2.0 techniques and standards, data can be published by a wide variety of actors. Authoritative sources such as research- and governmental institutes publish research and public data, news agencies publish news articles and the general public publishes comments, opinions and photographs on social media or discussion boards. These last examples are instances of so-called user-generated content (UGC). UGC does not have gatekeepers to assess and moderate the information or data that is published. Another source of UGC are citizen science projects that disseminate data to the public and their members. A subset of UGC is volunteered geographic information (VGI) (Goodchild, 2007).

Depending on the website or project, the data can have multiple sources and authors and is therefore considered heterogeneous. Furthermore, the data can be structured or unstructured, meaning it allows for computations with this data in a straightforward- or more comprehensive, extract, transform, load (ETL) way. This research aims to reuse geospatial data that is available on the internet.

One of the most well-known and analysed examples of VGI is OpenStreetMap (OSM). OSM is a collaborative mapping project launched by Steve Coast in 2004 to create and disseminate a global map of the world that is free to use, as long as OSM and its contributors are credited (OpenStreetMap Foundation, 2016). OSM functions as a basemap alternative to satellite- or aerial imagery and proprietary or national mapping agencies vector maps. VGI is characterised by having some form of citizen's participation in the collection of georeferenced information (Kamel Boulos et al., 2011). Over time, VGI has manifested itself in many different forms. The practice of using citizen's or volunteers as agents in a mapping process is not new (see Elwood, Goodchild, & Sui, 2013), but the technologies that are being employed are, and have given rise to a wide variety of VGI systems.

Some of the enabling technologies behind VGI have been identified in Goodchild (2007), namely: web 2.0 standards, georeferencing, geotags, the global positioning systems (GPS), improved computer graphics and broadband communication. The combination of these technologies has given rise to VGI in a multitude of domains.

In the emergency response domain, *Ushahidi* is a platform that allows for VGI creation, analysis and dissemination (Ushahidi, 2016). In the domain of public services, *BuitenBeter*, *FixMyStreet* and *Verbeterdebuurt* are smartphone applications that allows citizens to report graffiti, trash or other public nuisances to the municipality. In the transportation domain, *Inrix* collects data from trucks and fleets of vehicles to compile real-time information on congestion levels. In the environmental domain,

iObs is a smartphone application that allows volunteers to submit and share observations of animals with a community of roughly 70.000 users (Stichting Natuurinformatie, 2016).

In a 2009 survey, Elwood et al. (2013) discovered 99 VGI initiatives by keyword search, but acknowledge that there may be many more and some of the ones found are not be active anymore. This is to say that the field of VGI is diverse and ever changing.

This variety of systems has led to a plethora of methods and techniques to collect, process, analyse and disseminate (geographic) data. It has been established that VGI in general is heterogeneous (Elwood, 2008). Heterogeneity is one of the main difficulties when reusing VGI.

The reason for this heterogeneity appears to be inherent to the concept of VGI itself. Employing the diversity of citizens, 'produsers', or scientific communities leads to heterogeneity on multiple levels. Creators of a VGI system might be well versed in a certain dissemination method and will choose a certain method over another even if the second method would allow for later reuse of the data. Furthermore, there can be ambiguity in the semantics when collecting data. What one person might call a 'mountain', another may call a 'hill'. This obviously depends on the kind of data that is gathered and the level of standardization that is used by the creators. But the human understanding can be a large part of a VGI projects.

Another disadvantage of VGI that hinders reuse is the quality of the data. Accuracies can vary significantly because of differences in equipment, the quality can be undocumented, its coverage can be incomplete and VGI projects do not necessarily apply scientific rigour when it comes to sampling. In general, these are valid assumptions about the nature of VGI however as Goodchild & Li (2012) have argued there are methods to circumvent these issues. Furthermore, several researchers suggest that solving heterogeneity according to one schema should not be the main goal, quite the opposite; the diversity of VGI lends itself to a more diverse view on the world. Exploiting the diversity of VGI can provide new insight into complex multi-disciplinary problems.

Reusability of data is particularly important for public organizations that have to abide to freedom of information (FOI) laws. FOI laws grant citizens a right to certain data. Furthermore, public organizations have a 'collect once, use many times' incentive towards other public organizations. Reusing data is also of particular interest to the academic field, where reproducibility of results is of fundamental importance. Another field of interest for the reusability of data is the data integration community. Researchers and developers that try to combine multiple sources of data in order to inspect or update datasets, georeference data, enhance the geometry of objects or semantically enrich data (Butenuth et al., 2007).

In order to reuse data from multiple VGI-systems there needs to be an understanding of the context and semantics of the data. From this context, a secondary user can determine if data is applicable for a certain purpose. There is a collection of technologies that help facilitate this process of contextual information on the internet. So-called Semantic Web technologies are a stack of internet enabled data models, standards and query languages that can help enable reusability.

1.2 Semantic Web technologies

In 2001, Tim Berners-Lee, James Hendler and Ora Lassila published an article in *Scientific American* titled *The Semantic Web.* It describes their vision of a possible evolution of the world wide web. It provides a summary of the techniques and standards they envisaged that could help in this endeavour. By doing so it laid the groundwork for years of future development by the World Wide

Web Consortium (W3C). The goal of the Semantic Web is to have a web of data instead of the traditional web of documents (Shadbolt et al., 2006). Traditionally, data on the web are in formats such as comma separated values (CSV), Extensible Markup Language (XML) or Hypertext Markup Language (HTML) tables. Using these formats sacrifices much of the structure and the semantics of the original data (Bizer et al., 2012). HTML pages provide a document-based web, linked together by hyperlinks. In order to structure the web in a more comprehensive way, several standards and protocols are developed by the W3C.

The term Semantic Web technologies applies to several techniques, standards and protocols. The two-main building blocks of Semantic Web technologies are the Hypertext Transfer Protocol (HTTP) and Uniform Resource Identifiers (URI's). HTTP is part of the traditional web for hyperlinking; this makes Semantic Web technologies compatible with the existing web. However, hyperlinks do not specify what the relationship between two documents are. Semantic Web technologies use URI's to identify resources. A resource can refer to a multitude of things, for example, HTML pages, like Uniform Resource Locators (URL's), but it can also refer to strings or integers.

HTTP and URI's use a third building block named the Resource Description Framework (RDF) to encode data. The RDF data model consists of so-called triples. A triple is composed of a *subject*, a *predicate* and an *object*. The subject and object are resources or literals (commonly specified by URIs) and the predicate specifies the relation between the resources (also specified with an URI) (W3C, 1999).

To signify unique instances of data, Semantic Web technologies rely on ontologies (sometimes synonymously called vocabularies). These ontologies provide domain specific knowledge and create prefixes, which are identifiable by the URI's. These ontologies are written in knowledge representation languages such as Web Ontology Language (OWL). OWL supports the use of semantic reasoners *"that can make implicit facts explicit, discover incompatibilities, improve retrieval beyond keyword search, and provide the framework for complex integrity constraint checking that reduces the risk of combining incompatible data and models*."(Janowicz et al., 2012, p:322). Ontologies use relations, axioms, limits, domains and ranges to achieve this.

Data encoded in RDF needs methods to query, update and delete data. For this purpose, the SPARQL Protocol and RDF Query Language (SPARQL) was developed. SPARQL allows a user to post queries across multiple RDF databases.

In 2000, Tim Berners-Lee created a schematic overview of the components of his vision of the Semantic Web. Over time this Semantic Web layer cake (as it has become known) has evolved to incorporate more functions of the Semantic Web such as SPARQL and the need for cryptographic functions (see Figure 1).

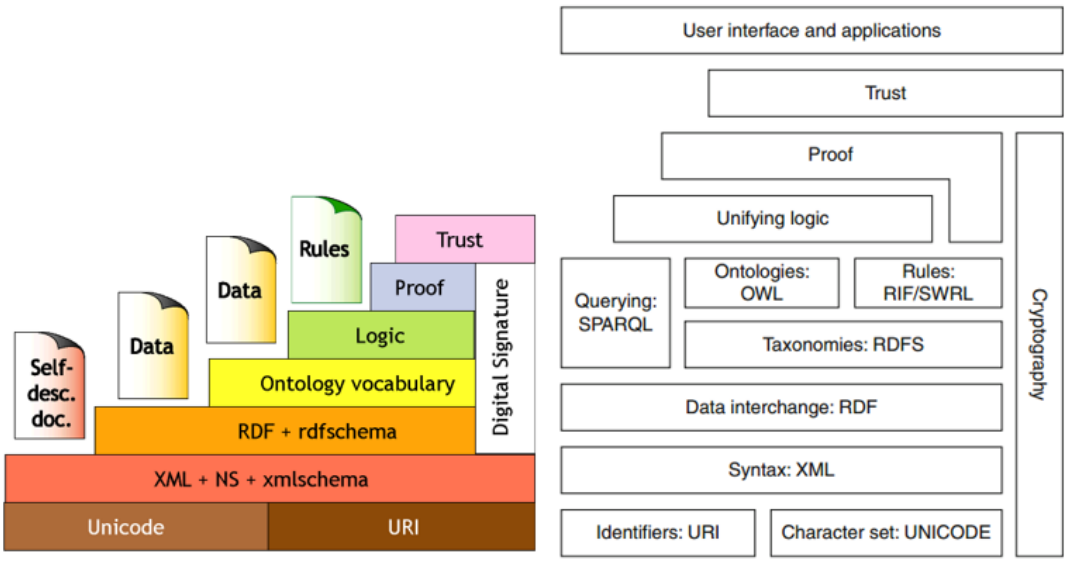

Figure 1. The original Semantic Web layer cake by Tim Berners-Lee (2000) on the left and a more recent adaptation by Domingue, Fensel and Hendler (2011) on the right.

Semantic Web technologies are already used for a wide range of applications. Notable fields are Linked Open Data (LOD) initiatives, search engine optimization, journalism, medicine, bioinformatics, geography, knowledge engineering and artificial intelligence. Some examples of applications that use Semantic Web technologies are recommender systems on Amazon, Netflix or Facebook, artificial intelligence bot development, rich snippets in Google query results, data integration and background information queries for journalism (Domingue et al., 2011; Shadbolt et al., 2006).

1.3 Research objectives

This research applies Semantic Web technologies to three use cases in the environmental domain. The use cases are composed of five VGI systems and one authoritative source, namely: TrashHunters and Melding Openbare Ruimte Amsterdam (MORA) for trash, hetweeractueel.nl and the Royal Netherlands Meteorological Institute (KNMI) for the climate and enviroCar and AiREAS for air quality.

The reusability and integration of heterogeneous data sources is becoming increasingly more important for solving multidisciplinary problems. Situations where a heterogeneous group of people from various (academic) backgrounds has to solve complex queries is becoming ever more prominent. Explicit semantics provide contextual information, which can put people on a level (knowledge) playing field. This thesis has developed a domain ontology for VGI that can be extended and further reused by other researchers and scientist. The following section discusses the research objectives of this thesis.

1.3.1 General research objective

The objective of this research is to develop a method to reuse VGI outside of its originating system. To assess this, the next chapter will address some of the characteristics of VGI and geographic information in general and their relation with reusability of data. This involves coping with syntactic and semantic heterogeneity and issues regarding metadata of datasources and datasets. By exploring and applying methods and techniques used in the field of the Semantic Web new insights can be

gained by inferencing the metadata and determining which existing ontologies can be used to model both metadata and data of the selected data sources.

The general research objective of this thesis is ***to develop a method that improves the reusability of volunteered geographic information by using Semantic Web technologies in the domain of environmental applications.***

1.3.2 Research objectives and corresponding questions

The general research objective has been divided into multiple research objectives and corresponding research questions as follows.

Objective 1: To identify the requirements for the reuse of VGI and application domains in which reuse is relevant.

1. *What are the elements of reusing VGI?*

   Research into this question shows the relevant aspects of data reuse. Some of the topics to be discussed are metadata, provenance, syntactic and semantic heterogeneity, interoperability, data quality, data collection, granularity, data retrieval and licensing.

2. *Which VGI application domains are applicable for (re)use cases?*

   Answering this question provides insight into the characteristics of the VGI application domains that were considered and which domains are the most likely to produce reusable data.

3. *Which VGI systems are available within the application domain and what are their characteristics?*

   This question will take into account the findings of question 2 and select multiple exemplary VGI systems to perform a case study.

Objective 2: To develop a method for improved syntactic and semantic interoperability between VGI systems.

4. *Which methods and tools are available to improve syntactic and semantic interoperability?*

   This question addresses the issues of syntactic and semantic heterogeneity in more depth. It will discuss methods such as data mapping, semantic enrichment, Extract Transform Load, data integration and ontologies.

5. *How can the data from the selected VGI systems be accessed and how is it structured?*

   This question discusses the current data access mechanism available on the web for reusing VGI. It provides the advantages and disadvantages of web APIs, download services and web scraping.

6. *What is the best way to structure data from VGI systems for reuse?*

   This answer combines the findings from question 4 and 5 to develop a method for reusing VGI.

7. *How to implement the use of semantic descriptions of VGI?*

   The answer to this question discusses the implementation methods for metadata, semantic enrichment and creating a VGI knowledge base that enhances reusability of data.

Objective 3: To construct a VGI knowledge base as a proof of concept in support of VGI reuse.

8. *What visualization tools for RDF are available and how can they assist in creating a knowledge base for reusable VGI?*

   This question provides insight into methods that allow the results in RDF to be visualized.

9. *What is the best method for the data from the VGI system to be in disclosed in a knowledge base?*

   The knowledge base functions as a method of validation of the research.

Objective 4: To recommend improvement of current standards and recommend reusage strategies to VGI application builders.

10. *Which improvements can be made to the existing metadata- and dissemination-/interface standards?*

    By the experience gained from developing the method and the construction of the proof of concept recommendations can be made regarding the standards that were considered and used.

11. *What reusage strategies can be employed by future VGI system developers and what are the consequences of certain decisions?*

    For future developers of VGI systems, recommendations can be made in the form of reuse strategies. These entails recommendations on the data input, data output, standards and licensing of data.

1.4 Relevance of research

As every thesis research project, the performed work must have some relevance. This section first provides the scientific relevance followed by the societal relevance of this thesis.

1.4.1 Scientific relevance

Reusable data enhances the scientific reproducibility of research. Even though verifiability is a cornerstone of scientific research, in practice it can still be cumbersome for researchers to reproduce the results or understand the data from other researchers. Reusable data allows researchers to perform aggregated studies of a certain phenomenon. Reusable data creates efficiency benefits for researchers for example, it saves them time in collecting data.

Preferably, computer readable scientific data should be interoperable, storable and reusable for many centuries. Many digital preservation initiatives have been initiated but there remain research challenges for which Semantic Web technologies can proof useful (Schlieder, 2010). Computer readable data has been around for decades, perhaps a century, but not much longer. The earlier we clarify methods and techniques that improve the long-term usability of data the better.

Furthermore, the evidence for climate change has become overwhelming. By investigating the reusability of data in the field of the environment, this thesis provides insight and support into VGI environmental research and science. It provides an RDF ontology that can be reused by future researchers that can be extended for future purposes.

1.4.2 Societal relevance

The collection of VGI is perceived as a societal beneficial activity, primarily referring to emergency response, crisis management and mapping of scarcely mapped areas (Feick & Roche, 2013). Also in the fields of tourism and health tracking there are several applications that benefit from VGI. If the data from these VGI applications can be reused for multiple purposes this will likely increase the exposure of VGI and in time perhaps the willingness of people to participate, reinforcing the societal beneficial activities. The motivation for participation in VGI is complex and determines on many factors, however exposure to more data-driven decision making can prove to be a positive influence on these factors.

1.5 Thesis outline

Chapter 1 introduced the research subject and stated the research goal, research questions and the thesis outline. Chapter 2 discusses some of the related work that has been performed and how this relates to the reusability of VGI. Chapter 3 describes the methods that have been applied throughout the research focusing on the case study and proof of concept approach. Chapter 4 provides a method that can be used to improve and analyse the reusability of VGI. Chapter 5 describes the application of the proof of concept. Chapter 6 provides conclusions on the stated research questions and recommendations for future developers of VGI systems.

2 Related work: Reusability of volunteered geographic information

This thesis addresses the reusability of VGI for a secondary purpose. In order to do so, the following chapter clarifies the notion of VGI and reusability. What are the elements we need to consider when reusing data for a secondary purpose? What are the current methods of reusing VGI, and what are their advantages and disadvantages? The answers to these questions are followed by theoretical solutions, methods and tools.

2.1 A definition of VGI

The concept of VGI is often traced back to Goodchild, (2007) however arguments have been made that VGI has been around for a longer time, in particular in the field of citizen science (Elwood et al., 2013; Kamel Boulos et al., 2011). A recent study on terms related to VGI by See et al., (2016) provides a review of the terminology of what in this research is considered VGI. See et al., (2016) trace 26 different definitions of VGI-related terms and keywords. They have reviewed their origin and relative importance to grasp the notion of VGI better. The point they make is that there is an ongoing shift taking place, where the creation of geospatial data is no longer solely in the hands of professional organizations and is proliferating to citizens. User-generated content (UGC), contributed geographic information (CGI), public participation in geographic information systems (PPGIS), neogeography, citizen-contributed geographic information (CCGI) – to name a few – are all terms that signify this shift. Overall, there are more commonalities between these terms than there are differences. Without the advent of new internet standards and miniaturization of sensors in smartphones this proliferation would not be possible.

In order to encompass the diverse landscape of VGI systems, this research uses a relatively broad definition of VGI, namely: *online data with a geospatial component (e.g. coordinates, place names) that is available (free of charge) publicly on the internet, structured and unstructured.* It is up to subsequent readers to assess whether the case study performed should have a narrower definition and if the selected projects fall under other definitions of VGI.

