DISCOVERY, INTEGRATION AND AGGREGATION OF SENSOR DATA USING THE SEMANTIC WEB

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of

Master of Science in Geomatics

by

Ivo de Liefde

June 2016

Ivo de Liefde: *Discovery, Integration and Aggregation of Sensor Data Using the Semantic Web* (2016)

The work in this thesis was made in the:

OTB - Research for the Built Environment
Faculty of Architecture & the Built Environment
Delft University of Technology

Supervisors: Drs. M. de Vries
dr.ir. B.M. Meijers

ABSTRACT

[Should fit on one page.]

Bacon ipsum dolour sit amet porchetta beef turkey, bacon turducken boudin hamburger venison ball tip. Brisket pork loin bresaola short loin ground round leberkas pastrami tongue jerky cow turducken beef ribs. Pork ribeye landjaeger prosciutto pig venison tenderloin. Swine beef ribs kielbasa, porchetta tenderloin salami venison pork belly tail. Bacon ipsum dolour sit amet porchetta beef turkey, bacon turducken boudin hamburger venison ball tip. Brisket pork loin bresaola short loin ground round leberkas pastrami tongue jerky cow turducken beef ribs. Pork ribeye landjaeger prosciutto pig venison tenderloin. Swine beef ribs kielbasa, porchetta tenderloin salami venison pork belly tail. Bacon ipsum dolour sit amet porchetta beef turkey, bacon turducken boudin hamburger venison ball tip. Brisket pork loin bresaola short loin ground round leberkas pastrami tongue jerky cow turducken beef ribs. Pork ribeye landjaeger prosciutto pig venison tenderloin. Swine beef ribs kielbasa, porchetta tenderloin salami venison pork belly tail.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

API	Application Programming Interface	2
CBS	Dutch central agency for statistics	15
CORINE	Coordination of Information on the Environment	15
EEA	European Environment Agency	21
EU	European Union	1
HTTP	HyperText Transfer Protocol	10
INSPIRE	Infrastructure for Spatial Information in Europe	1
IoT	Internet of Things	1
IRCEL-CELINE	Belgian interregional environment agency	vii
IRI	International Resource Identifier	11
ISO	International Organisation for Standardisation	1
OGC	Open Geospatial Consortium	1
O&M	Observations and Measurements	1
OWL	Web Ontology Language	1
RDF	Resource Description Framework	vii
REST	Representational State Transfer	5
RIVM	Dutch national institute for public health and the environment	vii
SEL	Semantic Enablement Layer	5
Sem-SOS	Semantically Enabled SOS	5
SensorML	Sensor Modelling Language	1
SIR	Sensor Instance Registry	2
SOR	Sensor Observable Registry	2
SOS	Sensor Observation Service	vii
SPARQL	SPARQL Protocol and RDF Query Language	1
SSNO	Semantic Sensor Network Ontology	6
SSW	Semantic Sensor Web	2
SWE	Sensor Web Enablement	1
UML	Unified Modeling Language	vii
URI	Uniform Resource Identifier	3
URL	Uniform Resource Locator	9
W3C	World Wide Web Consortium	2
WCS	Web Coverage Service	2
WKT	Well-Known Text	12
WFS	Web Feature Service	2
WMS	Web Map Service	2
WPS	Web Processing Service	8
XML	Extensible Markup Language	3

1 INTRODUCTION

From 2020 onwards all member states of the European Union (EU) should provide sensor data to the Infrastructure for Spatial Information in Europe (INSPIRE) in order to comply with annex II and III of the INSPIRE directive [INSPIRE, 2015]. For this a number of Sensor Web Enablement (SWE) standards are required to be used [INSPIRE, 2014]. The sensor web is a relatively new development and there are still many questions on how to structure it. This thesis aims to design a method to publish and link sensor metadata on the semantic web to improve the discovery, integration and aggregation of sensor data.

1.1 BACKGROUND

In 2008 the Open Geospatial Consortium (OGC) introduced a new set of standards called Sensor Web Enablement (SWE). These standards make it possible to connect sensors to the internet and retrieve data in a uniform way. This allows users or applications to retrieve sensor data through standard protocols, regardless of the type of observations or the sensor's manufacturer [Botts et al., 2008]. Among other standards SWE includes the Observations and Measurements (O&M) which is a model for encoding sensor data, the Sensor Modelling Language (SensorML) which is a model for describing sensor metadata and the SOS which is a service for retrieving sensor data [Botts et al., 2007]. O&M has also been adopted by the International Organisation for Standardisation (ISO) under ISO 19156:2011 [ISO, 2011].

Recently OGC has defined the role which their standards could play in smart city developments [Percivall, 2015]. Smart cities can be defined as "enhanced city systems which use data and technology to achieve integrated management and interoperability" [Moir et al., 2014, p. 18]. Research on smart cities has shown a great potential for using sensor data in urban areas. Often this is presented in the context of the Internet of Things (IoT) [Zanella et al., 2014; Wang et al., 2015a]. The IoT can be described as "the pervasive presence around us of a variety of *things* or *objects* ... [which] are able to interact with each other and cooperate with their neighbors to reach common goals" [Atzori et al., 2010, p. 2787].

Parallel to the development of the sensor web other research has focused on the semantic web, as proposed by Berners-Lee et al. [2001]. This is a response to the traditional way of using the web, where information is only available for humans to read. The semantic web is an extension of the internet which contains meaningful data that machines can understand as well. Rather than publishing documents on the internet the semantic web contains linked data using the RDF, also known as the *web of data* [Bizer et al., 2009]. Data in RDF can be queried using the SPARQL Protocol and RDF Query Language (SPARQL) at so-called SPARQL endpoints. The Web Ontology Language (OWL) is an extension of RDF and was designed "to represent rich and complex knowledge about things, groups of things, and relations between things" [OWL working group, 2012]. Originally, the semantic web intended to add metadata to the internet [Lassila and Swick, 1999]. However, today it is being used for linking any kind of data from one source to another in a meaningful way [Cambridge Semantics, 2015].

Sheth et al. [2008] proposes to use semantic web technologies in the sensor web. This Semantic Sensor Web (SSW) builds on standards by OGC and the World Wide Web Consortium (W3C) "to provide enhanced descriptions and meaning to sensor data" [Sheth et al., 2008, p. 78]. W3C responded to this development by creating a standard ontology for sensor data on the semantic web [Compton et al., 2012].

1.2 PROBLEM STATEMENT

Finding sensor data that can be retrieved using open standards is not easy. The implementation of the sensor web is still in an early stage. At the moment there are only a limited number of SOS implementations available on the web and they contain a limited amount of data. In the Netherlands the SOS by the RIVM is one of the first ones to be developed. It has only recently been launched and contains data on air quality. A number of other organisations still use a custom Application Programming Interface (API) to retrieve data from sensors connected to the internet. The problem of these custom APIs is that it is very hard to create an application that automatically retrieves data from them, because they have not implemented standards regarding the content of their service, the metadata models behind it or the kind of requests that can be made. It forces the application to have knowledge built in on the specifics of the individual APIs that are being used.

It has been researched to what extent a catalogue service could be useful for discovering sensor data from a SOS using the web service interfaces Sensor Instance Registry (SIR) [Jirka and Nüst, 2010] and Sensor Observable Registry (SOR) [Jirka and Bröring, 2009]. Catalogue services have already been available for example for the Web Map Service (WMS), Web Feature Service (WFS) or Web Coverage Service (WCS) [Nebert et al., 2007]. However, for the sensor data sources used in this paper no register or catalogue service has been implemented. Atkinson et al. [2015] also argues that catalogue services have a number of major disadvantages. It places a very high burden on the client to not only know where to find the catalogue service, but also to have knowledge on all kinds of other aspects (e.g. its organisation, access protocol, response format and response content) [Atkinson et al., 2015, p. 128]. Atkinson et al. suggest that linked data is therefore a much better solution for discovering sensor data.

However, for sensor data to be discovered on the semantic web there have to be inward links, from other sources linking towards the sensor (meta)data. Current research on the SSW has focused on publishing sensor data on the semantic web with links that point outwards [Atkinson et al., 2015; Janowicz et al., 2013; Pschorr, 2013]. This gives meaning to the data and is useful in order to work with the data, but it has a very limited effect on the discovery of the sensor data by others.

One of the challenges of using sensor data is the difficulty of integrating it from different sources to perform data fusion [Corcho and Garcia-Castro, 2010; Ji et al., 2014; Wang et al., 2015b]. Data fusion is "a data processing technique that associates, combines, aggregates, and integrates data from different sources" [Wang et al., 2015a, p. 2]. Even if the sources comply with the SWE standards it is challenging, since the data can be of a different granularity, both in time and space. Spatio-temporal irregularities are a fundamental property of sensor data [Ganesan et al., 2004].

The question arises to what extent the semantic web could be a better solution for publishing sensor data than the current geoweb solutions like SOS. The geoweb has some very good qualities, such as very structured approaches through which (sensor) data can be retrieved using well defined services. These standardised services have been accepted by large organisations as OGC and ISO. Furthermore, they are often based on years of discussion. This is different from for example web pages where content can be completely unstructured. The response of a SOS also contains some semantics about sensor data. There can be x-links inside the Extensible Markup Language (XML) with Uniform Resource Identifier (URI)s that point to semantic definitions of objects.

Still, the semantic web could be beneficial for the geoweb. Since data on the web has a distributed nature it can be questioned whether centralised catalogue services are feasible to create. It places a burden on the owner of the SOS to register with a catalogue service. Also, there could be multiple of these services on the web creating issue regarding the discovery of relevant catalogues. The semantic web could solve this issue by getting rid of the information silos and storing data directly on the web instead. This allows the interlinking and reuse of data on the web, which makes it easier to find related data. For automatic integration and aggregation it could be useful that the semantic web is machine understandable.

In conclusion, the problem to be addressed is the lack of knowledge on how to exploit the full potential of the sensor web using the semantic web. Creating the right links could greatly enhance the discovery, integration and aggregation of sensor data. However, there is no method yet to establish this linked metadata for sensors, while the standardised nature of a SOS should allow for generating it in an automated process. This thesis will create a design for such an automated process, research how to establish inward links and explore the advantages and disadvantages of publishing sensor metadata on the semantic web with a proof-of-concept implementation.

1.3 SCIENTIFIC RELEVANCE

Sensor data ties together many different fields of research. On the one hand there is research on how to create the most efficient sensor networks that uses the least amount of power to transfer the observed data over long distances [Korteweg et al., 2007; Xiang et al., 2013]. This involves academic fields such as mathematics, physics and electrical engineering. On the other hand there is research that uses sensor data to gain insights into real world phenomenon. This involves academic fields such as geography, environmental studies and urbanism. In order to connect these scientific fields, studies have focused on the use of computer science and standardisation for transferring sensor data over the internet.

In the future more sensor data is expected to be produced [Price Waterhouse Coopers, 2014]. Both experts and non-experts will be involved in this development. Experts will produce more data because of European legislation (INSPIRE). Non-experts will be involved more often via smart cities and IoT developments where users or consumer electronics produce sensor data as well. This vast amount of data could be very useful for academic research, provided researchers are able to find the data they need online and are able to integrate and aggregate data from heterogeneous sources. Publishing sensor metadata on the semantic web could make it easier to find what you need through related data on the internet. Having a automated process for this and being able to seamlessly integrate and aggregate data from different sources could be of great use for research such as van der Hoeven et al. [2014], Van der Hoeven and Wandl [2015] and Theunisse [2015]. They are examples of studies that try to understand phenomenon in the built environment using sensor data. Currently data collection and processing takes up a large part of the research, while with the implementation of SWE standards and the use of the semantic web this might be significantly reduced.