2.2 Elements of reusability for VGI

This section describes the elements that are of importance when reusing VGI. It first describes some frequently mentioned problems of VGI that hinder reusability. This is followed by elements that aid the reuse of VGI.

2.2.1 Heterogeneity, interoperability and quality of VGI

The first problem of reusing VGI is that VGI is inherently heterogeneous. This has numerous causes for example, the diversity (on the lowest granular level) of multiple contributors submitting entries to a dataset, the use of different systems (including standards, equipment, units and measurement techniques) and the different semantic interpretations humans have of the real world and the provided data. This heterogeneity complicates the ability to reuse VGI for different purposes

than for what it was originally collected. These problems can be referred to as heterogeneity and interoperability problems.

According to Pagano et al., (2013), the term interoperability is too often related with the technical issues that arise when data is exchanged or integrated. However, they argue that interoperability can also arise from organizational and semantic issues. See et al., (2016) finalize their article with a list of areas for further research, which also mentions data interoperability as an important problem that VGI faces. Interoperability is a beneficial feature of reusable data because it increases the number of people that are able to use it. Based on a six levels of interoperability between two GIS systems Bishr (1998) identifies three types of heterogeneity, namely: semantic-, schematic and syntactic heterogeneity. A common method of creating interoperability between systems is to develop standards. However, standards are primarily concerned with creating interoperability on a syntactic and schematic level (Harvey et al., 1999). For data to be actually reused for a secondary purpose, there needs to be interoperability on the semantic level.

Semantic interoperability is not just concerned with the syntax and data format of data, but also how data relates to other concepts and its inherent meaning. There appears to be no agreed upon definition of what semantic interoperability exactly entails. Kuhn (2005) calls semantic interoperability the only useful form of interoperability, and therefore creating definitions for it might be a bit redundant. However, the European Commission has noticed its importance and made semantic interoperability a key interoperability area. It provides the following definition:

"This aspect of interoperability is concerned with ensuring that the precise meaning of exchanged information is understandable by any other application that was not initially developed for this purpose. Semantic interoperability enables systems to combine received information with other information resources and to process it in a meaningful manner." (European Communities, 2004, p:16)

According to this definition reusing data for secondary purpose is a vital part of semantic interoperability. It must be said however that this document is not solely or specifically related to geographic data, but for all sorts of information in the field of eGovernance. Harvey et al. (1999) provides perhaps the most comprehensive overview of semantic interoperability in the field of GIS. However, the article revolves mostly around authoritative GIS systems, formats and standards.

Goh (1997) has identified three main reasons for semantic heterogeneity:

- **Naming conflicts** consist of synonyms and homonyms among attribute values. This means similar real-world concepts are referred to by different conventions, for example IBM and I.B.M. refers to the same organization.
- **Scaling and unit conflicts** occur when different units or measures are used in multiple sources, for example different currencies or temperature degrees.
- **Confounding conflicts** occur when the same term is used to describe something, but in the real world they refer to different things. For example, the term "SDI" can refer to a "Spatial Data Infrastructure" or the "Strategic Defense Initiative".

All three of these causes for semantic heterogeneity are found in VGI systems.

Another problem of reusing VGI is its data quality. Because of the lack of moderation, standardization, relatively low-quality equipment, vandalism and other reasons, there are serious concerns regarding the data quality of VGI in general. These concerns should be kept in mind when reusing data from VGI projects for a secondary purpose.

Standards have been proposed to assess the data quality of VGI. Antoniou & Skopeliti (2015) provide an overview of VGI quality elements and propose quality indicators by using VGI's unique

features as proxies for data quality. Antoniou & Skopeliti (2015) provide several astute observations about VGI such as the notion that VGI is of particular interest to researchers and governments, but not much to other geo-information professionals. They believe data quality of VGI is one of the largest reasons for the slow uptake by geo-information professionals. Also, their notion of participation biases should be acknowledged; VGI is heavily reliant on (high speed) internet access, knowledge of languages (English in particular), a user's available time and technical abilities. These issues where first found by Holloway et al. (2007) for Wikipedia but they also hold true for VGI projects.

Fonte et al. (2015) have posited guidelines that can be helpful when assessing the data quality of VGI systems, these entail guidelines regarding provenance information, automated consistency checks and allowing revisions by users, amongst many other suggestions. However, their work is mostly focussed on VGI systems that gather geographic features, such as OSM.

Goodchild & Li (2012) have posited three alternative approaches by which VGI data quality can be assessed. They focus on a wider variety of VGI systems, underlining that the appropriate method will depend on the VGI system that is assessed. Their approaches try to make use of the crowd as much as possible. The 'crowdsourcing approach' relies on the 'law of large numbers' where given enough eyes the crowd will approach the correct value. Their 'social approach' relies on moderators and gate-keepers to negate users that provide bad quality data. And the 'geographic approach' is based on logical rules for example, if a theatre feature is placed in a national park the contribution should be flagged, because it is highly unlikely that there would be a theatre in a national park. Goodchild & Li (2012) acknowledge that this approach is still experimental especially for VGI projects, but has a lot of potential for VGI systems that gather geographic features.

In a recent academic workshop, Mocnik et al. (2017) argued that the context in which data - and in particular VGI because of its heterogeneity - is collected can be a proxy of the data quality that is associated with it. To know the context in which the data is grounded, improves the quality of that data. A critique on research about VGI is that much empirical work, in particular on data quality, is based on empirical work with OSM, which is not representative of VGI projects.

Apart from heterogeneity, interoperability and data quality, there are two other issues regarding the reusability of VGI, namely: findability and accessibility that hinder the reuse of VGI.

2.2.2 Findability and accessibility of VGI

Two other aspects to reusing data on the internet are the findability and the accessibility of it. For over two decades the Open Geospatial Consortium (OGC) has been developing standards in order to publish and disseminate data on the internet, such as the WMS, WFS and the WCS. However, as mentioned by Taylor & Parson (2015), much of the valuable information behind these services are not accessible to most internet users, both human or automated. They provide very little explanation on why this is the case in their opinion. However, the frames that are used to display OGC maps are distinctly different from text on common HTML pages. This means that web crawlers that index the internet for search engines do not understand the frames and the data within them that are created by OGC standard maps.

The data availability of VGI projects can vary significantly. From being an unavailable for reuse (used internally), to downloadable as files (CSV, KML, XML) and availability via a web API. Sometimes authorization or API keys are required. See et al. (2016) assume that the availability of a web API reflects a significant external demand of data. While an external demand for data might be a factor in developing a web API, it should be kept in mind that APIs are already used internally to develop

applications. Some project might expose their data without any sign of external data demand, but for internal developmental or educational purposes.

These problems of findability and accessibility have long been recognized by the academic- and governmental community. The development of service oriented architecture (SOA) spatial data infrastructures (SDIs), such as nationaalgeoregister.nl and inspire-geoportal.ec.europa.eu has reduced these barriers for certain datasets. These often government-backed initiatives, have the resources to use OGC standards. However, VGI projects are often much smaller in scale and do not have the resources it requires to adhere to comprehensive standards that allow the reuse of their data.

2.2.3 Metadata, provenance and licensing of VGI

Metadata in terms of ISO:19115, INSPIRE or the Federal Geographic Data Committee (FDGC) standards are not common among VGI projects. Metadata of VGI is often provided on the website of the project or the available query methods of the available web API (Elwood et al., 2013; See et al., 2016).

Comber et al., (2007) argue for metadata that relate to the productive use of data. According to them the current metadata standards are too producer-centric. Instead of focussing on the usability, metadata should focus on the usefulness for secondary users. They provide seven recommendations to improve metadata. (Comber et al., 2007) focus their experience primarily on authoritative datasets for land cover. However, their findings on the different semantics of public organizations also holds for the different semantics of VGI projects. In the epilogue to *Quality aspects of Spatial Data Mining*, Goodchild (2009) expands on the call for more user-centric metadata and proposes a framework based on web 2.0 methods and techniques.

The term provenance (also known as lineage) has been around in the field of geo-information for quite some time. It refers to information about the sources and production processes of a geospatial product (He et al., 2015). Lanter (1991) first described the advantages of having this information for the purposes of data quality. GIS have purposely built modules that track the changes of a dataset or geodatabase. In the Semantic Web technology field, there are experimental tools to that address the issue of provenance (Hoekstra & Groth, 2015). Especially because of the multitude of content creators on the web, it could be argued that provenance information for VGI is even more important than for traditional datasets. For geographical data this means the registration of the data, collection methods and the operations that have been performed on it (Di et al., 2013). Frew (2007) argues that modelling provenance in VGI can improve the quality of the data. Depending on one's point of view or applied definition, provenance can be seen as a part of metadata.

Another element of reusing data is licensing. The production of geographical datasets can cost lots of resources, especially when it comes to high quality-, high resolution- and timely data. The owners or license holders to datasets sell licenses in order to retrieve those resources and perhaps turn a profit. When a dataset is licensed in such a way, it becomes costly for a secondary user to reuse that dataset. This same problem is relevant in the case of VGI, a point made by Heinzelman & Waters, 2011, in the case of the Haiti earthquake relief effort. However, because of the crowdsourced nature of VGI the overall costs of production are spread across a wide user base and because most VGI projects rely relatively cheap equipment (i.e. smartphones) it would be fair to assume that VGI projects are less likely to license their datasets.

2.3 Current methods for reusing VGI

Data from VGI projects is already reusable to a certain extent in various ways. This section will explore the most common methods that can be used to reuse data from VGI projects.

2.3.1 Application programming interfaces

The term application programming interface (API) has been used in computer science for a long time. In general, it refers to methods where software components can interact with each other. This can be within an operating system, between programming libraries or over the internet. In this thesis, we are interested in APIs over the internet, so-called web APIs.

Web APIs use HTTP request- and post messages to communicate between a server and a client. The client in this case can be an internet browser or an application that initiates a request to a web API endpoint. This web API endpoint is published by developers on the server side, along with some documentation that shows which requests the web API will respond to. Once the client has send a valid request, the server will respond with the data that has been requested. Some common formats that are used for web APIs are JSON and XML on both the request as the response side.

Web APIs allow software developers to use data from a project programmatically for a different purpose. For example, Google Maps has several (paid) web APIs that allow Airbnb, Citibike and Expedia to use their data (Google, 2017). This has the added benefit that when Google updates features in Google Maps, the applications that use their web API also use the most recently updated data.

2.3.2 Download services

A second method to reuse data from VGI projects is by using a download service. The VGI project creates an interface or website where users can download (pieces) of a dataset with their internet browser. The benefit of this method is that it is simpler to setup than creating a web API that responses to specific requests. The downsides are that it is not possible to specify which (pieces) of data you want. The client is dependent on the granular level of the download service. Furthermore, download services create local copies of data so if the original data updates the reuser remains with inaccurate information. Common data formats when using download services are CSV, PDF, and JSON.

2.3.3 Web scraping

The final method to reuse data from VGI projects is a method that can be applied to any webpage, namely web scraping. Web scraping is essentially copying a website and making a local copy of it. Web scraping software use parsers to read the HTML of a website to determine its structure. Once it has parsed the website, the user can tell the scraper which sections to scrape and store locally. There is a wide variety of tools available for web scraping, which tool to use depends on the website (Mitchell, 2015).

One benefit that web scraping has, is that it does not require any setup on the VGI projects side, besides hosting the website where the data is displayed. A downside is that there may be relevant information not displayed on the page (such as units of measurements) that are not being copied when scraping. Furthermore, web scraping has the same downsides as a download service; local copies and the scraper has to make sure each change in the website is scraped.

Scraping generally works well for non-perishable data, data that is not subject to change. For example, a list of books published by an author in the 19th century is easy to scrape and won't change anymore. Perishable data, like stock prices on the other hand would be less useful to scrape since it changes constantly and would need to be scraped at every iteration (Grimes, 2014; Morrison, 2015).

Of these methods web APIs are the most desirable method of disclosing information; however, it also requires the most resources on the side of the VGI project. Web scraping should be a last resort option since a lot of the semantic information of the data could get lost.

The implementation of the different methods for this research are available in Chapter 3 of this thesis.

2.4 Frameworks and standards

The first chapter of this thesis provided a brief introduction to the basics of Semantic Web technologies. This section continues the explanation of these frameworks and explains how they can help improve the reusability of VGI by creating a web of data, instead of the traditional web of documents.

2.4.1 The Resource Description Framework

RDF is a standard model by the W3C that allows data exchange on the web. RDF allows data to be encoded and queryable over the internet, more specifically over HTTP. The benefit of this is that data is more easily found by so-called web crawlers, improving the findability of the data. Egenhofer (2002) made the point that with the increasing of growth of the World Wide Web, its complexity also grew. This complexity means that it is becoming increasingly more difficult to *"compare, query, analyse combine, or integrate data due to the lack of methods that make compatible information available"* (Egenhofer, 2002:p1). More than a decade later we are still struggling with web crawlers in the geographic domain (Huang & Chang, 2016). Huang & Chang (2016), explain the relevance of what they call The GeoWeb Long Tail, where data hosted by large geoportals and SDI's is indexed and used more often than data that is hosted by small portals, individual researchers and small institutes, even though the amount and relevance of the data in the long-tail may be just as relevant as the data in the large geoportals and SDIs.

Besides the triple structure mentioned in Chapter 1, the Semantic Web relies on ontologies to create the predicates between subjects and objects. Just like RDF data, ontologies are also made up of triple statements. Ontologies are written in the W3C's Web Ontology Language (OWL), these ontologies allow the creation of classes and allow the creation of relationships between classes. By using an OWL engine, a user can infer information that was not in an original dataset. For example, consider the triples [movies] [are watched by] [people]. [Bruce] [watched] [Blade Runner]. [Blade Runner] [is a] [movie]. With a proper OWL ontology, an OWL engine can infer that Bruce must be a person.

By using RDF, OWL and SPARQL this information can be inferred without the creation of additional tables and is considered one of the main benefits of storing data in triples opposed to relational databases. The Semantic Web uses OWL reasoners that search for certain predicates that allow the inferencing of information (DuCharme, 2013).

Brodeur (2012) distinguishes three levels of ontologies in the Semantic Web: global ontologies, domain ontologies and application ontologies. Global ontologies provide generic terms that can be used for any type of dataset. The Dublin Core Metadata Initiative (DCMI) is an example of a global ontology. It can be used for a wide variety of datasets and consists of broad terms independent of any specific field. Domain ontologies are more specific and often relate to a specific technology or scientific field. The National Aeronautics and Space Administration (NASA) has developed the Semantic Web for Earth and Environmental Terminology (SWEET) ontology, this can be considered a domain ontology from Brodeur's point of view, even though it is composed of several other, more specific ontologies. Application ontologies are ontologies that are actually used within a Semantic Web application. This means that both global- and domain ontologies can retroactively become application ontologies once they are being used within an application. This is why Janowicz et al. (2012) only distinguishes top-level and domain ontologies. Where top-level ontologies refer to ontologies such as DOLCE, CIDOC and WordNet. Domain-ontologies focus on a specific (academic) field of interest, and enrich the top-level ontologies, making the difference between them irrelevant.

2.4.2 SPARQL Protocol and RDF Query Language

SPARQL is the query language for the Semantic Web. It understands the triple patterns of RDF because the queries also consist of triples patterns, only there are variables in at least one of the subject, predicate or object positions. A SPARQL query is made up of five segments, namely:

1. Prefix declarations, in order to abbreviate URIs.
2. An optional dataset definition, to state which RDF file to query.
3. A result clause, to select what information to display from the query.
4. The query pattern, the triple pattern – including a variable - that the SPARQL engine needs to search for.
5. Query modifiers, to set the order or rearrange the results.

SPARQL queries can be posted against remote SPARQL endpoints over the internet or against local triplestores. The results that are returned by a SPARQL query can be displayed in a variety of formats including HTML, CSV, JSON and XML (Feigenbaum & Prud'hommeaux, 2013).

2.4.3 Linked Open Data Initiatives

Open data is a long-standing strain of thought that has taken new forms in recent times with the initiation of government-backed data.gov, data.co.uk and the recently launched europeandataportal.eu. The thought is that data or at least certain datasets should be available, useable and republished for everyone's use (Algemene Rekenkamer, 2015; Carrara et al., 2015). The availability of data can aid political decision-making, scientific research, education and other fields.

The open data community and governmental agencies have also adopted Semantic Web technologies. This has led to four principles of Linked Data as states by Berners-Lee (2006), namely:

1. Use 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.

In order to help public organizations, in 2010, Berners-Lee published a five-star rating system for Linked *Open* Data. This five-star rating system works as follows:

Table 1. Tim Berners-Lee's five-star rating system for public organizations.

★	Available on the web (whatever format) but with an open licence, to be Open Data
★★	Available as machine-readable structured data (e.g. Excel instead of an image scan of a table)
★★★	as (2) plus: Non-proprietary format (e.g. CSV instead of Excel)
★★★★	All the above plus: Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus: Link your data to other people's data to provide context

Source: Berners-Lee, (2010).

Open data is relevant to reuse because reuse forms an important pillar of open data initiatives and the economic benefits that result from it (Carrara et al., 2015). Linked Open Data targets the reuse of authoritative data, for example public service information.

2.5 Methods and tools to improve syntactic and semantic interoperability

Reusability of data and integration of multiple heterogenous data sources relies on achieving syntactic and semantic interoperability. The current information technology landscape has several ways in which it deals with syntactic and semantic heterogeneity and achieving interoperability. This section will provide an overview of some of the used methods and techniques.

The first method to deal with syntactic and semantic heterogeneity is by an Extract, Load, Transform (ETL) process. This method is often used within large organizations with legacy systems and a multitude of databases. Vassiliadis (2009) has conducted a survey of different ETL technologies, and discusses what the persistent problems of ETL are, such as the provenance issue. Once data has been extracted, without provenance data the subsequent iterations of that data will also be without the original provenance data. ETL primarily revolves around the creation of new data warehouses to answer specific queries.