1.4 RESEARCH QUESTION

This thesis aims to design a method that uses the semantic web to improve sensor data discovery as well as the integration and aggregation of sensor data from heterogeneous sources. The following question will be answered in this research: *To what extent can the semantic web improve the discovery, integration and aggregation of distributed sensor data?*

2 RELATED WORK

A number of research topics are relevant for this thesis: how to use existing standards for publishing sensor data to the semantic web, developing ontologies that are suitable for many different kinds of sensor data and how to aggregate sensor data based on geographical features and time. This chapter discusses the recent relevant literature on these topics.

2.1 SENSOR DATA CATALOGUE SERVICE

The SOR is "a web service interface for managing the definitions of phenomena measured by sensors as well as exploring semantic relationships between these phenomena" [Jirka and Bröring, 2009, p. vi]. This is a web service developed by OGC to enable semantic reasoning on sensor networks, especially concerning phenomenon definitions. This should make it easier to discover sensors that observe a certain phenomenon and to interpret sensor data.

Another web service interface specification by OGC is SIR. SIR is aimed at "managing the metadata and status information of sensors" [Jirka and Nüst, 2010, p. xii]. The goal of this web service is to close the gap between metadata models based on SensorML, which is used in SWE, and the metadata model used in OGC catalogue services. Furthermore, it provides functionalities to discover sensors, to harvest sensor metadata from a SOS, to handle status information about sensors and to link SIR instances to OGC catalogue services.

2.2 SEMANTIC SENSOR DATA MIDDLEWARE

Henson et al. [2009] and Pschorr [2013] suggest adding semantic annotations to a SOS which they call Semantically Enabled SOS (Sem-SOS). In Sem-SOS the raw sensor data goes through a process of semantic annotating before it can be requested with a SOS service. The retrieved data is still an XML document, but with embedded semantic terminology as defined in an ontology. The data retrieved from Sem-SOS is therefore semantically enriched.

Janowicz et al. [2013] has specified a method that uses a Representational State Transfer (REST)ful proxy as a façade for SOS. When a specific URI is requested the so-called Semantic Enablement Layer (SEL) translates this to a SOS request, fetches the data and translates the results back to RDF. In this method the sensor data is converted to RDF on-the-fly. This allows the data to be interpreted by both humans and machines.

Atkinson et al. [2015] have identified that "distributed heterogeneous data sources are a necessary reality in the case of widespread phenomena with multiple stakeholder perspectives" [Atkinson et al., 2015, p.129]. Therefore, they propose that methods should be developed to move away from the

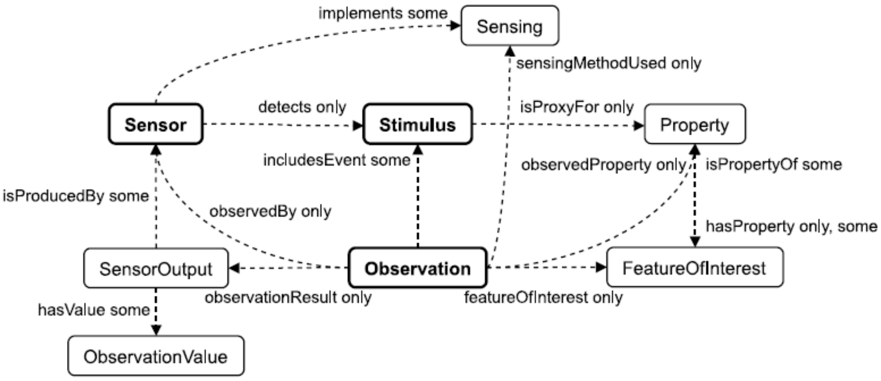


Figure 2.1: The stimulus–sensor–observation pattern [Compton et al., 2012, p. 28]

traditional dataset centric approaches and towards using linked data for cataloguing. This has the potential to bring together data and knowledge from different areas of research about the same (or similar) features-of-interest. It is also argued that using both linked data services and data-specific services could ease the transition into the linked data world.

2.3 SENSOR DATA ONTOLOGIES

Ontologies are necessary to provide meaning to data on the semantic web and to create semantic interoperability. Three recent efforts for developing a standard ontology for sensor data based on SWE standards will be discussed here.

2.3.1 Semantic sensor network ontology

W3C has developed an ontology for sensors and observations called the Semantic Sensor Network Ontology (SSNO). This ontology aims to address semantic interoperability on top of the syntactic operability that the SWE standards provide. To accommodate different definitions of the same concepts the broadest definitions have been used. Depending on the interpretation these can be further defined with subconcepts. The SSNO is based on the stimulus-sensor-observation pattern, describing the relations between a sensor, a stimulus and observations (Figure 2.1). Sensors are defined as "physical objects ... that observe, transforming incoming stimuli ... into another, often digital, representation", stimuli are defined as "changes or states ... in an environment that a sensor can detect and use to measure a property" and observations are defined as "contexts for interpreting incoming stimuli and fixing parameters such as time and location" [Compton et al., 2012, p. 28]. The ontology can be used to model sensor networks from four different perspectives (sensor, observation, system, and feature & property), which they discuss together with additional relevant concepts.