During the 'transform' phase, the data models of two or more data sources are matched and checked whether their content is semantically interoperable. This process of comparing multiple data models is also referred to as data mapping or schema matching. Data mapping is mostly done statistically or heuristically and by extension neglects the semantics of the data involved.

Another method to deal with syntactic and semantic heterogeneity are the use of ontologies. Ontologies explicitly state the relation between concepts. This allows them to be used for the integration of various data sources (Wache et al., 2001). According to Wache et al. (2001) ontologies are used to describe the semantics of the source data and use that information to see if multiple data sources match up with each other. Fonseca et al. (2002) have laid the ground work for an ontology-driven GIS that can be used for seamless and flexible data integration. In this case, every data source would have an ontology attached which could be explored or queried.

Semantic enrichment is the process of adding additional (meta)data to existing data, by linking data to already established concepts. These concepts can provide a secondary user with information about the syntax, conversion methods and other semantics of the data (Lemmens et al., 2016). This research combines ETL and semantic enrichment to propose a method to improve the reusability of VGI. This research provides an ontology for VGI to expose the semantics of the data sources and model their data.

The final method of dealing with semantic and syntactic heterogeneity is at present merely an academic exercise. In editorials for *Semantic Web* and *Int. Journal of Spatial Data Infrastructures Research*, Janowicz & Hitzler (2012) and Craglia et al. (2008), reinvigorate the Al Gore's notion of the Digital Earth (Gore, 1998). The Digital Earth emphasizes the need for the reuse, integration and application of geo-referenced data. Gore (1998) already mentions the need for interoperability and metadata, alongside computational power, satellite imagery and broadband network. As further research for the Digital Earth, Craglia et al. (2008) mentions the need for *"multi-source and heterogeneous, multi-disciplinary, multi-temporal, multi-resolution, and multi-media, multi-lingual"* information integration as a key field for computer and information science and domain disciplines.

Janowicz & Hitzler (2012), actually argue that Linked Data can provide a new approach to structure data for the Digital Earth. Furthermore, they argue that the answer does not lie in improving semantic interoperability, but that the semantic diversity should be encouraged as it provides us with new views on the world. By extension this leads to exposing the existing semantics of data and metadata as they are.

3 Methodology: A case study of environmental applications

This chapter discusses the research approach of this thesis, the multitude of VGI domains, the criteria to select a VGI domain, the use cases and corresponding data sources and the software tools that have been used in developed of the proof of concept.

3.1 Research approach

The approach to this thesis is that of a qualitative case study. The case study helps in order to develop a method for VGI reuse and a proof of concept to assess this method. In general, a case study refers to a detailed and intensive analysis of a single case (Bryman, 2008)**.** Yin (2003) distinguishes five types of case studies, namely: the critical case, the extreme or unique case, the representative or typical case, the revelatory case or the longitudinal case. This thesis most closely resembles a representative case, as it tries to use exemplary VGI projects and a variety of data access mechanisms.

VGI plays an increasingly important role in a number of domains. In crisis- and response management, it provides timely data (Okolloh, 2009). In tourism or city marketing, it provides an unprecedented wealth of qualitative data about hotels, attractions, restaurants and other facilities and amenities (Hauthal & Burghardt, 2016). The applicability of OSM to update cadastral maps is researched (Olteanu-Raimond et al., 2016). Fitness tracking applications use GPS to track subject's locations to assess their movement for health purposes (Griffin & Jiao, 2015). Transportation applications like Uber and Lyft rely on volunteered GPS locations for pickup and drop-off locations.

To research the reusability of the data from these kinds of applications one domain is chosen. This increases the chance that combining the data from multiple sources will actually be meaningful. The determination of this domain depends on the following criteria:

- The expected transparency and openness of data.
- Number of available applications.
- The availability of authoritative data.
- The integration of the data should provide a meaningful result.
- The expected applicability of cross-border research because this increases semantic differences between data sources.

These criteria initially lead to two relevant application domains: environmental- and smart-city applications (including tourism and health monitoring). Based on the availability of applications available and access to data, the choice was made to use environmental applications as the application domain.

3.2 Applications in the use case domain

Within the environmental domain, in the Netherlands, there are a variety of VGI applications available. iSPEX (http://ispex.nl) measures particulate matter for scientific research with the help of an attachment specifically developed for Apple's iPhones. BuitenBeter (http://www.buitenbeter.nl) provides notifications of nuisances in public space that is relayed to municipalities. MijnVismaat (http://www.mijnvismaat.nl) allows recreational fishers to share their catch with their friends and shares the information with the Dutch sport fishing associations, which uses the data for the creation

of publications. iObs is a smartphone application which is linked to the website waarneming.nl. Waarneming.nl conducts citizen-based censuses of animals. With help of an OBD-II adapter, enviroCar (http://www.envirocar.org) gathers information on the emissions of vehicles. The OBD-II adapter reads engine parameters and the smartphones uses GPS to determine its location. The data is processed to assess the relative air-quality. MORA (http://www.opdekaart.amsterdam.nl/mor) is a system created by the municipality of Amsterdam that is accessible via smartphone applications and the internet to report littering and trash. AiREAS (http://www.aireas.com) is a citizen science initiative that strives to improve the air quality in the city of Eindhoven. Hetweeractueel.nl (http://hetweeractueel.nl) is a network of weather enthusiasts that publishes climatological data from weather stations. TrashHunters (http://www.trashhunters.org) is an initiative to create awareness about littering. Contributors upload pictures and spatial information about litter that they have found and thrown away.

For the purpose of this research into reusability the aim was to find VGI applications that are somewhat related in order to reuse the data. The following use cases have been developed to illustrate the requirements that (re)users have.

3.3 Use case scenarios

Use cases are story-based scenarios, that exemplify behaviour of certain actors with a system. The method has been applied in various organisational contexts, including business and academia. A common method of creating use case scenarios is to identify the actors involved, set goals for the actors and develop a scenario in which these goals can be met (Cockburn, 2001; Jacobson et al., 2011). The following section will describe this process for three scenarios that include VGI.

The following use cases are developed to illustrate how VGI can be used outside of its originating system. Within each use case the systems deal with similar information, i.e. trash, weather and air quality. Furthermore, the choice was made to select data sources with a variety of data retrieval methods. A variety of retrieval methods provides a more representative view of VGI in general.

Table 2 provides an overview of the selected VGI applications and their characteristics. The goal was to have a diverse palette of VGI applications within the environmental domain. This means variety in their scope, initiators, licensing, geometry, data access mechanism and size of the community. For the most part, the selected applications fulfilled this requirement, except for the geometric features. All the information from these projects regard point geometry. However, this was also the case for alternative VGI projects in this domain, such as MijnVismaat and iSPEX.

Table 2. VGI applications and research characteristics for use cases.

Use cases	Use case 1		Use case 2		Use case 3	
Project name	TrashHunters	MORA	hetweeractueel.nl	KNMI	enviroCar	AiREAS
Research area	Amsterdam	Amsterdam	The Netherlands	The Netherlands	Eindhoven	Eindhoven
Project scope	National	Municipal	National	National	Global	Municipal
Public or private initiators	Private	Public	Private	Public	Both	Both
Data licensing	All rights reserved	Public	No information	Public	Public	Public
Geometry features	Points	Points	Points	Points	Points	Points
Data access mechanism	API	API and download service	Web scraping	Download service	API	Download service
Size of community	674	Unknown	316	30	500-1000	Unknown
Authoritative dataset	No	Yes	No	Yes	No	No

Use case 1: TrashHunters & Melding Openbare Ruimte Amsterdam (MORA)

TrashHunters is a sustainability project by the Plastic Soup Foundation, an organization striving for the reduction of plastic litter. The goal of TrashHunters is for volunteers to photograph plastic bottles, cans, cardboard packages and drink pouches on the street, take a picture – determining the brand and location - and dispose of the litter. Submissions to TrashHunters are posted on their Flickr account. TrashHunters uploads submissions in monthly batches to FlickR. A selection of 38 images in the area of Amsterdam-West is the dataset. The images are stored in the .jpg file format. The FlickR web API provides access to the Exif-data and coordinates. The data is available in JSON, JSONP, PHP and XML. JSON is the selected retrieval format for this thesis.

The municipality of Amsterdam has a smartphone application called CleanUp! which allows residents to submit incidents regarding trash that they would like to have cleaned up. The incidents are uploaded and published daily at 9:00am. Not all incidents are about picking up trash, also broken public collection points, graffiti and other nuisances are reported via the application. The data of the 11th of March 2016 was used in this thesis to develop the proof of concept.

Use case 2: The Royal Netherlands Meteorological Institute (KNMI) & hetweeractueel.nl

Hetweeractueel.nl is a website that functions as an aggregator for individuals with a weather station. It publishes the location, temperature, precipitation, wind speed, -power, -direction and air pressure for over 300 stations across the Netherlands. Furthermore, it provides historical data for each of the stations. In this case, the choice was made to use the data that is published on the overview page of the Netherlands. The historical data can be collected in a similar method as the overview page. Because there is no web API or download service, the data is scraped.

The data of amateur weather stations is combined with official measurements from the Royal Netherlands Meteorological Institute (KNMI). This data will not function as a source for validating the results of data sources but rather about the methods that are involved with reusing the data. The KNMI provides historical data in CSV format for all 30 of its active weather stations in the Netherlands.

Use case 3: enviroCar & AiREAS

enviroCar is a platform for citizen science by the 52°North research and development network and the Institute for Geoinformatica at Münster University (enviroCar, 2016). The enviroCar smartphone application uses an OBD-II Bluetooth adapter to connect to the smartphone and communicate the real-time on-board parameters. The smartphone interpreters these parameters and provides the driver with real time feedback on his driving. The data of enviroCar is available via a web API.

AiREAS is a citizen science association that strives to a cleaner city. Measuring air quality is the first implementation of this association. The data is processed and the results are posted on the website by a third party, named Scapeler (AiREAS, 2016). The unprocessed data is available via a download service.

3.4 VGI Ontology development

Ontologies are prevalent in two branches of science: philosophy where it deals with the nature of being and computer science where it defines concepts and the relations between them. This thesis will deal with ontologies in the sense of computer science. In computer science, ontologies are used for a variety of purposes namely, creating advanced computer applications, such as artificial intelligence programs (Gliozzo et al., 2013), to integrate information (Wache et al., 2001), or to structure the concepts and definitions of a certain domain (Noy & McGuinness, 2001).

In ontology engineering, there is an inverse relation between the usability and reusability of an ontology. This means that the more specific an ontology is the less reusable it becomes for other purposes. On the other hand, if an ontology is broad it becomes more reusable for secondary purposes, but it becomes too vague for actual applications (van Harmelen, 2011). A general rule for the Semantic Web is that it is better for an ontology to be reused than to be reinvented. This reduces the redundancy of ontologies (Heath & Bizer, 2011).

As the amount and variety of data on the internet keeps growing it becomes more important to create structure in this data. There are already many ontologies, however in the field of VGI their development has been limited. An internet search has revealed several projects and organizations that are working on or have worked on ontologies that are relevant to VGI.

The Open Geospatial Consortium (OGC) has put forth a request for participation for a Citizen Science Domain Working Group (Open Geospatial Consortium, 2016). The (draft) charter does not specify the use of RDF, OWL or other Semantic Web standards, but does mention semantics, metadata and reusability, as areas that need further research and ultimately standardization. As of yet, there is no defined ontology available.

Ramos et al., (2013) have demonstrated the use of a domain ontology to integrate VGI and authoritative data with each other. However, they only created a small proof of concept and not a full and available ontology. Since two of the three authors are now otherwise employed it is fair to assume their ontology development has also seized.

In Bakillah et al., (2013) the first steps are made to make a conceptual model that allows for the development of a VGI ontology. By aligning ISO standards with common attributes in VGI data they have created mappings that can help with the development of an VGI ontology. However, there is no ontology provided and the work remains mostly on a conceptual level.

The COST ENERGIC project has developed a VGI ontology in RDF (Lemmens, Falquet, & Métral, 2016). The ontology primarily focusses on metadata of VGI projects. (Meta)data is structured into three superclasses: information entity, research entity and VGI entity. For this thesis, inspiration was drawn from this ontology.

The developed ontology is used to expose the semantics of the data sources and their data. It is available online at: https://s3.eu-central-1.amazonaws.com/swarish-gima.com/index.html. Figure 2 provides a partial visual representation of the relationships in this ontology. Many of the classes and predicates are subclasses of the `VGI System` class. This visualization tool prefers to use `Thing` as a node when predicates do not refer to other classes. The visualization is made with WebVOWL (see section 4.4.2).

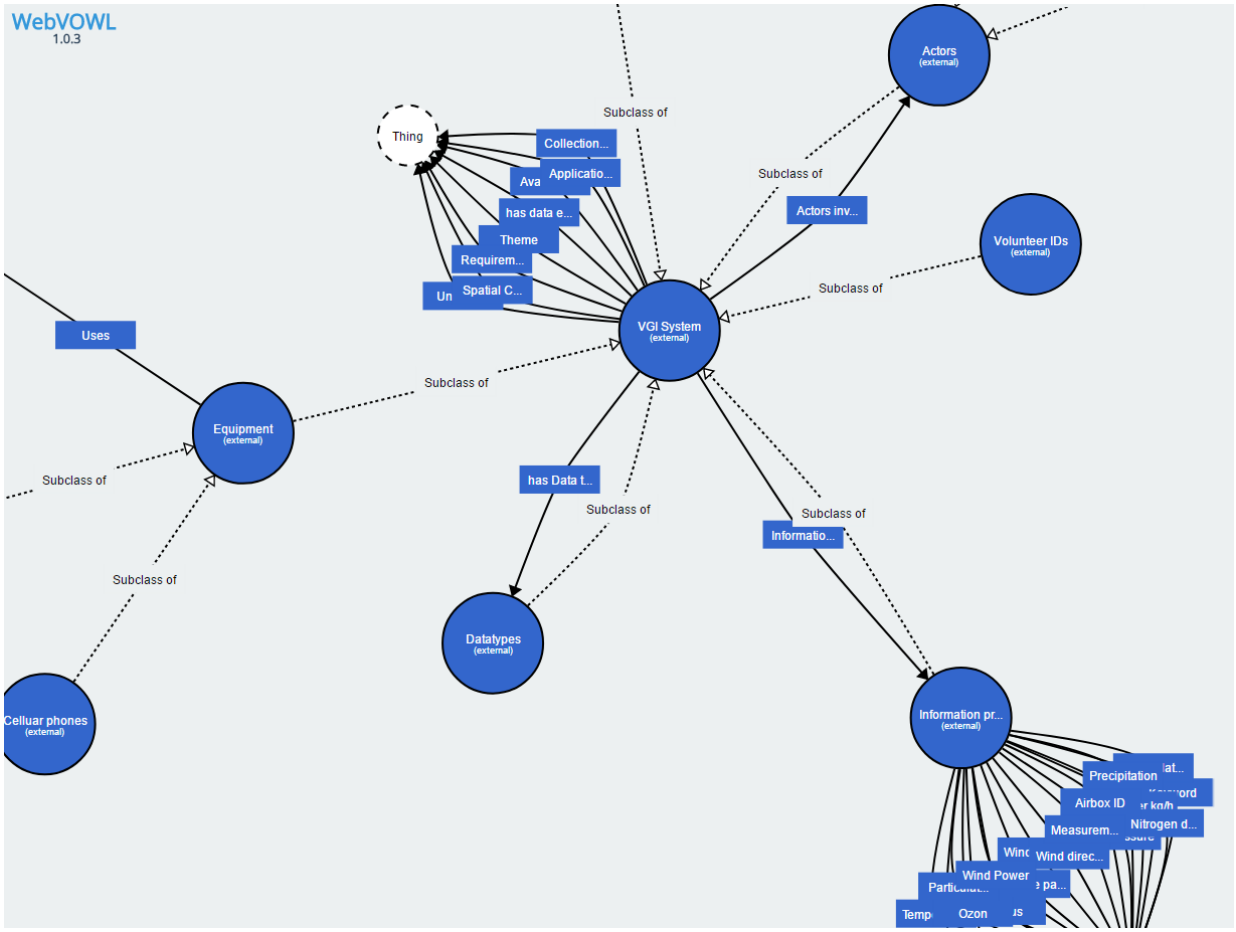

Figure 2. Excerpt of visual representation of the developed VGI ontology.

3.5 Software and tools

This section provides an overview of the software and tools that are used in the creation of the VGI ontology and the development of the proofs of concept.

rdfEditor (http://www.dotnetrdf.org/) is a Notepad replacement for editing RDF and SPARQL. It can parse multiple serialization of RDF to ensure the correct syntax is being used. It is part of the dotNetRDF Project. The application is used to write the VGI ontology, create the SPARQL queries and visually check the data of the VGI projects.

Parliament (http://parliament.semwebcentral.org/) is a triplestore developed by BBN Technologies and has been in use since 2001 (Kolas et al., 2009). It that supports GeoSPARQL which was considered for the third use case of this thesis.

Google Refine 2.5 (http://openrefine.org/) is a web browser based data management tool formerly developed by Google. It allows the editing, conversion and filtering of a wide variety of data formats. In combination with the RDF Refine (http://refine.deri.ie/) extension by the Digital Enterprise Research Institute (DERI), I it can be used to create RDF files from CSV, HTML and JSON files.

Protégé (http://protege.stanford.edu/) is an open source ontology editor and framework for building intelligent system. It is developed by the University of Stanford and first released in 1999. It and provides the user with a graphical user interface that allows the user to focus less on the syntax and more on the content of their ontology. There is a web- and desktop version available. Protégé is a relatively complex tool with a steep learning curve and for the purpose of this thesis too elaborate.