2.3.2 Observation capability metadata model

Hu et al. [2014] have reviewed a number of metadata models (including SensorML and SSNO) for the use of earth observation (including remote sensing). They argue that all of the current metadata models are not sufficient for sensor data discovery. This conclusion is based on an evaluation of six criteria. Three steps were identified in the process of obtaining relevant sensor data for earth observation, which have been used to derive criteria for their evaluation framework. These steps are sensor filtration, sensor optimisation and sensor dispatch. The filtration of sensors should result in a set of sensors that meets the requirements of the application: It should measure the right phenomenon, be active, be inside the spatial and temporal range, and have a certain sample interval. In sensor optimisation the selected sensors should be combined to complement or enhance each other. To do this, the observation quality, coverage and application is relevant. In the last step – sensor dispatch – the data should be retrieved, stored and transmitted. In every evaluated model the same sensors can be described in different ways or only partially, which affects the outcome of the sensor dispatch.

Therefore, a metadata model is proposed that "reuses and extends the existing sensor observation-related metadata standards" [Hu et al., 2014, p. 10546]. It is composed of five modules: observation breadth, observation depth, observation frequency, observation quality and observation data. They should be derived from metadata elements described using the Dublin Core metadata element set. These five modules can then be formalised following the SensorML schema which can be queried by users via a 'Unified Sensor Capability Description Model-based Engine'.

2.3.3 Om-lite & sam-lite ontologies

Cox [2015b] has been working on new semantic ontologies based on O&M. Previous efforts, such as the SSNO have been using pre-existing ontologies and frameworks. However, there are already many linked data ontologies that could be useful for describing observation metadata, such as space and time concepts. Also, the SSNO does not take sampling features into account. Therefore, Cox [2015b] proposes two new ontologies: OWL for observations or om-lite [Cox, 2015a], which defines the concepts from O&M regarding observations and OWL for sampling features or sam-lite, which defines the sampling feature concepts [Cox, 2015d]. A mapping of the SSNO to om-lite is also provided.

Cox [2015b] describes how the PROV ontology [Lebo et al., 2013] can be directly used inside om-lite. The PROV ontology is "concerned with the production and transformation of Entities through time-bounded Activities, under the influence or control of Agents" [Cox, 2015b, p. 12]. This is a very convenient ontology for modelling real world entities, such as sensors, observation processes and sampling processes. Many other ontologies could be implemented in combination with om-lite and sam-lite, depending on the kind of observations that are being modelled and the data publisher's preference.

2.4 SENSOR DATA AGGREGATION

Sensor data aggregation can be performed for two purposes: To reduce the energy constraint of sensor networks [Korteweg et al., 2007] or to sample a feature-of-interest in space and/or time [INSPIRE, 2014]. Sampling is performed when a feature-of-interest is not accessible, in which case "observations are made on a subset of the complete feature, with the intention that the sample represents the whole" [Cox, 2015a]. Stasch et al. [2011a] proposes a Web Processing Service (WPS) that retrieves sensor data from a SOS service in order to aggregate it based on features-of-interest. The approach by Stasch et al. [2011b] is similar, but takes sensor data as input that is already published on the semantic web.

Ganesan et al. [2004] stresses that spatio-temporal irregularities are fundamental to sensor networks. Irregular sampling can have a potentially large influence on the accuracy of the aggregated outcome. For example, averaging sensor data from a feature-of-interest that is being sampled densely in some parts and more sparsely in other parts could lead to inaccurate results. To counter this the values of the densely sampled area should have a lower weight than the values from the sparsely sampled area. The same holds true for temporal irregularities [Ganesan et al., 2004]. Also, Stasch et al. [2014] argue that in order for automatic aggregation to work there needs to be semantics on which kind of aggregation methods are appropriate for a specific kind of sensor data. Not all kinds of aggregation are meaningful (e.g. taking the sum of temperature values). This requires a formalisation of expert knowledge which they call semantic reference systems.

3 METHODS

A number of studies related to this thesis have been reviewed in Chapter 2. This chapter discusses why the semantic web will be used for linking sensor metadata and which methods will be used to achieve this. The SWE standards, the om-lite and sam-lite ontologies, and RDF will be described.

3.1 SENSOR METADATA ON THE SEMANTIC WEB

Sem-SOS [Henson et al., 2009; Pschorr, 2013] as well as SEL [Janowicz et al., 2013] focus on combining the sensor web with the semantic web, but do not address the integration and aggregation of sensor data. Similarly, Atkinson et al. [2015] proposes to expose sensor data to the semantic web in order to find other kinds of related data about the same feature-of-interest. Data that can be collected for another area of research. However, Atkinson et al. [2015] do not mention the integration of complementary sensor data from heterogeneous sources either. Stasch et al. [2011b] and Stasch et al. [2011a] suggest interesting methods for aggregating sensor data based on features-of-interest. However, also these studies use sensor data from only a single source into account. Moreover, Corcho and Garcia-Castro [2010] and Ji et al. [2014] argue that methods for integration and fusion of sensor data on the semantic web is still an area for future research. Data fusion is "a data processing technique that associates, combines, aggregates, and integrates data from different sources" [Wang et al., 2015a, p. 2].

Jirka and Nüst [2010] and Jirka and Bröring [2009] present methods for including SOS services in an OGC catalogue service using SOR and SIR. Making sensor metadata available in a catalogue service will improve the discovery. However, discovery through the semantic web is likely to be more effective, since links can be created towards the sensor data from many different sources of related information. Another advantage is that links can be created by everybody that publishes linked data on the web, allowing sensor data to be used for implementations that were not identified beforehand by the publisher. Also, the semantic web will be easier to access, while the catalogue service can only be requested at a certain Uniform Resource Locator (URL) which has to be known to potential users.

Since data on the web has a distributed nature it can be questioned whether centralised catalogue services are feasible to create. It places a burden on the owner of the SOS to register with a catalogue service. Also, there could be multiple of these services on the web creating issue regarding the discovery of relevant catalogues. The semantic web could solve this issue by getting rid of the 'dataset-centric' approach and adding metadata directly to the web instead.

3.2 SENSOR OBSERVATION SERVICE

There are three core requests that can be made to retrieve sensor (meta)data from a SOS: GetCapabilities, DescribeSensor and GetObservation. Get Capabilities returns a complete overview of what the SOS has to offer. The DescribeSensor request returns detailed information about individual sensors. These three core requests are mandatory in a SOS under the 2.0 specifications [Bröring et al., 2012]. There are also a number of optional extensions to a SOS. Requests can be made as a HyperText Transfer Protocol (HTTP) GET request or a HTTP POST request. There can be different response formats. Usually there is at least the option to retrieve the response as an XML document. Based on the specification by Bröring et al. [2012] this paragraph describes the core and optional requests of a SOS, as well as the structure of their responses.

3.2.1 Get capabilities

The GetCapabilities request is the first step in communicating with a SOS. The request is made by taking the HTTP address of the SOS and adding service=SOS&request=GetCapabilities. It returns a document including information on what the service has to offer. The document contains a number of sections: service identification, service provider, operations metadata, filter capabilities and contents.

In the service identification section there is general information about the service, such as the title and supported SOS versions, but also whether there are fees or access constraints. The service provider section contains details on which organisation provides the SOS and lists their contact information. The operations metadata section lists the supported request types. It also contains an overview of all features-of-interest, observed properties, procedures and offerings. Offerings are similar to layers in a WMS, grouping together observations collected by one procedure.

The contents section describes the data that can be retrieved, grouped in offerings. Each offering has an identifier, together with information on the procedure, observable properties and the feature-of-interest type. Which filters can be applied in a request is described in the filter capabilities section. The supported parameters for both spatial and temporal filters are listed here.

3.2.2 Describe sensor

The DescribeSensor request gives detailed information on a specific sensor. The request is built by taking the HTTP address of the SOS and adding service=SOS&version=2.0.0&request=DescribeSensor&procedure=aprocedure&proceduredescriptionformat=aformat where the procedure and procedure description format have to contain values defined in the capabilities document.

3.2.3 Get observation

Using GetObservation actual measurements can be retrieved. The request is made by taking the HTTP address of the SOS and adding service=SOS&version=2.0.0&request=GetObservation. This returns a response with the

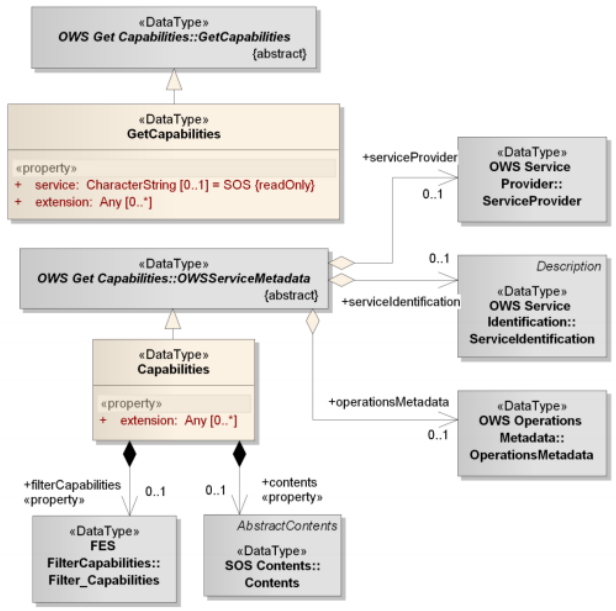


Figure 3.1: Data model behind the capabilities document of a SOS [Bröring et al., 2012]

default parameters, which can differ from one SOS to another. To further specify the request, optional paramaters can be added such as: observed property, procedure, feature-of-interest, offering and outputformat. Spatial and temporal filters can be added if these are supported by the service.

3.2.4 Optional extensions

3.3 RESOURCE DESCRIPTION FRAMEWORK

For publishing static geographic data on the semantic web a conversion of Shapefiles to RDF is required. For this the method by Missier [2015] will be used. First the Shapefile is loaded into a Postgres database with the Postgis extension. After that a Python script retrieves the records from the database. Attributes of the records will be mapped to classes from predefined ontologies. Then the script creates an RDF graph and serialises it to a certain RDF notation. This is written to a file. The final step is to publish the RDF on the web and create a SPARQL endpoint to query the data [Missier, 2015].

In RDF data is stored as so-called 'triples'. These triples are structured as: subject, predicate and object [Berners-Lee et al., 2001]. The subject and the object are things and the predicate is the relation between these two things. For example, to define a geographic feature such as the municipality of Delft on the semantic web a number of triples can be made. Figure 3.2 shows how Delft can be defined as a municipality with a certain geometry using triples of subject, predicate and object.

Three types of data can make up these triples. The first type is an International Resource Identifier (IRI). This is a reference to a resource and can be used for all positions of the triple. A URL is an example of an IRI, but IRIs can also refer to resources without stating a location or how it can be accessed. An IRI is a generalisation of an URI, and also allows non-ASCII characters. In the example of the municipality of Delft, IRIs can be used to define 'Delft'

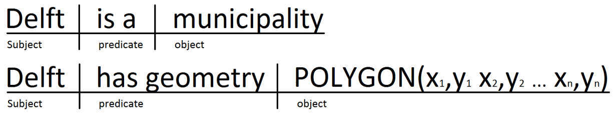
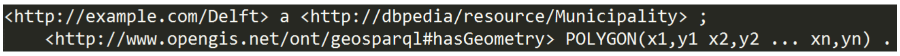


Figure 3.2: Triples of object, predicate and subject define Delft as a municipality with a geometry



```
<http://example.com/Delft> a <http://dbpedia/resource/Municipality> ;
    <http://www.opengis.net/ont/geosparql#hasGeometry> POLYGON(x1,y1 x2,y2 ... xn,yn) .
```

Figure 3.3: Triples of Figure 3.2 in the Turtle notation

and 'Municipality', but also for the predicates 'is a' and 'has geometry'. The second type of data is a literal. A literal is a value which is not an IRI, such as strings, numbers or dates. These values can only be used as object in a triple. In the example of Delft, a literal could be used to store the actual geometry of the boundary: POLYGON(x_1, y_1 x_2, y_2 ... x_n, y_n). Sometimes it is useful to refer to things without assigning them with a global identifier. The third type is the blank node and can be used as an subject or object without using an IRI or literal [Manola et al., 2014].