4 A method for the improvement of VGI reuse with Semantic Web technologies

This chapter develops a method that can be used to increase the reusability of VGI by employing Semantic Web technologies. The method is divided into four steps: gather metadata, gather data, model the data in RDF and querying the (meta)data by using SPARQL. Each step has a number of intermediate steps that are explained in the following sections. These steps are not an exhaustive list. Depending on the goals of the reuser there may be other characteristics that are relevant. This method is used in Chapter 5 to develop the mashups of (meta)data for the selected use cases.

1. Gather metadata
    i. Explore the website
    ii. Identify licensing situation
    iii. Identify data access mechanisms
    iv. Identify the purpose- and methods of data collection
    v. Identify coverages
2. Gather data
    i. Query web APIs, use download services or apply web scraping
    ii. Determine the relative quality of the data
3. Model the (meta)data in RDF
    i. Apply DCAT, DQV and VGI ontologies to metadata
    ii. Apply PROV, GEO, XSD and VGI ontologies to data
4. Upload and query the (meta)data
    i. Upload the (meta)data into a triplestore
    ii. Visualizing and exploring the results
    iii. Query the (meta)data using SPARQL

4.1 Gather metadata

Reusing data starts with an intended goal and domain in mind. Currently, there is no centralised place, like a geoportal or an SDI, where VGI projects can submit their (meta)data. Finding VGI projects that contain data of interest will most likely start at a search engine or by the reuser being aware of a VGI project in advance. Once a candidate VGI project is identified there are several steps to take in order to reuse the data.

4.1.1 Explore the website

The first step in determining the reusability of a VGI project starts with exploring the website. Unlike authoritative datasets from SDIs, VGI projects often do not provide standardised metadata according to ISO standards. The website of a VGI project is therefore a good place to start searching information for a reuse purpose. The website will clarify the intentions of the VGI project and provide the basic parameters to determine whether is it worth continue exploring the data.

4.1.2 Identify the licensing situation

Unless data is published under an open license, reusing data without a license agreement, written permission or proper accreditation is unlawful, unless otherwise stated. Examples of open licenses are the Creative Common license and the Open Data Common License (Open Data Institute, 2017).

4.1.3 Identify data access mechanisms

Determining the availability of the data is an important part of gathering metadata. In the case of Semantic Technologies, ideally the VGI project has a SPARQL endpoint. However, the adoption of Semantic Web technologies is limited and therefor it is useful to understand other methods of data gathering. For reasons mentioned in section 2.3, web APIs have certain advantages over download services. However, in this case the data will be extracted and reloaded into a triplestore so web APIs lose their benefit of being continuously up to date over a download service. If neither a web API nor a download service is available, but the data is visible on the website, web scraping is the only option left.

Web APIs generally come with instructions on how to get the parameters from the endpoint, therefor it might take some time to come to grips with the terminology of a web API. Download services are fairly straightforward to operate, but might require more post-processing work because it is not possible to input parameters (such as a specific user, date or time). The difficulty of applying web scraping depends heavily on the website where the data is displayed. Static HTML tables are fairly straight forward to scrape with existing tools and libraries on the web. However, when websites are highly dynamic and have many subpages, it becomes more difficult. Furthermore, setting up a database that keeps track of all the changes on such a website also requires efficient data management skills.

4.1.4 Identify the purpose- and methods of data collection

Understanding the context for which data was originally collected can help assess whether data fits the reuse purpose. It means gathering information about the level of skill and standardization that were involved in gathering the data in the first place.

To determine if the data fits the reusers purpose it is important to know which methods were used in the collection of the data. Certain methods or equipment might result in inaccuracies that are deemed too large for the reuser.

4.1.5 Identify spatial and temporal coverages

When reusing data, it is important to know whether or not two datasets will overlap with each other. Before actually gathering the data, it is therefore useful to find out what the temporal and spatial coverages are of the project.

4.2 Gather data

Once the metadata seems appropriate for reuse the actual data can be gathered. Which method to apply depends on the VGI project. The following section will explain three prevalent data gathering methods.

4.2.1 Query web APIs, use download services or apply web scraping

Querying web APIs consist of creating a correct HTTP request that an API endpoint responds to. An application like Postman (https://www.getpostman.com/) can help in the construction of APIs requests. In an attempt to limit abuse of an API, VGI developers may require you to obtain some form of authorization. Generally, you will need to send them a message and inform them how and why you will be using their API. The VGI project will provide you with a token that can be embedded in the HTTP request. The remainder of the steps to retrieve the data will depend on the extensiveness of the web API.

Much like web APIs, the functionality of download services depends highly on the VGI developers side. Download services have less flexibility than web APIs and provide data in dumps. Often this requires more post-processing activities, such as selecting the correct timeframe from within a CSV file.

Web scraping comes in many forms, there are commercial solutions that use graphical interfaces on websites such as Import.io (https://www.import.io/). There are command line tools such as Wget (https://www.gnu.org/software/wget/) and cURL (https://curl.haxx.se/). These are commonly free for use but also more cumbersome to use. There are internet browser add-ons such as Data Toolbar (http://datatoolbar.com/). Finally, there are web scraping libraries for specific programming languages. Some prominent Python libraries are: Beautiful Soup 4 (https://www.crummy.com/software/BeautifulSoup/), Pandas (http://pandas.pydata.org/), lxml (http://lxml.de/), Selenium (http://www.seleniumhq.org/) and Scrapy (https://scrapy.org/). Which tool is the most appropriate depends on the content that needs to be scraped. In this thesis, the Pandas Python library was used to scrape data for one use case.

4.2.2 Determine the relative quality of the data

After gathering the data, but before modelling it in RDF, it is important to assess whether the (meta)data suits the purpose of reuse. This means ensuring the spatial- and temporal coverage are correct and whether the data has been consistently gathered.

4.3 Model the (meta)data in RDF

RDF allows the creation of self-describing data in a web based format. For the purpose of this thesis a lightweight VGI ontology is developed to enhance the reusability of data from VGI systems.

4.3.1 Apply DCAT, DQV and VGI ontologies to metadata

The metadata about the VGI projects was not available in a standardized format, therefore it was developed from the ground up in rdfEditor. Whenever it was possible existing ontologies were used to limit the amount of redundancy. However, there are plenty of attributes that were not accounted for in existing ontologies. For that reason, a new ontology was developed, namely `vgi: <http://www.vgiprojects.com/>`. It has classes and properties such as `vgi:actorsInvolved`, `vgi:collectionPurpose`, `vgi:equipment`, `vgi:applicationDeveloper` etc. that are not found in other ontologies. The VGI ontology has been used in this research for metadata that was deemed relevant but was not modellable in other ontologies.

The DCAT ontology provides relevant properties that can be used to model metadata such as, `dct:distribution`, `dct:landingpage` and `dct:license`.

The DQV provides classes and properties to model the data quality of a VGI project or dataset. In this case the properties `dqv:value` and `dqv:UserQualityFeedback` have been used to model the data quality of two VGI projects (enviroCar and AiREAS). The values that have been used in the proofs of concept are demonstrative, because no actual data quality values were present in the metadata.

4.3.2 Apply PROV, GEO, XSD and VGI ontologies to data

The PROV ontology is structured around three basic concepts, namely the entity, the activity and the agent. In the scope of this thesis an entity refers to metadata and data. Activities are the means by which entities come into existence. Agents are the actors that perform the activity by which entities are created. Because of the importance of provenance information for the reusability these conventions were used in the modelling of the data of two VGI projects (TrashHunters and MORA).

The Basic Geo ontology is an ontology that allows for the modelling of longitudinal and latitudinal coordinates in the World Geodetic System (WGS) 84 coordinate system. If coordinates in a dataset are provided in a different coordinate system, a different ontology has to be used, or the coordinates must be converted to the WGS84 standard.

XSD is the XML Schema Datatypes ontology which is useful in modelling dates and times to specific data entries.

The VGI ontology is used to model the remaining data that was necessary for the use cases.

4.4 Upload and query the (meta)data

The remaining steps of this method consists of inserting the created (meta)data into a triplestore and querying it with SPARQL.

4.4.1 Upload the (meta)data into a triplestore

Ideally, every VGI projects would disclose their (meta)data by using the same ontologies and publish the SPARQL endpoint. This would allow for federated queries without uploading the (meta)data into a local triplestore. To take advantage of inferencing it is necessary to use a triplestore that has an OWL reasoner. The reasoner searches for specific properties and classes that allow inferencing of information (DuCharme, 2013). Figure 3 shows an example of metadata in a triplestore. This was achieved by retrieving all the subject, predicates and objects available in the triplestore.

SPARQLer Query Results

Home **Operations:** Query Explore SPARQL/Update Insert Data Export Indexes Admin

Count: 100

subject	predicate	object
http://www.vgiprojects.com/ID1	http://www.w3.org/2000/01/rdf-schema#label	"TrashHunters"
http://www.vgiprojects.com/ID1	http://www.w3.org/2000/01/rdf-schema#comment	"TrashHunters is a smartphone application created by the Plasic Soup Foundation and strives to reduce litter."
http://www.vgiprojects.com/ID1	http://www.w3.org/ns/dcat#landingPage	http://www.trashhunters.org/
http://www.vgiprojects.com/ID1	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/ns/dcat#Dataset
http://www.vgiprojects.com/ID1	http://www.w3.org/ns/dcat#Distribution	http://www.vgiprojects.com/api2
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/theme	"Trash"
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/actorsInvolved	http://www.vgiprojects.com/Organizations
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/actorsInvolved	http://www.vgiprojects.com/Volunteers
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/collectionPurpose	"The data is collected for a public awareness campaign by volunteers."
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/uses	http://www.vgiprojects.com/Smartphones
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/Smartphones	"iPhone 5"
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/Smartphones	"iPhone 6"
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/Smartphones	"Samsung S5"
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/hasDataTypes	"JSON"
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/hasDataTypes	"XML"
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/hasDataTypes	"JPG"
http://www.vgiprojects.com/ID1	http://www.vgiprojects.com/dataEntries	"34669" ^^<http://www.w3.org/2001/XMLSchema#integer>

Figure 3. Example of query results in a triplestore.

The benefit of this method compared to conventional (meta)data sharing methods, such as web portals, is that the VGI project does not have to adhere to the standards set by the web portal. Furthermore, it does not have to update the metadata at a third-party website and update the data there.

4.4.2 Visualizing and exploring the results

Visualizations can help a user to get a better understanding of the links between concepts, attributes and values. There are several tools that can aid this process of visualization of RDF. Many of these tools are however still in a prototype stage. WebVOWL (http://visualdataweb.de/webvowl/) is a web-based visualization tool for ontologies. Ontologies can be uploaded (to the browsers cache, not an external server) and WebVOWL creates a visual overview of the relations between concepts. The SPatiotemporal EXplorer (or SPEX) (http://giv-lodum.uni-muenster.de/spex/) is developed by the University of Munster and allows a user to construct exploratory queries in order to explore RDF data in a web browser. The interface has a map and timeline that allows a user to browse through the data and shows the query that is constructed in the bottom right corner. Gephi (https://gephi.org/) is a desktop application that can be used to visualize both ontologies and data. The application was primarily developed for social network- and biological network analysis which are also represented in graphs. The SemanticWebImport plugin however makes it possible to visualize RDF.

Depending on the use case and the user's familiarity with RDF, OWL and SPARQL they can improve a user's understanding of the data and thereby improve the reusability.

4.4.3 Query the (meta)data using SPARQL

For this method to be successful, a basic understanding of SPARQL is required. Without the correct query, a triplestore will not provide any meaningful result. The basic structure of a SPARQL query is provided in section 2.6.2. There is a wide variety of resources available to learn SPARQL. *Learning SPARQL* by Bob DuCharme (2013) deserves a commendation since without it this work would not have been possible. In most cases the subjects the subjects and objects will originate from the existing data. The SPARQL queries will mostly revolve around usages of the defined predicates.

The following chapter provides three use cases were this method was applied.

5 Proof of concept

This chapter takes the reuse framework from the previous chapter and demonstrates it for the chosen VGI projects, by implementing the use cases from Chapter 3. These proofs of concept will show the practical implementation of reusing the (meta)data from two heterogeneous datasources by using Semantic Web technologies. The reusability was attempted on two levels for each use case namely, the metadata- and the data level.

5.1 Use case 1: TrashHunters & MORA

The first use case consists of the (meta)data from TrashHunters and MORA. Both are VGI projects that are related to trash in the public space. TrashHunters has a spatial coverage of the whole of the Netherlands, whereas MORA is only used in Amsterdam.

The metadata of TrashHunters and MORA was structured by using the DCAT and VGI ontologies. The metadata is derived from the websites of the VGI projects and the website FlickR where the data of TrashHunters is hosted. The data on the FlickR website is licensed. The data of MORA is available via a download service and is available for reuse because it consists of public data by the municipality of Amsterdam. The collection purpose of MORA is to support the municipal trash services whereas the purpose of TrashHunters is to create public awareness about littering and making the companies responsible for the pollution.

5.1.1 Metadata level

At the metadata level, several reuse scenarios were considered. The metadata level provides insight into the datatypes that are used by a project. If a reuser is well versed in the data handling of XML or JSON for example, projects that use these formats may have their preference. Another scenario that was considered used the spatial coverage of the projects.

The final reuse scenario for the presented query is a researcher who is interested in the littering of trash in Amsterdam. From the VGI projects he or she is interested in understanding why the data was collected, who the organizations behind the projects are, how he can access the data and what the licensing status is. Furthermore, an up to date estimation of the size of the population of Amsterdam is required. All these questions can be answered by using the query in Listing 1.

Listing 1. Use case 1: TrashHunters & MORA metadata query.

```
prefix vgi: <http://www.vgiprojects.com/>
prefix dct: <http://www.w3.org/ns/dcat#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix dbo: <http://dbpedia.org/ontology/>

SELECT ?label ?purpose ?organizations ?URL ?requirements ?license ?population

WHERE
{

 ?s rdfs:label ?label .
 ?s vgi:collectionPurpose ?purpose .
 ?s vgi:Organizations ?organizations .
 ?s dct:Distribution ?download .
 ?download dct:accessURL ?URL .
 ?download vgi:requirements ?requirements .
 ?download dct:license ?license .

    SERVICE <http://DBpedia.org/sparql>
    { <http://dbpedia.org/resource/Amsterdam> dbo:populationTotal ?population . }

}
```

The query first finds the subjects related to `rdfs:label`, `vgi:collectionPurpose`, `vgi:Organizations` and `dct:Distribution`. The `dct:Distribution` object variable contains additional information which can be accessed by using it in the subject position. By requesting the `dct:accessURL`, `vgi:requirements` and `dct:license` from the `?download` variable this additional information is retrieved. The `SERVICE` function retrieves the population count from the DBPedia resource.

label	purpose	organizations	URL	requirements	license	population
"MORA"	"The data is used for municipal trash services. "	"Municipality of Amsterdam"	https://kaart.amsterdam.nl/datasets/datasets-item/t/mor-afval-1/export/rss	"None"	"Open Data. "	"834119" ^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger>
"TrashHunters"	"The data is collected for a public awareness campaign by volunteers."	"Plastic Soup Foundation"	https://www.flickr.com/services/api/flickr.photos.getExif.html	"FlickR API key & Photo ID's"	"All rights reserved."	"834119" ^^<http://www.w3.org/2001/XMLSchema#nonNegativeInteger>

Figure 4. Results of the query in Listing 1, displaying metadata from MORA, TrashHunters and Amsterdam.

The result in Figure 4 provides the researcher with an overview of both VGI projects. The most important finding in this case is the licensing information. The data of TrashHunters is not licensed under an open-date license which prohibits further reuse. The collection purpose, organizations and requirements will help the researcher determine if these data sources suit his/her purpose.

If this information was structured in RDF and disclosed by a SPARQL endpoint it would significantly reduce the time a researcher would have to spend on finding out this information.

5.1.2 Data level

The relevant extracted data from both projects is rather limited and constitutes the coordinates of contributions, along with a status indicator and a keyword designation. Therefore, additional information was added to the data to showcase the provenance ontology. A random name was added to submissions from both TrashHunters and MORA. This use case uses the PROV and FOAF ontologies. The FOAF profiles were added to demonstrate the abilities of the PROV ontology. Without the addition of this provenance information the scenario would remain limited to a summation of all the trash related incidents that are available.

By adding the provenance information, a researcher can be interested in knowing if volunteers have uploaded data to both VGI systems and if so, what that volunteers name is and when the submissions were made. This can be achieved by using the query in Listing 2.

Listing 2. Use case 1: TrashHunters & MORA data query.

```
prefix prov: <http://www.w3.org/ns/prov#>
prefix foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?submission ?volunteer ?name ?date

WHERE
{

 ?submission prov:wasGeneratedBy ?activity .
 ?activity prov:wasAssociatedWith ?volunteer .
 ?volunteer a foaf:Person .
 ?volunteer foaf:name ?name .
 ?activity prov:startedAtTime ?date .

}

ORDER BY ?name
```

Because of the structure of the PROV ontology (outlined in section 4.3.2) the `?submission` variable first needs to know which activities generated the submissions. The activities can then be associated with the volunteers with the help of the `prov:wasAssociatedWith` predicate. The volunteers that have a FOAF profile will have their name and corresponding activity displayed.

Count: 15

submission	volunteer	name	date
http://www.vgiprojects.com/TrashHunters/17235971463	http://www.vgiprojects.com/volunteerID1	"Bruce Wayne"	"2016-07-14T12:41:51" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/TrashHunters/17235983443	http://www.vgiprojects.com/volunteerID1	"Bruce Wayne"	"2016-07-14T12:41:51" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/TrashHunters/17235989323	http://www.vgiprojects.com/volunteerID1	"Bruce Wayne"	"2016-08-14T13:41:51" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/TrashHunters/17236001853	http://www.vgiprojects.com/volunteerID1	"Bruce Wayne"	"2015-07-16T06:45:45" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/MORA/0	http://www.vgiprojects.com/volunteerID1	"Bruce Wayne"	"29-09-2015" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/MORA/12	http://www.vgiprojects.com/volunteerID1	"Bruce Wayne"	"24-11-2015" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/MORA/1	http://www.vgiprojects.com/volunteerID1	"Bruce Wayne"	"20-01-2016" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/TrashHunters/17856904411	http://www.vgiprojects.com/volunteerID3	"Clark Kent"	"2015-08-11T11:31:51" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/TrashHunters/17273739914	http://www.vgiprojects.com/volunteerID3	"Clark Kent"	"2016-03-15T13:11:51" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/TrashHunters/17275768533	http://www.vgiprojects.com/volunteerID3	"Clark Kent"	"2014-07-15T14:41:51" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/TrashHunters/17668606168	http://www.vgiprojects.com/volunteerID3	"Clark Kent"	"2016-06-28T16:25:51" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/MORA/3	http://www.vgiprojects.com/volunteerID3	"Clark Kent"	"31-01-2016" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/TrashHunters/17856898281	http://www.vgiprojects.com/volunteerID2	"Peter Parker"	"2016-07-16T16:41:51" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/TrashHunters/17273623164	http://www.vgiprojects.com/volunteerID2	"Peter Parker"	"2017-01-14T16:41:51" ^^<http://www.w3.org/2001/XMLSchema#date>
http://www.vgiprojects.com/MORA/2	http://www.vgiprojects.com/volunteerID2	"Peter Parker"	"26-01-2016" ^^<http://www.w3.org/2001/XMLSchema#date>

Figure 5. Results of the query in Listing 2, displaying VGI contributions by volunteers and their submission date.

The `submission` column in Figure 5 shows the VGI system to which the submission was made. For the modelled dataset, there were three volunteers that have made submissions to both TrashHunters and MORA. The final column shows when these submissions were made.

Even though the FOAF profiles were added for demonstrative purposes this use case displays an important advantage of using RDF. The FOAF profiles can be defined in a separate triplestore, but by linking the volunteer identification numbers with the profiles, it provides an instant overview of the submissions that have been made.

5.2 Use case 2: hetweeractueel.nl & KNMI

The second use case regards climatological data from a VGI project, namely hetweeractueel.nl and The Royal Netherlands Meteorological Institute (KNMI) an authoritative data source. Both sources use weather stations to collect their data. The data in this use case is of 5th of May 2016. Both project have a spatial coverage of the Netherlands.

The data of hetweeractueel.nl is not available via a web API of download service therefore web scraping was applied. The Python library Pandas, in combination with the html5lib parser was used to extract the data from the website. The Python script is available in Appendix A. The data of the KNMI is available via a download service at https://www.knmi.nl/nederland-nu/klimatologie/daggegevens.

Hetweeractueel.nl consist of over 300 climate enthusiasts that collect and publish their data online. Most have their own website and hetweeractueel.nl functions as an aggregator of climatological data. The KNMI is the Dutch metrological institute of which the climatological service operates 35 weather stations. The weather stations operated by the KNMI are obviously more professional, with more sensitive sensors, better calibration, better positioning. Climate enthusiast generally cannot compete with the resources of a public institute. However, hetweeractueel.nl does have a higher spatial resolution than the KNMI.

5.2.1 Metadata level

The metadata of both projects was structured by using the DCAT, VGI and RDFS ontology. The use case on the metadata level consists of making a comparison between the available parameters of both projects. This overview provides a researcher with the opportunity to decide which source of data has the more relevant information. To make this assessment initially, the query was more elaborate. Ideally, the researcher would also prefer to know information about the equipment and calibration methods that are used. It can be expected that the KNMI has more advanced equipment than weather enthusiasts for example. The researcher might need to take this information into account when determining which data to use. Unfortunately, this information was not available for the weather stations of hetweeractueel.nl.

Listing 3. Use case 2: hetweeractueel.nl & KNMI metadata query.

```
prefix vgi: <http://www.vgiprojects.com/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>


SELECT ?project ?name ?parameters


WHERE
{

 ?project vgi:informationType ?informationType .
 ?project rdfs:label ?name .
 ?informationType vgi:availableParameters ?parameters .

}

ORDER BY ?parameters
```

The predicates `vgi:availableParameters` are modelled as an object of a blank node of the predicate `vgi:informationType`, therefore the information of the `?project` variable is queried first, followed by the name with `rdfs:label`. To gather the actual parameters the `?informationType` variable is used in the subject position.

project	name	parameters
http://www.vgiprojects.com/ID4	"KNMI - The Royal Netherlands Meteorological Institute"	"Air pressure (hPa)"
http://www.vgiprojects.com/ID3	"hetweeractueel.nl"	"Air pressure (hPa)"
http://www.vgiprojects.com/ID3	"hetweeractueel.nl"	"Precipitation (mm)"
http://www.vgiprojects.com/ID4	"KNMI - The Royal Netherlands Meteorological Institute"	"Relative humidity (%)"
http://www.vgiprojects.com/ID3	"hetweeractueel.nl"	"Station type"
http://www.vgiprojects.com/ID4	"KNMI - The Royal Netherlands Meteorological Institute"	"Temperature (°C)"
http://www.vgiprojects.com/ID3	"hetweeractueel.nl"	"Temperature (°C)"
http://www.vgiprojects.com/ID4	"KNMI - The Royal Netherlands Meteorological Institute"	"Visibility (m)"
http://www.vgiprojects.com/ID4	"KNMI - The Royal Netherlands Meteorological Institute"	"Weather description"
http://www.vgiprojects.com/ID4	"KNMI - The Royal Netherlands Meteorological Institute"	"Wind direction"
http://www.vgiprojects.com/ID3	"hetweeractueel.nl"	"Wind direction"
http://www.vgiprojects.com/ID3	"hetweeractueel.nl"	"Wind power (bft)"
http://www.vgiprojects.com/ID3	"hetweeractueel.nl"	"Wind speed (km/u)"
http://www.vgiprojects.com/ID4	"KNMI - The Royal Netherlands Meteorological Institute"	"Wind speed (m/s)"

Figure 6. Results of the query in Listing 3, displaying the available parameters and units of measurement for hetweeractueel.nl and KNMI.

Ordering the outcome by the `?parameters` variable makes it clear what the different parameters of each system are (see Figure 6). Overall, both systems provide similar data, in similar units, such as temperature in Celsius degrees, air pressure in hectopascals and wind direction. Both measure wind speed, however in different units. The KNMI also provides the relative humidity, visibility indicators and a weather description. Hetweeractueel.nl also provides unique data such as precipitation, the station type and wind power on the Beaufort scale.

These parameters are quite easily found on the websites of the respective systems. Extracting, transforming and loading the metadata in this instance might not be the most time effective method.

5.2.2 Data level

On the data level, the use case consists of data about the city of Hoek van Holland. This city was chosen for this use case because it is one of the few cities that occurs in both datasets. Initially the goal was to gather the coordinates of the weather stations and although technically this would not be difficult, the information was not provided by either system. The nearest estimation of the position of the weather stations is the place name.

The scenario that was finally opted for entails a researcher interested in climatological information about Hoek van Holland. Furthermore, the researcher is interested in general information about Hoek van Holland. The query in Listing 4 collects this data from the triplestore and DBPedia.

Listing 4. Use case 2: hetweeractueel.nl & KNMI data query.

```
prefix vgi: <http://www.vgiprojects.com/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>


SELECT *

WHERE
{

?station vgi:station ?place .
FILTER (regex(?place, "Hoek van", "i")) .
?station vgi:temperature ?temperature .
?station vgi:windspeed ?windspeed .
?station vgi:winddirection ?winddirection .
?station vgi:airpressure ?airpressure .
OPTIONAL { ?station vgi:rv ?humidity } .
OPTIONAL { ?station vgi:precipitation ?precipitation } .

    SERVICE <http://DBpedia.org/sparql>
    { <http://dbpedia.org/resource/Hook_of_Holland> rdfs:comment ?comment . }

FILTER (lang(?comment) = "en" )

}
```

The query uses the regular expression (`regex`) function of SPARQL to search within strings. By binding the `?place` variable to the weather stations it's possible to select the relevant parameters: temperature, wind speed and direction and air pressure. Because the relative humidity and precipitation only occur in one of the datasets the `OPTIONAL` function is used for those parameters. The `SERVICE` function in SPARQL retrieves additional information about Hoek van Holland from its DBPedia page.

station	place	temperature	windspeed	winddirection	airpressure	humidity	precipitation
http://www.vgiprojects.com/knmi/25	"Hoek van Holland"	"20.5 °C"	"6.0 m/s"	"ZO"	"1013.1 hPa"	"46.0"	
http://www.vgiprojects.com/hetweeractueel/145	"Hoek van Holland (oost)"	"18.9 °C"	"14.4 km/u"	"OZO"	"1012.8 hPa"		"0.0 mm"
http://www.vgiprojects.com/hetweeractueel/144	"Hoek van Holland"	"20.4 °C"	"18.9 km/u"	"ZO"	"1012.9 hPa"		"0.0 mm"

comment
"The Hook of Holland (Dutch: Hoek van Holland) is a town in the southwestern corner of Holland, at the mouth of the New Waterway shipping canal into the North Sea. The town is administered by the municipality of Rotterdam as a district of that city. Its district covers an area of 16.7 km2 (of which 13.92 km2 is land). On 1 January 1999 it had an estimated population of 9,400.On the north side of the New Waterway, to the west of the town, is a pier part of which is accessible to pedestrians and cyclists." @en

Figure 7. Results of the query in Listing 4, displaying the data for Hoek van Holland on May 5th 2016 (edited for legibility)

This provides the result in Figure 7. It shows three measurements stations and their respective data. KNMI does not provide data regarding precipitation and hetweeractueel.nl does not have data on the relative humidity, therefore those respective fields remain empty. The information from DBPedia is posted for each station, SPARQL currently does not have a method to reduce this redundancy (the screenshot above was edited for legibility and does not show this).

This use case displays clear advantages to the usage of RDF and SPARQL. Although it would have been possible to gather this information from both systems separately, RDF and SPARQL simplify this process by standardizing the information with the help of the VGI ontology.

5.3 Use case 3: AiREAS & enviroCar

The third use case concerns data about the air pollution. This use case combines the metadata from the VGI projects AiREAS and enviroCar. Both projects have a different approach when it comes to data collection. AiREAS uses stationary airboxes that sample the air at ten minute intervals. enviroCar collects data of the emissions of cars. The enviroCar community register their movement and the parameters of their car along that track. AiREAS is a citizens' initiative that strive to improve the air quality. The data of enviroCar needs to be aggregated and processed in order to be used.

AiREAS data is available via a download service at http://data.aireas.com/csv/. Whereas enviroCar's data is available through a web API at https://envirocar.org/api/stable/. The spatial coverage of AiREAS is only in the city of Eindhoven. The coverage of enviroCar is primarily clustered in the city of Münster, Germany where the project originated. Data with enviroCar was collected in Eindhoven on the 25th of June 2016 to create a spatial overlap between the datasets. Neither project has any information regarding the licensing of their data on their respective websites.

5.3.1 Metadata level

The metadata of both projects was modelled by using the DQV, DCAT VGI and RDFS ontologies. The data quality information was added for demonstrative purpose to exemplify the usage of data quality characteristics in RDF and to explore the structure of the data quality ontology. Because these two projects have a wide variety of public and private stakeholders the scenario for a stakeholder analysis was also considered. And even though stakeholders may have some influence on the reusability of data, data quality was considered to be more relevant to VGI.

The use case requires a researcher that is searching for information about air pollution and want to know in which units of measurements the data are gathered and what the user generated quality assessments are.

Listing 5. Use case 3: AiREAS & enviroCar metadata query.

```
prefix vgi: <http://www.vgiprojects.com/>
prefix dct: <http://www.w3.org/ns/dcat#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix dqv: <http://www.w3.org/ns/dqv#>

SELECT ?name ?units ?download ?qualityvalue ?usercomment


WHERE
{
?s rdfs:label ?name .
?s vgi:informationType ?info .
?info vgi:unitsOfMeasurement ?units .
?s dct:distribution ?download .
?download dqv:value ?qualityvalue .
?download dqv:UserQualityFeedback ?usercomment .

}
```

The query first retrieves the names of the projects. Followed by the units of measurements, for which it first has to query the `vgi:informationType`. The data quality values (`dqv:value` and `dqv:UserQualityFeedback`) are part of the `dct:distribution` part of the graph, which is why the `?download` variable is used in the subject position for these values.

name	units	download
"AiREAS"	"NO2"	http://www.vgiprojects.com/download1
"AiREAS"	"Ozon"	http://www.vgiprojects.com/download1
"AiREAS"	"Ultra Fine Particles"	http://www.vgiprojects.com/download1
"AiREAS"	"Parts per million (PM1)"	http://www.vgiprojects.com/download1
"AiREAS"	"Parts per 2.5 million (PM2.5)"	http://www.vgiprojects.com/download1
"AiREAS"	"Parts per 10 million (PM10)"	http://www.vgiprojects.com/download1
"enviroCar"	"CO2 (kg/h)"	http://www.vgiprojects.com/api1

qualityvalue	usercomment
"85%"	" Project has a wide variety of sensors. There 35 Airboxes in the area of Eindhoven, the Netherlands. The data is available via downloadservice."
"85%"	" Project has a wide variety of sensors. There 35 Airboxes in the area of Eindhoven, the Netherlands. The data is available via downloadservice."
"85%"	" Project has a wide variety of sensors. There 35 Airboxes in the area of Eindhoven, the Netherlands. The data is available via downloadservice."
"85%"	" Project has a wide variety of sensors. There 35 Airboxes in the area of Eindhoven, the Netherlands. The data is available via downloadservice."
"85%"	" Project has a wide variety of sensors. There 35 Airboxes in the area of Eindhoven, the Netherlands. The data is available via downloadservice."
"85%"	" Project has a wide variety of sensors. There 35 Airboxes in the area of Eindhoven, the Netherlands. The data is available via downloadservice."
"75%"	"Data is primarily focussed in Germany and the units of measurements require further computation."

Figure 8. Results of the query in Listing 5, displaying the units of measurements and data quality indicators for AiREAS and enviroCar (edited for legibility).

Figure 8 shows the measurement units of the two projects. While AiREAS uses conventional air quality assessment methods, enviroCar uses CO_2 emissions in kilograms per hours as a unit. This unit is derived from the revolutions per minute of the car's engine. Even though the data quality information is demonstrative it provides an insight into the possibility of using user generated statistics and comments to supplement the reusers experience.

5.3.2 Data level

As can be seen in the retrieval of the metadata in Figure 8, there is a significant difference between the measurement units of AiREAS and enviroCar. AiREAS uses conventional parts per million units to discern what the relative air quality is. An airbox takes a direct sample measurement from the air, analyses it and reports the outcome. enviroCar measures the CO_2 emissions from a car (in kg/h) which needs to be aggregated with other measurements in the same area and processed to develop a relative cold- and hotspot analysis of high- and low concentrations of CO_2 from cars (EnviroCar, 2017). There is no conversion method between either the data collection methods or the units they result in. This results in a semantic gap, even though both projects are concerned with data regarding air quality, their data is not compatible.

Several attempts were made to combine the data from both projects including using GeoSPARQL to query data in the Eindhoven area. However, because of the semantic gap between both data sources it is too difficult and unrealistic to develop a relevant use case.

6 Discussion, conclusions and recommendations

This chapter provides a discussion of the results of this thesis. It starts by discussing the results from Chapter 4 and 5. Section 6.3 provides the conclusions that can be made after this thesis. Followed by recommendations and further research areas in sections 6.4 and 6.5.

6.1 Discussion of proposed method

The proposed method to improve the reusability of data from VGI systems entails four main steps: gather metadata, gather data, model the (meta)data in RDF and query the (meta)data with SPARQL. At first glance, the steps appear to be rather straightforward however there are some difficulties involved. This section will evaluate the proposed method.

The first step does not have a steep learning curve. The five intermediate steps are clear should not pose any problems for researchers that are interested in data from VGI systems.

The second step is more difficult. The three techniques mentioned throughout this thesis, web APIs, download services and web scraping, each have their own learning curve, depending on the data source and the researcher's skills and experiences. Web APIs require some knowledge about HTTP requests and understand the nature of the formats in which responses are given. Download services function simple, but this simplicity has a downside. Data from download services often requires a lot of processing before the data is reusable. Web scraping can have a substantial learning curve. It depends heavily on the website that retains the data. This research benefitted from having a website that posted its data in relatively simple HTML tables on one page. There are plenty of tools of to scrape all sorts of websites (see section 4.2.1), each with their own documentation. For computer scientists, this would not pose a significant problem, geo-information professional or researcher might have to invest some time in order to get the data they want.

The third and fourth step requires a basic understanding of Semantic Web technologies. The third step requires a reuser to understand the RDF and, if a domain ontology is needed, some ontology engineering skills are needed. Furthermore, when using existing ontologies, such as DCAT and PROV, their structure is important to understand in order to be in accordance with other resources on the Semantic Web. Besides the Semantic Web community these technologies and ontologies are quite obscure, so understanding them will cost some time. This thesis used rdfEditor to develop a domain ontology from the ground up and Google Refine with DERI's RDF Refine extension to model the extracted data.

The fourth step requires an understanding of SPARQL to query the data. For those familiar with other query languages, such as Structured Query Language (SQL), adjusting to SPARQL should not be too difficult. SPARQL has some distinctive features and operators, but there are many tutorials on the web to explain them. If SPARQL is your first query language it may require some trial and error attempts but once you understand the logic behind RDF triples, SPARQL is not so difficult.

Some of the other existing ontologies that were considered for modelling the metadata and the data were: DataCube, GeoNames, SWEET and the Semantic Sensor Network (SSN). These ontologies were either found to be too detailed or too general for the use case purposes.