There are a number of different notations for writing down these triples (serialisation), such as XML [Gandon and Schreiber, 2014], N3 [Berners-Lee and Connolly, 2011] and Turtle [Beckett et al., 2014]. Turtle will be used in this thesis, because it is is used for the object 'Municipality'. The 'is a' predicate is represented by a built-in RDF predicate which can be written simple as 'a'. The second predicate is 'hasGeometry' for which the GeoSPARQL IRI is used. The geometry is a literal in the Well-Known Text (WKT) format. Note that the subject is only written once when there are multiple triples with the same subject. Triples that shares the same subject are divided by semicolons. A point marks the end of the last triple with a specific subject.

The sensor metadata will also be published on the semantic web. To do this an XML document is automatically retrieved from a SOS by a Python script. This script then extracts the relevant data from the XML and maps it to an ontology. It outputs an RDF file that will be published online. When new sources of sensor data are added the RDF documents will be updated.

3.4 ONTOLOGY MAPPING

When publishing data on the semantic web, ontologies are required to specify what things are and how they relate to other things. The evaluation of observation metadata ontologies by Hu et al. [2014] is interesting, since it exposes what the relevant aspects are in the process of observation discovery. However, their proposed model focusses mainly on including remote sensing and imagery data in metadata models that were not originally created for this kind of data. The SSNO is an ontology that clearly describes the process between sensor, stimulus and observation. However, Cox [2015b] points out that an important aspect of describing a sensor network is missing in this ontology: the sampling. Also, the om-lite and sam-lite ontologies by Cox [2015b] are lightweight ontologies that can be complemented by already existing linked data ontologies. They do not rely on the (heavy) ISO specifications that date from before the semantic web, unlike the SSNO. The om-lite and sam-lite ontologies will therefore be used in this thesis.

The UML diagram (Figure 3.4) describes different components of a SOS. The SOS has a number of metadata attributes such as the service provider's details (including contact information), its spatial and temporal extent (spatialFiler & temporalFilter) and the capabilities to query a subset of this extent. It receives data from a sensor which makes observations. An observation can be defined as "an action whose result is an estimate of the value of some property of the feature-of-interest, obtained using a specified procedure" [Cox, 2015a].The sensor is placed at a sampling point. The sampling point is part of a sampling feature which intents to resemble the feature-of-interest. In the case of air quality the feature-of-interest is the bubble of air surrounding the sensor, therefore the sampling point equals the feature-of-interest [INSPIRE, 2014]. The design is that an observation of the sampling feature describes the feature-of-interest through measuring one of its properties. The measurement procedure is described by a short string of text, input and output parameters and the units of measurement of the ouput. The relation between feature-of-interest and administrative units is added to improve the discovery of sensor data on the semantic web.

To publish data on the semantic web ontologies are required to specify the different classes and their relations. An ontology for static geographic data has to be connected to an ontology for sensor metadata. From the UML diagram in Figure 3.4 the classes Observation, Process, ObservedProperty and FeatureOfInterest can be mapped to classes belonging to OWL for observations [Cox, 2015c]. SamplingFeature and Sampling point can be mapped to classes from OWL for sampling features [Cox, 2015d]. GeoSPARQL can be used for the administrativeUnit class [Perry and Herring, 2011] and the PROV ontology for the sensor and sensor observation service classes [W3C Semantic Sensor Network Incubator Group, 2011].

3.5 SENSOR DATA AGGREGATION

There are many different ways to aggregate sensor data, for example by taking the minimum value, the maximum value, the average value, the sum, etc. Also, spatial aggregation techniques (based on neighbourhood analysis) can be considered to adjust for spatio-temporal irregularities as mentioned by Ganesan et al. [2004]. In order to determine which method of aggregation is applicable for a specific kind of sensor data the sensor metadata will contain links to appropriate aggregation methods. However, which methods are appropriate should be based on expert knowledge.