The same inverse relation between usability and reusability that appears in ontologies, also appears to be true for the reusability of VGI. The more specific the data that is collected, the less applicable it is for other purposes. The more generic the data the easier it is to reuse, when the complexity increases it becomes more complex to reuse it for secondary purposes. Simple features and measurements are more easily reusable than data that needs additional equipment for collection.

The proposed method relies on an ETL process. Ideally, the first three steps should not be necessary and reusers can gather both the metadata and data from remote SPARQL endpoints with federated queries. This relies on the VGI projects to be willing to share their (meta)data and structure it in accordance with existing ontologies.

6.2 Discussion of the proof of concept results

The proof of concept that is developed for this thesis was based on use case scenarios. The VGI projects that were selected came from the environmental domain. These scenarios are in the fields of trash, the climate and air quality. This section discusses the results of the proof of concept that was developed in Chapter 5.

6.2.1 General discussion on proof of concept

The data access mechanisms of the sampled VGI systems were fairly simplistic. Reuse by secondary parties was not a primary objective of any of these projects. As opposed to a project like OSM for example where reuse by secondary parties is the main purpose of the project. More research is needed into the motivations of VGI projects that publish their data.

The most time-consuming part of developing the proof of concept was understanding the variety of methods and tools that were available. Standardization is one way in which data from VGI projects can become more reusable. The first method to share data is by exposing an API endpoint with corresponding documentation that allows reusers to query the data dynamically.

However, a problem is that VGI system developers may not realise that the application that they have developed belongs to the academic notion of VGI. VGI projects are generally small-scale projects with limited resources. Modelling their data in accordance with an existing ontologies or developing an API endpoint with safeguards in place, may consume too many of their resources.

To alleviate the stress on resources it is recommended that VGI projects, that want their historic data to be available, do so via a download service. This places a higher burden on the reuser in terms of processing the data but less so on the VGI projects. Since historic data is less dynamic the advantages of an API are negated.

A third recommendation to share their data, is the use of Semantic Web technologies. If a VGI project has the resources and knowledge available to develop a SPARQL endpoint and model their data in accordance with ontologies, using Semantic Web technologies has the benefits of an API and the data can easily be linked with other semantically enriched sources of data.

6.2.2 Use case 1: Trash

The metadata from TrashHunters and MORA was initially collected from their respective websites. The information was available in an unstructured fashion, the semantic enrichment structures the metadata and if the triplestore was populated with metadata from additional VGI

projects that regard trash, it would save a researcher time. Furthermore, the inferencing capabilities of the Semantic Web would allow researchers to find related VGI projects once they are familiar with the subjects and predicates of the VGI ontology.

On the data level, provenance information was added to the data to demonstrate the benefits of having this information in RDF. By applying identical volunteer identification numbers between two VGI projects it was demonstrated that it is possible to retrieve the names and contribution dates of multiple projects at the same time, with one query. The structure of the PROV ontology, which separates entities, agents and events creates a structure in the data that was not there before.

However, there can be various reasons why VGI projects do not want to provide this provenance information. One reason is to maintain the privacy of contributors. Agents would need to consent to their personal information being available on the web and this might form an entry barrier.

6.2.3 Use case 2: Weather

The weather metadata use case showcases the scaling and units conflicts that causes heterogeneity of VGI projects (Goh, 1997). The KNMI measures windspeed in meters per seconds whereas hetweeractueel.nl standardizes wind speed in kilometres per hours. In this case, the conflict can be resolved relatively easily. Of the three use cases, this use case shows the reusability of both metadata and data the best. Besides the windspeed unit conflict, the data is interchangeable between the two systems and the metadata scenario provides multiple stations within the same area.

However, even though the modelled data is interchangeable between the systems, remarks should be made about the measurements themselves. The equipment and calibration methods of the KNMI are more advanced than those of most weather enthusiasts. The measurements of the KNMI are therefor probably more accurate than those of hetweeractueel.nl. This refers back to the issues of data quality of VGI in section 2.2.1.

By structuring the data from both systems with the VGI ontology, the data is queryable over the internet, where previously it was only accessible in a HTML table or as CSV. Furthermore, if the hypothetical SPARQL endpoints of these project were published on their websites it would also improve the findability of this data. Once a secondary user understands SPARQL and the used ontologies, he or she can start to query, explore and export the data.

6.2.4 Use case 3: Air quality

The air quality use case is the most unsatisfactory. After collecting the data from both respective systems. It appeared that there is a semantic gap between the measurements from both projects (see section 5.3.2). The query used on the metadata level visualizes this semantic gap. The units and methods used by both projects are too different too combine. Where AiREAS uses an airbox to collect a direct air sample to determine the air quality, the data from enviroCar needs to be processed into a hotspot analysis to determine relative air quality. Furthermore, AiREAS measures air quality in part per million units, whereas enviroCar measures CO_2 emissions in kg/h. There is no numerical conversion method possible between these two units.

Several attempts were made to achieve meaningful results. This includes modelling the data into a GeoSPARQL appropriate format, however to properly model the data would be too time consuming for this thesis. The result would only show the technical abilities of GeoSPARQL, not result into any meaningful outcome.

6.3 Conclusions

The following sections presents the conclusions of this thesis. First an answer is provided to the main research objective of this thesis, followed by answers on the formulated research questions.

6.3.1 Main conclusions

The main objective of this thesis is ***to develop a method that improves the reusability of volunteered geographic information by using Semantic Web technologies in the domain of environment applications.***

The developed method is described in *Chapter 4*. Four steps delineate the process of extracting (meta)data from VGI systems, semantically enriching it with a new and existing ontologies and potentially reusing it for a secondary purpose. This method is applied to three use cases which are described in *Chapter 5*.

The general conclusions that can be drawn from this thesis is that Semantic Web technologies work well for the reuse of metadata and can also work well for the reuse of data from VGI systems. Reusing data depends on the requirements of the reusers and therefore if there is a large semantic gap between two data sources, they will not be compatible. Such a gap is more likely to occur at the data level than at the metadata level. The metadata level is primarily used for exploratory queries, whereas data might actually be used in further research or applications and therefor has stricter requirements from a reusers side. This semantic gap can occur because of the data collection and processing methods or the units that have been used. The former happened in the third use case, the latter in the second use case.

In the second use case, one system measured wind speed in km/h, the other in m/s. Even though the data is not directly compatible, a simple computation can convert one in to the other. This can be considered a small semantic gap. In the third use case, the difference in data collection and processing methods (see section 5.3.2) resulted in different units (ppm and kg/h) which were not convertible. This resulted in a semantic gap that was too large for a meaningful result by combining the data sources. Overall, when it concerns a simple conversion of units a semantic gap is bridgeable, but when the data collection and processing methods differ significantly, is becomes too difficult.

Although it was not encountered with the selected VGI systems. The same issue arises, to some degree, when there is room for human interpretation in data collection, for example in mapping applications like OSM. One person calls a feature a hill, another a mountain, this doesn't directly mean the data is not reusable but it does create space for further interpretations.

This thesis provides a domain ontology that can be reused for the modelling of other VGI systems. However, the ontology is not exhaustive and will require continuous upkeep. Furthermore, the more specific an ontology is, the less reusable it becomes for secondary purposes this is a characteristic that ontologies and data share with each other.

The main benefits of using Semantic Web technologies to disclose (meta)data, compared to conventional methods like a web portal are that, it does not require a third party that maintains the web portal. The (meta)data does not have to comply with the standards that are used by the web portal. The VGI project can choose to use existing ontologies to structure their data, but do not have to do so.

Even though there are methods and techniques that can be applied to enhance the data quality of VGI it remains a point of contention, also for the use cases in this thesis. Semantic Web technologies can provide a structure that assists in enhancing data quality but are not the final solution.

6.3.2 Answering research questions

Objective 1: To identify requirements for the reuse of VGI and application domains in which reuse is relevant.

1. *What are the elements of reusing VGI?*

   The elements of reusing VGI are identified in *Chapter 2*. The heterogeneity of VGI complicates the reusability of the data from VGI systems. Heterogeneity leads to interoperability and data quality issues. Heterogeneity in VGI systems occurs on multiple levels. In general, the syntactic heterogeneity of VGI systems can be dealt with computationally however, sematic heterogeneity is more difficult to deal with.

   Other elements of reusing VGI are, the findability and accessibility of VGI. These differ from traditional geo-information, which complicates reuse. This thesis focusses on three data access mechanisms, APIs, download services and web scraping to reuse data that is provided by VGI systems. The advantages and disadvantages are discussed in *section 2.4*.

   Provenance, metadata and licensing are elements that can help improve the reusability of data, however the research on these topics for VGI is limited.

   In general, a comprehensive theoretical framework work on reusability is currently not available. This research has provided a lot of the crucial elements when it comes to reusing VGI, however the exact interplay between all these variables is still unknown.

2. *Which VGI application domains are applicable for (re)use cases?*

   VGI applications are at hand in a wide variety of domains. *Section 3.1* mentions VGI applications in the domains of health monitoring, tourism, smart-city, mapping and the environment. In order to have a representative sample of VGI projects in a domain there need to be enough systems to investigate. Several criteria were formulated to narrow down the domains. As stated in *section 3.1* the final two domains that were considered were smart-city- and environmental applications. The number of available applications, the availability of their data and the fact that environmental applications are inherently spatial gave the advantage to environmental applications.

   This choice does not make other VGI domains less applicable in producing reusable data. The first use case in its essence consists of two notification applications regarding trash. Users notify and upload trash instances to an interested party, either for the municipal trash services or for a public campaign. These types of notifications can also be made in other domains, such as incident reporting for traffic accidents or emergency response.

   For other domains, it may be more difficult to produce reusable data. The level of privacy of the contributors will influence the reusability of the data. For example, people might be more reluctant to share their location continuously or other health-related information in health monitoring applications, or be reluctant to review amenities in a city with their name attached in a tourism application. Safeguarding privacy appeared to be less of an issue in the environmental domain, but it can influence the reusability of data.

A VGI project can deal with this by anonymize or aggregating the data before publishing it. It is up to the VGI project to determine how to balance this trade-off between privacy protection and publishing the most crude, reusable data possible. The more crude the data is, the more reusable it will be; however, the less likely people will be willing to contribute. On the other hand, data can also be anonymized or aggregated so much that is becomes unusable for secondary purposes.

3. *Which VGI systems are available within the application domain and what are their characteristics?*

   A number of examined VGI systems in the environmental domain are listed in *section 3.2*. An initial assessment was made to ensure that the projects were related to similar domains, otherwise a reuse scenario would not be likely. Characteristics of the VGI projects are provided in Table 2.

   Two of the three use cases worked out as expected. The data from both systems was compatible with each other in a realistic and meaningful way. For the third use case this was not the case. After gathering the data from AiREAS and enviroCar and examining their measurement methods there appeared to be a large semantic gap between the two projects. Although they are both concerned with air quality and there was data with spatial- and temporal overlap, there was no meaningful reuse case available.

   Objective 2: To develop a method for improved syntactic and semantic interoperability between VGI systems.

4. *Which methods and tools are available to improve syntactic and semantic interoperability?*

   Section 2.5 discusses several methods that can improve syntactic and semantic interoperability, namely, ETL, ontologies for data integration, semantic enrichment and the Semantic Geoweb. This research proposes a combination of ETL and semantic enrichment to improve the reusability of VGI. The data from the VGI projects was extracted, semantically enriched and loaded into a triplestore. The method is described in *Chapter 4*.

   When it comes to data that is available on the internet syntactic interoperability does not appear to resemble a large problem. Most of the data, especially from VGI projects, is published in non-proprietary formats (CSV, JSON, XML). Achieving semantic interoperability also encompasses taking into account the human understanding and meaning of data and although Semantic Web technologies can help, they are not a silver bullet.

5. *How can the data from the selected VGI systems be accessed and how is it structured?*

   Current methods of reusing data from VGI systems is discussed in *section 2.3*. This thesis focusses on three data access mechanisms, APIs, download services and web scraping to reuse data that is provided by VGI systems. Their respective advantages and disadvantages are discussed in *section 2.4*. For most researchers with a moderate understanding of computer technology, APIs provide by far the most benefits.

   Appendix C shows examples and excerpts of the sampled projects. The structure and data formats of the projects is varied. The currently favoured data formats for VGI projects are JSON, XML and CSV. The biggest advantage of these formats is that they are non-proprietary. Which means any one should be able to retrieve and open them.

6. *What is the best way to structure data from VGI systems for reuse?*

In general, the best way to access data depends on the skills of the reuser and the capabilities of the VGI project. However, from a technical point of view, from the three current reuse methods, APIs have the most benefits. They allow the querying of specific parameters and when used in an application the data is directly updated, whereas download services and web scraping always create local copies of data. However, they are the most difficult to implement from the VGI developers side.

VGI projects also have to take precautions when it comes to making their data available. They may not want to share all the personal information of their contributors to respect their privacy. This might mean processing their data before making it available. Too many requests to an API can hamper the functionality of the API. VGI projects need to protect themselves from people with malicious intents.

7. *How to implement the use of semantic descriptions of VGI?*

   This thesis has developed a VGI ontology that has been used to model the (meta)data from both VGI and authoritative data sources. The ontology functions as a domain ontology that clarifies metadata and structures data in a comprehensive way. The ontology is available online at: https://s3.eu-central-1.amazonaws.com/swarish-gima.com/index.html. In line with Semantic Web best practices, existing ontologies were used when applicable.

   Besides using a SPARQL endpoint to disclose their data, VGI projects can model their metadata in RDF and embed this metadata in HTML pages.

   Objective 3: To construct a VGI knowledge base as a proof of concept in support of VGI reuse.

8. *What visualization tools for RDF are available and how can they assist in creating a knowledge base for reusable VGI?*

   Section 4.4.2 discusses three tools that have been explored during this thesis. These applications all use graph interfaces to model ontologies and data. This appears to be a helpful feature to aid reusability, especially when a user's experience with Semantic Web technologies is limited. Visual methods (graphs) are in general more easily understood than textual methods (queries).

   A website that provides a SPARQL endpoint to several or one data sources could help inexperienced users with an initial (interactive) graphical representation of the data. A user, on the other hand, could use the applications in *section 4.4.2* to visualize their final results.

9. *What is the best method for the data from the VGI systems to be in disclosed in a knowledge base?*

   As mentioned in *section 4.4.1*, the preferred method for the data from VGI systems to be available for reuse is by employing Semantic Web technologies. VGI projects could provide a SPARQL endpoint and limited documentation on the used ontologies, subjects and predicates. This would entail the VGI projects structuring their data in accordance with those ontologies. These ontologies could be already existing once or specifically developed ones for more complex projects. This would resemble the structures that APIs currently have, providing an endpoint and documentation. The benefit of applying RDF is that it allows posting federated queries and there is no need to adhere to the standards that are applied by a third-party web portal.

A third-party web portal would be an alternative for using Semantic Web technologies. So far however, there is no centralized place on the web where (meta)data from VGI projects is being collected and shared. The development of such a portal could also encourage the adoption of Semantic Web technologies.

Objective 4: To recommend improvement of current standards and recommend reusage strategies to VGI application builders.

10. *Which improvements can be made to the existing metadata- and dissemination-/interface standards?*

    VGI does not adhere to standards. Many VGI projects are grassroots citizen science projects. In order to aid reusability, standards should emphasize the reuse more. This point was made in *section 2.2.3*, which argues for a shift from producer-centric metadata to user-centric metadata. The metadata modelled in *section 5.1.1* suffers from the same problem, although it provides accurate and relevant metadata about the projects it is still not clear at a glance which purposes this data could fulfil. In the case of Semantic Web technologies and VGI this would entail that the ontology should have classes that enable this, such as 'potential purpose' and 'possible aggregate use'. This however, requires more effort from the producers-side.

    No dissemination and interface standards were used in the creation of the final proof of concept.

11. *What reusage strategies can be employed by future VGI system developers and what are the consequences of certain decisions?*

    The advantages and disadvantages of the current reuse methods are discussed in *section 2.3*. Besides the current methods Semantic Web technologies should not be disregarded by VGI projects. Eventually, it will depend on the resources available and the reuse potential of the data that is created by the system.

    Section 6.2.1 mentions three strategies - developing an API endpoint, creating a download service or employ Semantic Web technologies - that VGI system developers can employ depending on the resources that they have available. Each method has its own advantages and disadvantages. The choice which one to choose becomes an interplay between the dynamics and nature of the data, the available resources and the necessity of sharing the data for reuse.

6.4 Recommendations

This thesis has resulted in various recommendations for both VGI system developers and standardization organizations. This section recalls the important recommendations from this thesis.

6.4.1 For VGI system developers

VGI system developers can adopt the existing VGI ontology to start modelling their metadata. RDF does not necessarily have to be disclosed via a SPARQL endpoint. It can also be embedded in existing HTML documents. Currently the disclosure of metadata of VGI projects is very limited. Providing context is essential for any type of secondary reuse. VGI system developers play a crucial role in creating this understanding for reuse.

Data quality of VGI will remain a contested issue for some time. Therefore, it is of importance to first of all be aware of the granular heterogeneity that a VGI project has. Try to clarify this process as much as possible with the use of provenance information about data. Be mindful of ethical issues, such as privacy, and be clear on for example, the type of equipment that has been used and the calibration methods that have been applied.

VGI projects often don't bother with providing information regarding licensing on their website. For data to be reusable, VGI system developers should be clear and explicit on the licensing of their data.

6.4.2 For standardization organizations

The W3C and the OGC have been making strides when it comes to developing standards for the web. Their current Citizen Science Domain Working Group charter (Open Geospatial Consortium, 2016) will definitely overlap with VGI projects. Even though a centralized web portal is not needed when using Semantic Web technologies, the OGC can start to popularize the methods and ontologies throughout the geospatial (VGI) community. The OGC has experimented and developed with Semantic Web technologies since 2005 (Open Geospatial Consortium, 2005), but clear guidelines, ontologies and best practises appear to be lacking.

Furthermore, in general to aid the reusability this thesis supports the notions of Comber et al. (2007) and Goodchild (2009) that in particular metadata standards should become more (re)user-centric, instead of the current producer-centric metadata models.

6.5 Further research

Even though this research has provided many insights and possibilities of the usage of Semantic Web technologies, more questions have also been raised. This section provides interesting research areas that can be explored by other researchers.

This research took for granted the fact that some VGI projects are willing to share their data with reusers. The exact motivations of why they do this is unknown. Further research into their motivations can also shed a light on the practicality of using Semantic Web technologies by VGI projects.

The domain ontology that is developed for this thesis is limited to only five VGI systems. In order to be more representative of the whole spectrum of VGI applications more applications, and their semantics need to be examined.

The academic literature regarding the reusability at this point appeared to be scattered across several subjects. Data interoperability, data quality, heterogeneity, semantics and provenance are some of the terms that relate to reusability, however comprehensive theoretical frameworks are lacking. Developing such a framework can assist future researchers in narrowing down the variables for reusability of data.

7 References

AiREAS. (2016). AiREAS. Retrieved May 25, 2017, from http://scapeler.com/SCAPE604/app/leaflet?city=Eindhoven&sensor=PM10&time=0

Algemene Rekenkamer. (2015). *Open Data Trend Report 2015*. Voorhout.

Antoniou, V., & Skopeliti, A. (2015). Measures and Indicators of VGI Quality: an Overview. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *II-3/W5*, 345–351. https://doi.org/10.5194/isprsannals-II-3-W5-345-2015

Bakillah, M., Liang, S., Zipf, A., & Arsanjani, J. (2013). Semantic Interoperability of Sensor Data with Volunteered Geographic Information: A Unified Model. *ISPRS International Journal of Geo-Information*, *2*(3), 766–796. https://doi.org/10.3390/ijgi2030766

Berners-Lee, T. (2006). Design Issues. Retrieved June 6, 2016, from https://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T. (2010). Berners-Lee W3C Linked Data page. Retrieved May 18, 2017, from https://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34–43. https://doi.org/10.1038/scientificamerican0501-34

Bishr, Y. (1998). Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science*, *12*(October 2014), 299–314. https://doi.org/10.1080/136588198241806

Bizer, C., Heath, T., & Berners-Lee, T. (2012). Linked Data - The Story So Far. Retrieved January 21, 2016, from http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

Brodeur, J. (2012). Geosemantic Interoperability and the Geospatial Semantic Web. *Springer Handbook of Geographic Information SE - 15*, 291–310. https://doi.org/10.1007/978-3-540-72680-7_15

Bryman, A. (2008). *Social Research Methods* (3rd ed.). Oxford University Press.

Butenuth, M., Gösseln, G. v., Tiedge, M., Heipke, C., Lipeck, U., & Sester, M. (2007). Integration of heterogeneous geospatial data in a federated database. *ISPRS Journal of Photogrammetry and Remote Sensing*, *62*(5), 328–346. https://doi.org/10.1016/j.isprsjprs.2007.04.003

Carrara, W., Chan, W. S., Fischer, S., & Steenbergen, E. van. (2015). *Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources*. https://doi.org/10.2759/328101

Cockburn, A. (2001). Writing Effective Use Cases, 246. https://doi.org/10.1145/505894.505918

Comber, A. J., Fisher, P. F., & Wadsworth, R. A. (2007). User-focused metadata for spatial data, geographical information and data quality assessments. In *10th AGILE International Conference on Geographic Information Science* (pp. 1–13). Aalborg, Denmark. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.7278&rep=rep1&type=pdf

Craglia, M., Goodchild, M. F., Annoni, A., Camara, G., Gould, M., Kuhn, W., … Parsons, E. (2008). Editorial: Next-Generation Digital Earth. *International Journal of Spatial Data Infrastructures Research*, *3*, 146–167. https://doi.org/10.2902/1725-0463.2008.03.art9

Di, L., Yue, P., Ramapriyan, H. K., & King, R. L. (2013). Geoscience Data Provenance: An Overview. *IEEE*

Transactions on Geoscience and Remote Sensing, *51*(11), 5065–5072. https://doi.org/10.1109/TGRS.2013.2242478

Domingue, J., Fensel, D., & Hendler, J. (2011). *Handbook of Semantic Web Technologies*. (J. Domingue, D. Fensel, & J. Hendler, Eds.), *Handbook of Semantic Web Technologies* (Vol. 1). Berlin, Heidelberg: Springer. https://doi.org/10.1017/CBO9781107415324.004

DuCharme, B. (2013). *Learning SPARQL*. (S. St. Laurent & M. Blanchette, Eds.) (2nd ed.). Sebastopol: O'Reilly Media Inc.

Egenhofer, M. J. (2002). Toward the semantic geospatial web. *Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems - GIS02*, 1–4. https://doi.org/10.1145/585147.585148

Elwood, S. (2008). Volunteered geographic information: Future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal*, *72*(3–4), 173–183. https://doi.org/10.1007/s10708-008-9186-0

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2013). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, *102*(3), 571–590. https://doi.org/10.1080/00045608.2011.595657

enviroCar. (2016). About Us. Retrieved March 30, 2016, from https://envirocar.org/about.php

EnviroCar. (2017). enviroCar. Retrieved June 4, 2017, from http://www.arcgis.com/home/webmap/viewer.html?webmap=5db4e1ea445e4b4b8612443e7ba76119

European Communities. (2004). European Interoperability Framework for pan-European eGovernment Services. *European Commission*, *Version 1*, 1–25. https://doi.org/10.1109/HICSS.2007.68

Feick, R., & Roche, S. (2013). Understanding the Value of VGI. In *2nd Workshop on Value of Geoinformation* (pp. 1–15).

Feigenbaum, L., & Prud'hommeaux, E. (2013). Cambridge Semantics. Retrieved May 10, 2017, from https://www.cambridgesemantics.com/semantic-university/sparql-by-example

Fonseca, F., Egenhofer, M., Agouris, P., & Camara, G. (2002). Using ontologies for integrated geographic information systems. *Transactions in GIS*, *6*(3), 231–257. https://doi.org/citeulike-article-id:384456

Fonte, C. C., Bastin, L., See, L., Foody, G., & Estima, J. (2015). Good Practice Guidelines for Assessing VGI Data Quality. In *AGILE 2015* (pp. 2–5). Lisbon.

Frew, J. (2007). Provenance and Volunteered Geographic Information. Retrieved May 15, 2017, from http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Frew_paper.pdf

Gliozzo, A., Biran, O., Patwardhan, S., & McKeown, K. (2013). Semantic Technologies in IBM Watson. In *Proceedings of the Fourth Workshop on Teaching Natural Language Processing* (pp. 85–92). Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W/W13/W13-34.pdf#page=95

Goh, C. H. (1997). Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Systems. *Technology*. https://doi.org/dl.acm.org/citation.cfm?id=925146

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Goodchild, M. F. (2009). Putting Research Into Practice. In A. Stein, W. Shi, & W. Bijker (Eds.), *Quality Aspects in Spatial Data Mining*.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. https://doi.org/10.1016/j.spasta.2012.03.002

Google. (2017). Google Maps. Retrieved May 9, 2017, from https://developers.google.com/maps/showcase/

Gore, A. (1998). The Digital Earth: Understanding our planet in the 21st Century. *Australian Surveyor*, *43*(2), 89–91. https://doi.org/10.1080/00050326.1998.10441850

Griffin, G. P., & Jiao, J. (2015). Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *Journal of Transport and Health*, *2*(2), 238–247. https://doi.org/10.1016/j.jth.2014.12.001

Grimes, S. (2014). Semantic Web Business: Going Nowhere Slowly. Retrieved June 4, 2017, from http://www.informationweek.com/software/information-management/semantic-web-business-going-nowhere-slowly/d/d-id/1113323

Harvey, F., Kuhn, W., Pundt, H., Bishr, Y., & Riedemann, C. (1999). Semantic interoperability: A central issue for sharing geographic information. *Annals of Regional Science*, *33*(2), 213–232. https://doi.org/10.1007/s001680050102

Hauthal, E., & Burghardt, D. (2016). Using VGI for analyzing activities and emotions of locals and tourists. In *AGILE 2016*. Helsinki, Finland.

He, L., Yue, P., Di, L., Zhang, M., & Hu, L. (2015). Adding Geospatial Data Provenance into SDI—A Service-Oriented Approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(2), 926–936. https://doi.org/10.1109/JSTARS.2014.2340737

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space* (1st ed.). Morgan & Claypool. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Heinzelman, J., & Waters, C. (2011). *Crowdsourcing Crisis Information in Disaster-Affected Haiti*. Washington D.C. Retrieved from http://scholar.google.com/scholar? q=related:0Xm_BoEh54wJ: scholar.google.com/&hl=en&num=20&as_sdt=0,5

Hoekstra, R., & Groth, P. (2015). PROV-O-Viz - Understanding the Role of Activies in Provenance. In B. Ludascher & B. Plale (Eds.), *5th International Provenance and Annotation Workshop* (pp. 215–220). Cologne: Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-540-89965-5

Holloway, T., Bozicevic, M., & Börner, K. (2007). Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity*, *12*(3), 30–40. https://doi.org/10.1002/cplx.20164

Huang, C.-Y., & Chang, H. (2016). GeoWeb Crawler: An Extensible and Scalable Web Crawling Framework for Discovering Geospatial Web Resources. *Isprs International Journal of Geo-Information*, *5*(8). https://doi.org/10.3390/ijgi5080136

International Telecommunication Union. (2015). ICT Facts & Figures. The world in 2015. *ITU 150 Años (1865 - 2015)*, 6. Retrieved from http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf

Jacobson, I., Spence, I., & Bittner, K. (2011). *Use-Case 2.0 The Guide To Succeeding with Use Cases*. https://doi.org/10.1145/2890778

Janowicz, K., & Hitzler, P. (2012). The Digital Earth as Knowledge Engine. *Semantic Web*, *0*(1), 1–10.

Janowicz, K., Scheider, S., Pehle, T., & Hart, G. (2012). Geospatial semantics and linked spatiotemporal data-past, present, and future. *Semantic Web*, *3*(4), 321–332. https://doi.org/10.3233/SW-2012-0077

Kamel Boulos, M. N., Resch, B., Crowley, D. N., Breslin, J. G., Sohn, G., Burtner, R., … Chuang, K.-Y. (2011). Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. *International Journal of Health Geographics*, *10*(1), 67. https://doi.org/10.1186/1476-072X-10-67

Kolas, D., Emmons, I., & Dean, M. (2009). Efficient Linked-List RDF Indexing in Parliament. *CEUR Workshop Proceedings*, *517*, 17–32. Retrieved from http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-517/ssws09-paper2.pdf

Kuhn, W. (2005). Geospatial Semantics: Why, of What, and How? *Journal on Data Semantics*, *III*, 1–24.

Lanter, D. P. (1991). Design of a Lineage-Based Meta-Data Base for GIS, *9844*(October). https://doi.org/10.1559/152304091783786718

Lemmens, R., Falquet, G., De Sabbata, S., Jiang, B., & Bucher, B. (2016). Querying VGI by semantic enrichment. In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, & R. Purves (Eds.), *European Handbook of Crowdsourced Geographic Information* (pp. 185–194). London: Ubiquity Press. https://doi.org/http://dx.doi.org/10.5334/bax.n

Lemmens, R., Falquet, G., & Métral, C. (2016). Towards Linked Data and ontology development for the semantic enrichment of volunteered geo-information. In *AGILE*. Helsinki, Finland. Retrieved from http://www.cs.nuim.ie/~pmooney/LinkVGI2016/Towards_Linked_Data_and_ontology_development_for_the_semantic_enrichment_of_volunteered_geo-information.pdf

Mitchell, R. (2015). *Web Scraping with Python: Collecting data from the modern web*. (S. St. Laurent & A. MacDonald, Eds.) (4th ed.). Sebastopol: O'Reilly Media Inc.

Mocnik, F.-B., Zipf, A., & Fan, H. (2017). The Inevitability of Calibration in VGI Quality Assessment. In *VGI-Analytics Workshop at AGILE 2017* (p. 4). Wageningen.

Morrison, A. (2015). The rise of immutable data stores. Retrieved June 4, 2017, from http://usblogs.pwc.com/emerging-technology/the-rise-of-immutable-data-stores/

Noy, N. F., & McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory*, 25. https://doi.org/10.1016/j.artmed.2004.01.014

Okolloh, O. (2009). Ushahidi, or "testimony": Web 2.0 tools for crowdsourcing crisis information. In *Change at Hand: Web 2.0 for Development* (pp. 65–69). International Institute for Environment and Development.

Olteanu-Raimond, A. M., Hart, G., Foody, G. M., Touya, G., Kellenberger, T., & Demetriou, D. (2016). The Scale of VGI in Map Production: A Perspective on European National Mapping Agencies. *Transactions in GIS*, (April). https://doi.org/10.1111/tgis.12189

Open Data Institute. (2017). Open Data Institute. Retrieved May 10, 2017, from https://theodi.org/guides/publishers-guide-open-data-licensing

Open Geospatial Consortium. (2005). OGC to Begin Geospatial Semantic Web Interoperability Experiment. Retrieved May 4, 2017, from http://www.opengeospatial.org/pressroom/pressreleases/420

Open Geospatial Consortium. (2016). OGC Citizen Science Domain Working Group. Retrieved May 17,

2017, from http://external.opengeospatial.org/twiki_public/CitizenScienceDWG/DraftCharter

OpenStreetMap Foundation. (2016). OpenStreetMap. Retrieved April 11, 2016, from https://www.openstreetmap.org/about

Pagano, P., Candela, L., & Castelli, D. (2013). Data interoperability. *Data Science Journal*, *12*(July), GRDI19-GRDI25. https://doi.org/10.2481/dsj.GRDI-004

Ramos, J. M., Vandevasteele, A., & Devillers, R. (2013). Semantic Integration of Authoritative and Volunteered Geographic Information (VGI) using Ontologies. In *AGILE 2013*.

Schlieder, C. (2010). Digital heritage: Semantic challenges of long-term preservation. *Semantic Web*, *1*(1–2), 143–147. https://doi.org/10.3233/SW-2010-0013

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., … Rutzinger, M. (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS International Journal of Geo-Information*, *5*(5), 55. https://doi.org/10.3390/ijgi5050055

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems*, *21*(3), 96–101. https://doi.org/10.1109/MIS.2006.62

Stichting Natuurinformatie. (2016). waarneming.nl. Retrieved April 11, 2016, from http://waarneming.nl/statistiek.php

Ushahidi. (2016). Ushahidi. Retrieved April 11, 2016, from https://www.ushahidi.com/features

van Harmelen, F. (2011). 10 Years of Semantic Web research: Searching for universal patterns. Retrieved March 15, 2017, from https://www.youtube.com/watch?v=13w_adm4zWg

Vassiliadis, P. (2009). A survey of Extract – transform – Load technology. *International Journal of Data Warehousing & Mining*, *5*(3), 1–27. Retrieved from http://bit.ly/15KE6p1

W3C. (1999). Resource Description Framework (RDF) Model and Syntax Specification. Retrieved July 4, 2016, from https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., & Hubner, S. (2001). Ontology-Based Information Integration: A Survey of Existing Approaches. *International Joint Conference on Artificial Intelligence; Workshop: Ontologies and Information Sharing*, 108–117. https://doi.org/citeulike-article-id:593500

8 Appendix A: Web scraping script

The web scraping scripts are written in the Python programming language with the use of the *pandas* library and the *html5lib* parser.

hetweeractueel.nl

```
    __author__ = 'Swarish Marapengopi'
# -*- coding: utf-8 -*-
import pandas as pd

pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)

df =
pd.read_html('http://www.hetweeractueel.nl/actueelweer/nederland',
index_col=0, flavor='html5lib', header=0)[3]

#print (type(df))
#print (len(df))
#for table in df:
#print(len(table))

df.columns = ['Station Type', 'Temperature','Precipitation','Wind
Speed','Wind Power','Wind Direction','Air Pressure'
              ,'Hide station']
df.index.name='Station Name'

df.drop(['Hide station'],inplace=True,axis=1)

print (df)

csv = df.to_csv('080516.csv')
```

9 Appendix B: Overview of RDF Ontologies, namespaces, publishers and use cases

prefix	namespace	name	publisher	use case
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	Resource Description Framework	W3C	All
rdfs	http://www.w3.org/2000/01/rdf-schema#	Resource Description Framework Schema	W3C	All
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema Definition	W3C	All
geo	http://www.w3.org/2003/01/geo/wgs84_pos#	Basic Geo (WGS84 lat/long) Vocabulary	W3C	Trash, Air quality
prov	http://www.w3.org/ns/prov#	Provenance Ontology	W3C	Trash
dct	http://purl.org/dc/terms/	Dublin Core	Dublin Core	All
dbr	http://dbpedia.org/resource/	DBPedia Resources	DBPedia	All
dbo	http://dbpedia.org/ontology/	DBPedia Ontology	DBPedia	All
vgi	http://www.vgiprojects.com/	VGI Ontology	Swarish Marapengopi	All

10 Appendix C: Data structures of sampled projects.

10.1 Use case 1: TrashHunters and MORA.

Data about photographs of TrashHunters is available via the FlickR Web API. The data was queried in JSON. Figure 9 shows an excerpt of data of the FlickR Web API.


```
{ "photo": { "id": "17233920844", "tag": "Unknown",
    "location": { "latitude": 52.368050, "longitude": 4.865241, "accuracy": 16, "context": 0,
      "locality": { "_content": "Amsterdam", "place_id": "xQ4tawtWULlNrOY", "woeid": "727232" },
      "county": { "_content": "Amsterdam", "place_id": "nmbnjNtQUL_iOTHdPg", "woeid": "12592040" },
      "region": { "_content": "North Holland", "place_id": "F86XYCBTUb6DPzhs", "woeid": "2346379" },
      "country": { "_content": "Netherlands", "place_id": "Exbw8apTUb6236fOVA", "woeid": "23424909" },
      "place_id": "xQ4tawtWULlNrOY", "woeid": "727232" } }, "stat": "ok" },
      "keyword": { "_content": "Red Bull" }
```


Figure 9. Excerpt from TrashHunters in JSON format (edited).