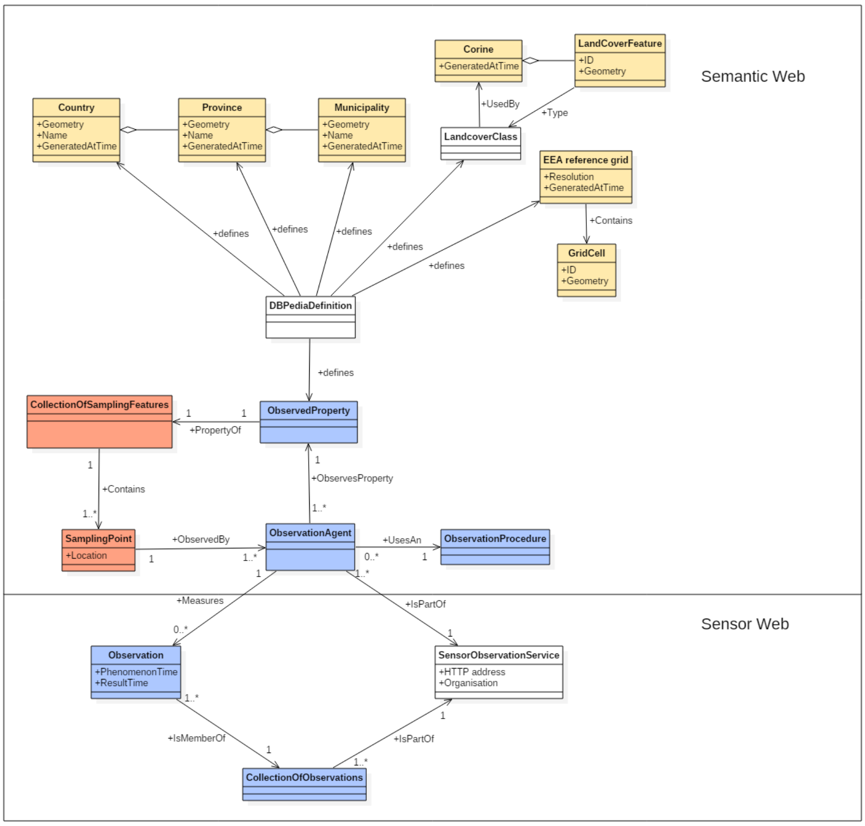


Figure 3.4: UML diagram showing sensor collections and their relation to the SOS

4 DATA

The data can be divided into static geographic data and sensor data. For static data the topographic data sets from the Dutch central agency for statistics (CBS) and the Dutch cadaster will be used. The dataset of Dutch provinces (provincies, Figure 4.2) and municipalities (gemeenten, Figure 4.1) has been downloaded from https://www.pdok.nl/nl/producten/pdok-downloads/basis-registratie-kadaster/bestuurlijke-grenzen-actueel. It is difficult to obtain data of administrative boundaries of Belgium (even from the INSPIRE data portal). Therefore, all data for Belgium was retrieved from http://www.gadm.org/. The geographic data contains the name of the administrative units and their (polygon) geometry.

Data on landcover will be used to complement the data of administrative units. A section of the 2012 dataset from the Coordination of Information on the Environment (CORINE) programme will be used (Figure 4.3). This dataset contains polygons (Figure 4.4) with a unique identifier, a code that defines the type of landcover and the size of the polygon's surface. It was downloaded from http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012.

Air quality sensor data will be used from the RIVM (http://inspire.rivm.nl/sos/) and from the IRCEL-CELINE (http://sos.irceline.be/). Both of these organisations have a SOS where data can be retrieved according to the SWE standards. The one of the RIVM has been online since the 21st of August, 2015. IRCEL-CELINE already made the SOS available on the first of January, 2011. Figure 4.5 and Figure 4.6 show the sensor networks of both organisations. They provide different kinds of sensor data, such as particulate matter (PM_{10}), nitrogen dioxide (NO^2) and ozone (O^3). Figure 4.7 shows one of the sensor locations in the city center of Amsterdam.

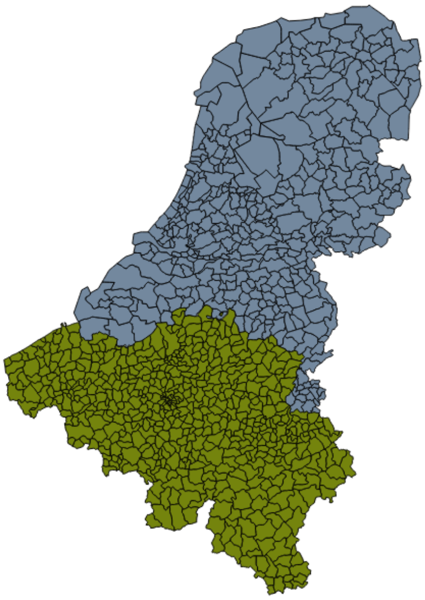


Figure 4.1: Dataset of municipalities in the Netherlands and Belgium in 2015 (from Dutch cadaster and GADM.org)

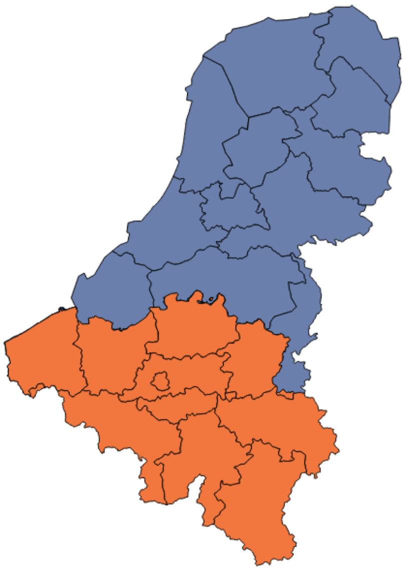


Figure 4.2: Dataset of provinces in the Netherlands and Belgium in 2015 (from Dutch cadaster and GADM.org)