The data of MORA is available via a download service in CSV, GeoRSS and GeoJSON (https://kaart.amsterdam.nl/datasets/datasets-item/t/mor-afval-1). Figure 10 shows how the attributes of MORA are structured in JSON.


```
{
   "type":"FeatureCollection",
   "features":[
      {
         "type":"Feature",
         "geometry":{
            "type":"Point",
            "coordinates":[
               4.79343272958,
               52.3627275737
            ]
         },
         "properties":{
            "titel":"MORA-0510131",
            "titel_key":"mora_0510131_2",
            "datum":"2015-09-21 00:00:00",
            "adres":"Willemskerkestraat 7",
            "postcode":"",
            "omschrijving":"",
            "locatie":"POINT (4.79343272958 52.3627275737)",
            "website":"",
            "plaats":"",
            "email":"",
            "telefoonnummer":"",
            "date_created":"2017-04-24 09:06:04",
            "date_modified":"2017-04-24 09:06:39",
            "datum_eind":"0000-00-00 00:00:00",
            "type":"Openstaand",
            "Adres":"Willemskerkestraat 7",
            "Categorie_naam":"Afval",
            "Categorie_nummer":"9170",
            "Categorie_status":"Afval - In behandeling",
            "Datum":"21-09-2015",
            "Status":"In behandeling",
            "Subrubriek":"Puin \/ sloopafval",
            "id_2":"point_3437253",
            "opmerking":"Deze melding wordt \u00e9\u00e9n keer per dag om 09.00 uur bijgewerkt.",
            "title":"MORA-0510131",
         }
      },
```


Figure 10. Report made in MORA in JSON format.

10.2 Use case 2: hetweeractueel.nl & KNMI

The data of hetweeeractueel.nl was scraped with the script in appendix A. Figure 11 is the result of one scraping instance.

Station Name	Station Type	Temperature	Precipitation	Wind Speed	Wind Power	Wind Direction	Air Pressure
's-Gravenzande	S	20.2 °C	0.0 mm	14.5 km/u	3 bft	ZO	1012.9 hPa
't Zandt	S	19.1 °C	0.0 mm	14.4 km/u	3 bft	ZO	1015.9 hPa
Alkmaar (Oudorp)	S	19.6 °C	0.0 mm	0.0 km/u	0 bft	NNO	1013.7 hPa
Alkmaar (centrum)	S	21.3 °C	0.0 mm	9.7 km/u	2 bft	NO	1014.0 hPa
Alkmaar (noord)	S	21.6 °C	0.0 mm	11.3 km/u	2 bft	OZO	1012.9 hPa
Alkmaar (westerhoutkwartier)	S	19.7 °C	0.0 mm	16.4 km/u	3 bft	OZO	1013.7 hPa
Alkmaar west	S	19.8 °C	0.0 mm	5.4 km/u	1 bft	ONO	1013.8 hPa
Almelo	S	20.8 °C	0.0 mm	5.0 km/u	1 bft	ONO	1014.0 hPa
Almere Buiten	S	22.3 °C	0.0 mm	11.2 km/u	2 bft	ONO	1013.2 hPa
Almere-Stad	S	8.1 °C	0.0 mm	0.0 km/u	0 bft		1029.2 hPa
Almkerk	B	20.1 °C	0.0 mm	9.9 km/u	2 bft	ZZW	1013.6 hPa
Alphen aan den Rijn (Weteringpark)	S	18.9 °C	0.0 mm	14.8 km/u	3 bft	O	1011.6 hPa
Amersfoort (Kattenbroek)	S	20.4 °C	0.0 mm	4.7 km/u	1 bft	ZO	1013.8 hPa
Amersfoort Soesterkwartier	S	22.3 °C	0.0 mm	8.3 km/u	2 bft	N	1013.1 hPa
Amsterdam (Nieuwendam)	B	19.5 °C	0.0 mm	2.4 km/u	1 bft	ZO	1012.6 hPa
Amsterdam (Noord)	S	19.6 °C	0.0 mm	6.4 km/u	2 bft	ZZO	1013.3 hPa
Amsterdam (westerpark)	S	21.1 °C	0.0 mm	13.4 km/u	3 bft	ONO	1013.5 hPa
Amsterdam Holendrecht	S	20.3 °C	0.0 mm	8.4 km/u	2 bft	OZO	1013.7 hPa
Andelst	S	20.6 °C	0.0 mm	11.3 km/u	2 bft	O	1013.5 hPa
Apeldoorn (Zuidwest 2)	S	24.4 °C	0.0 mm	7.0 km/u	2 bft	ZO	1017.0 hPa
Apeldoorn (oost)	S	18.9 °C	0.0 mm	11.9 km/u	3 bft	ONO	1013.7 hPa
Apeldoorn (zuidwest)	S	20.1 °C	0.0 mm	10.8 km/u	2 bft	ZO	1013.4 hPa
Appelscha	S	21.9 °C	0.0 mm	12.1 km/u	3 bft	O	1015.3 hPa
Arnhem	S	20.2 °C	0.0 mm	7.9 km/u	2 bft	NO	1013.5 hPa
Arnhem (presikhaaf)	S	20.0 °C	0.0 mm	8.0 km/u	2 bft	OZO	1014.0 hPa
Assen (marsdijk)	S	21.2 °C	0.0 mm	4.3 km/u	1 bft	OZO	1015.5 hPa
Assen (zuidwest)	S	20.4 °C	0.0 mm	4.3 km/u	1 bft	OZO	1015.6 hPa
Asten	S	19.8 °C	0.0 mm	4.8 km/u	1 bft	O	1013.6 hPa
Baarland	B	18.8 °C	0.0 mm	3.6 km/u	1 bft	NNW	1011.5 hPa
Badhoevedorp	S	20.2 °C	0.0 mm	11.3 km/u	2 bft	ZZO	1014.0 hPa
Balkbrug	S	22.0 °C	0.0 mm	6.4 km/u	2 bft	O	1014.5 hPa
Bedum	S	20.4 °C	0.0 mm	7.9 km/u	2 bft	ZZW	1014.7 hPa
Beegden	B	20.8 °C	0.0 mm	3.4 km/u	1 bft	O	1013.6 hPa
Belfeld	S	20.2 °C	0.0 mm	6.4 km/u	2 bft	ZO	1013.4 hPa
Bergen (NH)	S	21.4 °C	0.0 mm	12.9 km/u	3 bft	OZO	1012.9 hPa
Berghem	S	20.8 °C	0.0 mm	12.7 km/u	3 bft	O	1013.3 hPa
Berkel en Rodenrijs	S	18.9 °C	0.2 mm	6.4 km/u	2 bft	ZO	1010.9 hPa

Figure 11. Result of web scraping hetweeractueel.nl on May 5th 2016.

The data of het KNMI was downloaded via a download service. The data was edited to ensure a temporal overlap with the data of hetweeractueel.nl. The data from various weather stations was collated. Figure 12 shows the result of data on the 5th of May 2016.

Station	Weer	Temp (°C)	RV (%)	Wind	Wind (m/s)	Zicht (m)	Druk (hPa)
Lauwersoog		19.1	55.0	OZO	6.0		
Nieuw Beerta		19.2	55.0	O	5.0		
Terschelling		16.1	78.0	O	5.0	15100.0	1015.4
Vlieland	onbewolkt	17.0	66.0	O	6.0	25200.0	1014.9
Leeuwarden	onbewolkt	19.4	56.0	O	5.0	30700.0	1015.3
Stavoren	licht bewolkt	19.0	60.0	OZO	5.0	22300.0	
Houtribdijk				O	5.0		
Eelde	onbewolkt	19.6	54.0	O	4.0	26200.0	1015.8
Hoogeveen	licht bewolkt	19.0	56.0	OZO	5.0	28400.0	1015.5
Heino		19.6	52.0	O	5.0		
Twente	licht bewolkt	19.2	55.0	ZO	5.0	27900.0	1015.4
Deelen	onbewolkt	18.7	50.0	OZO	5.0	35000.0	1014.6
Hupsel		18.7	59.0	O	5.0		
Herwijnen	onbewolkt	19.2	53.0	OZO	5.0		1013.9
Marknesse		19.3	53.0	O	6.0	29500.0	
Lelystad	licht bewolkt	19.3	56.0	OZO	5.0	35800.0	1014.5
De Bilt	licht bewolkt	19.1	51.0	OZO	6.0	27500.0	1013.8
Cabauw		18.4	60.0	OZO	5.0	27300.0	1013.8
Den Helder	onbewolkt	17.5	67.0	O	5.0	23600.0	1014.4
Berkhout		18.1	60.0	OZO	6.0	21100.0	
IJmuiden				O	7.0		
Wijk aan Zee		20.3	47.0				
Schiphol	onbewolkt	19.7	52.0	OZO	5.0	33100.0	1013.6
Voorschoten	onbewolkt	18.6	60.0	OZO	6.0	24200.0	1013.5
Rotterdam	onbewolkt	18.9	57.0	OZO	4.0	16500.0	1013.5
Hoek van Holland		20.5	46.0	ZO	6.0		1013.1
Vlissingen	licht bewolkt	16.3	63.0	OZO	6.0	18700.0	1012.5
Westdorpe		19.0	57.0	O	3.0	36400.0	1012.3
Woensdrecht	licht bewolkt	19.1	50.0	O	4.0	29600.0	1012.9
Gilze Rijen	licht bewolkt	19.7	51.0	OZO	4.0	29600.0	1013.5
Volkel	onbewolkt	19.3	53.0	OZO	5.0	20900.0	1014.2
Eindhoven	onbewolkt	19.2	51.0	OZO	3.0	31800.0	1013.9
Ell	licht bewolkt	18.7	56.0	O	3.0	29800.0	
Arcen		19.0	53.0	OZO	4.0		
Maastricht-Aachen Airport	licht bewolkt	20.2	52.0	O	3.0	35600.0	1013.3

Figure 12. Climatological data from the KNMI on May 5th 2016.

10.3 Use case 3: AiREAS & enviroCar

The data of AiREAS is available via a download service in CSV format. Figure 13 shows an excerpt of data on the 22nd of May 2016. The shown data has a temporal overlap with the data from enviroCar.

id	AmbHum	AmbTemp	lat	lon	NO2	Ozon	PM1	PM10	PM25	RelHum	Temp	UFP	type	measured
15	0	0	51.28150329	5.38085083	0	0	2	6	3	69.03	14.67	0	ten	22-5-2016 11:00
21	77.07	18.81	51.26647008	5.31340704	14.4	51.2	2	5	3	57.36	18.09	3920	ten	22-5-2016 11:00
24	0	0	51.26640226	5.26064107	0	23.4	3	8	4	66.74	16.16	0	ten	22-5-2016 11:00
38	80.72	18.87	51.35354183	4.44486631	0	0	4	8	5	77.74	15.29	0	ten	22-5-2016 11:00
31	79.92	18.54	51.23652284	5.27786155	3.7	23.5	2	5	3	63.15	16.61	0	ten	22-5-2016 11:00
18	78.74	19.35	51.35165952	4.46477971	0	0	6	10	7	54.26	18.03	0	ten	22-5-2016 11:00
17	0	0	51.28866095	5.25794301	0	48.1	3	5	3	70.07	15.55	0	ten	22-5-2016 11:00
9	0	0	51.28783528	5.28504929	0	32.4	2	4	2	79.72	15.39	0	ten	22-5-2016 11:00
7	80.73	17.87	51.26248697	5.27099183	23.6	28.7	2	5	3	60.98	17.77	0	ten	22-5-2016 11:00
27	0	0	51.25703797	5.29670989	0	25.6	2	6	3	59.51	19.41	0	ten	22-5-2016 11:00
20	0	0	51.26380775	5.29636774	0	21.5	2	4	2	66.72	16.03	0	ten	22-5-2016 11:00
10	77.69	19.17	51.35467193	4.46887843	0	0	6	10	8	55.75	18.71	0	ten	22-5-2016 11:00
2	0	0	51.26866988	5.27982109	0	33.7	3	6	3	65.09	17	0	ten	22-5-2016 11:00
3	76.52	18.63	51.26218134	5.28684143	19.1	46	3	6	3	68.62	15.48	0	ten	22-5-2016 11:00
8	73.13	19.57	51.25278014	5.29720348	4.1	50	2	6	3	61.33	17.2	0	ten	22-5-2016 11:00
19	78.3	18.88	51.29484309	5.2634997	1.5	56.2	2	4	2	71.48	15.19	0	ten	22-5-2016 11:00
36	0	0	51.27368202	5.29091555	16.5	41.3	2	5	2	61.2	17.47	4655	ten	22-5-2016 11:00
12	79.16	18.77	51.25256807	5.27703299	8.1	33.1	2	4	2	61.24	17.44	0	ten	22-5-2016 11:00
28	0	0	51.27801794	5.27705783	0	42.1	3	6	3	74.17	14.7	0	ten	22-5-2016 11:00
14	81.03	18.6	51.24822748	5.25888245	11.9	17.6	3	5	3	62.64	17.38	0	ten	22-5-2016 11:00
1	0	0	51.25081999	5.30829058	0	42.8	3	5	3	67.41	18.43	0	ten	22-5-2016 11:00
34	77.84	18.97	51.27146323	5.28599157	5.9	25.9	2	4	2	63.1	20.4	0	ten	22-5-2016 11:00
6	76.77	19.13	51.28418095	5.29747937	8.3	33.1	3	5	3	55.5	18.66	0	ten	22-5-2016 11:00
4	80.7	18.37	51.25798047	5.27417798	7.2	43.3	2	5	3	61.34	17.59	0	ten	22-5-2016 11:00
30	79.48	18.43	51.26252506	5.284783	10.4	40.9	2	5	3	60.2	18.2	3675	ten	22-5-2016 11:00
35	83.73	17.7	51.27277401	5.27686146	10.5	66.2	4	6	4	67.24	18.72	0	ten	22-5-2016 11:00
13	83.28	17.28	51.26482916	5.2522282	2.5	67.3	3	6	3	73.67	14.96	0	ten	22-5-2016 11:00
29	77.17	18.86	51.27238161	5.29192739	7	29.2	3	7	3	62.27	17.7	0	ten	22-5-2016 11:00
22	77.1	19.07	51.26718066	5.30508484	6.9	39.3	2	8	3	59.46	18.95	0	ten	22-5-2016 11:00
26	80.13	17.84	51.26322511	5.28941871	11.3	75.3	2	5	2	72.4	14.89	0	ten	22-5-2016 11:00
16	82.43	18.52	51.2677628	5.26925307	11.3	24.9	2	6	2	61.9	17.29	0	ten	22-5-2016 11:01
24	0	0	51.26642895	5.2606355	0	23	3	6	4	64.9	16.31	0	ten	22-5-2016 11:10
9	0	0	51.28783528	5.28504929	0	34	2	6	3	75.61	15.74	0	ten	22-5-2016 11:10
17	0	0	51.28866095	5.25794301	0	46.5	3	6	3	68.46	15.82	0	ten	22-5-2016 11:10
7	77.69	18.12	51.26248517	5.27099605	18.5	25	2	4	3	58.57	18.03	0	ten	22-5-2016 11:10
27	0	0	51.25703797	5.29670989	0	21.2	2	6	3	58.97	19.58	0	ten	22-5-2016 11:10
38	79.57	18.85	51.3535407	4.44486825	0	0	5	8	6	76.93	15.24	0	ten	22-5-2016 11:10

Figure 13. Excerpt of CSV data from AiREAS on 22nd of May 2016.

Figure 14 is an excerpt of data from the enviroCar track measured on 22nd of May 2016 in the area of Eindhoven in the vicinity of the airboxes of AiREAS.

```
{
    "type": "FeatureCollection",
    "properties": {
        "id": "576e8022e4b0ea2463fc619e",
        "sensor": {
            "type": "car",
            "properties": {
                "engineDisplacement": 2997,
                "model": "740i",
                "id": "576911e7e4b091b7fce195e9",
                "fuelType": "gasoline",
                "constructionYear": 2013,
                "manufacturer": "BMW"
            }
        },
        "length": 32.592065818929925
    },
    "features": [
        {
            "type": "Feature",
            "geometry": {
                "type": "Point",
                "coordinates": [
                    5.4134792246851084,
                    51.50290123281994
                ]
            },
            "properties": {
                "id": "576e8022e4b0ea2463fc61a0",
                "time": "2016-06-25T09:16:16Z",
                "phenomenons": {
                    "Speed": {
                        "value": 0,
                        "unit": "km/h"
                    },
                    "Consumption": {
                        "value": 2.946150199894973,
                        "unit": "l/h"
                    },
                    "GPS Accuracy": {
                        "value": 6.000000178813934,
                        "unit": "%"
                    },
                    "GPS HDOP": {
                        "value": 0.9842857122421266,
                        "unit": "precision"
                    },
                    "GPS Bearing": {
                        "value": 149.58769671880282,
                        "unit": "deg"
                    },
                    "GPS VDOP": {
                        "value": 1.6892857193946837,
                        "unit": "precision"
                    },
                    "GPS Speed": {
                        "value": 0,
```


Figure 14. Excerpt of JSON data from enviroCar on 22nd of May 2016.