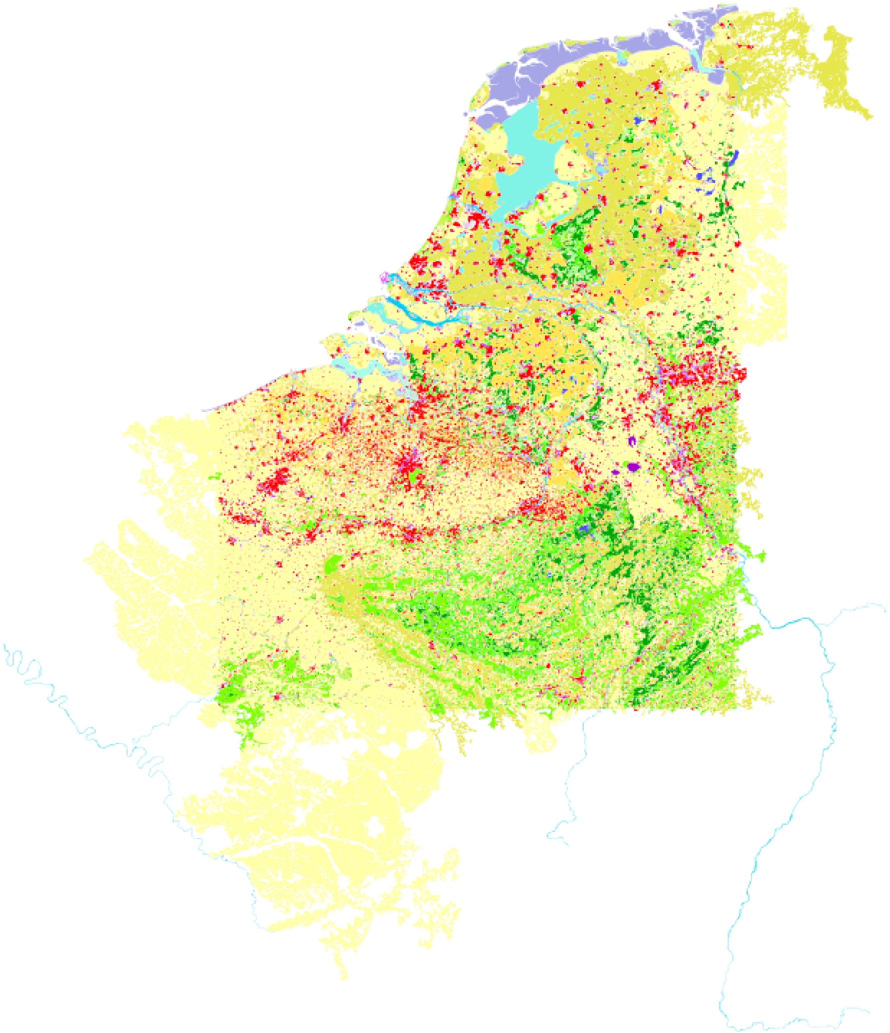


Figure 4.3: Dataset of landcover in the Netherlands and Belgium in 2012 (from Copernicus The European Earth Observation Programme)

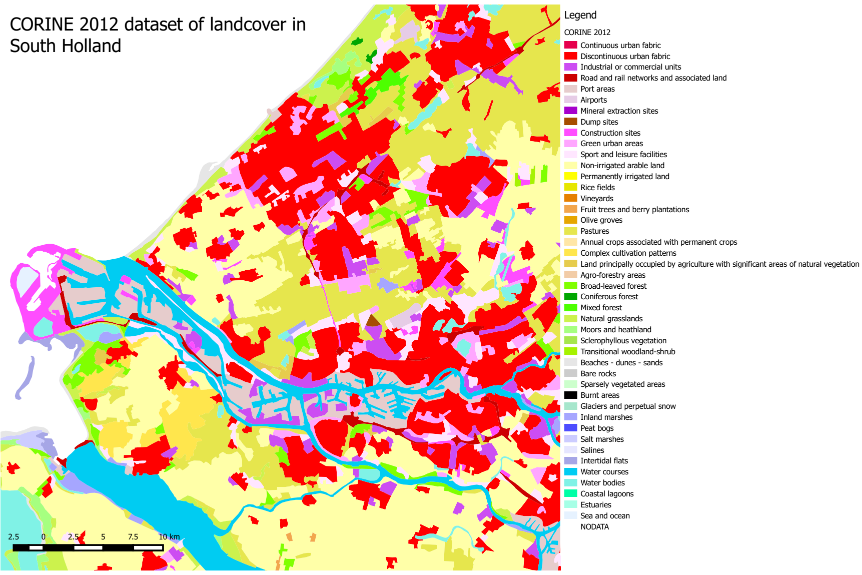

CORINE 2012 dataset of landcover in South Holland
Legend
CORINE 2012
Continuous urban fabric
Discontinuous urban fabric
Industrial or commercial units
Road and rail networks and associated land
Port areas
Airports
Mineral extraction sites
Dump sites
Construction sites
Green urban areas
Sport and leisure facilities
Non-irrigated arable land
Permanently irrigated land
Rice fields
Vineyards
Fruit trees and berry plantations
Olive groves
Pastures
Annual crops associated with permanent crops
Complex cultivation patterns
Land principally occupied by agriculture with significant areas of natural vegetation
Agro-forestry areas
Broad-leaved forest
Coniferous forest
Mixed forest
Natural grasslands
Moors and heathland
Sclerophyllous vegetation
Transitional woodland-shrub
Beaches - dunes - sands
Bare rocks
Sparsely vegetated areas
Burnt areas
Glaciers and perpetual snow
Inland marshes
Peat bogs
Salt marshes
Salines
Intertidal flats
Water courses
Water bodies
Coastal lagoons
Estuaries
Sea and ocean
NODATA
2.5 0 2.5 5 7.5 10 km

Figure 4.4: Landcover of the province of South Holland (subsection of the dataset from Figure 4.3)

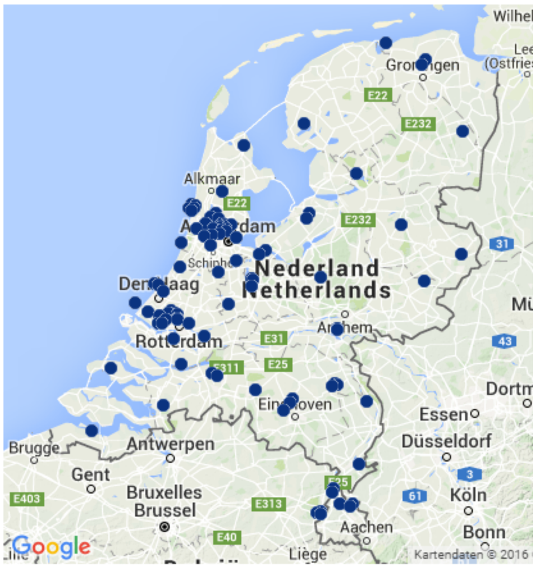


Figure 4.5: Webmap by the RIVM showing their air quality sensor network (http://www.lml.rivm.nl/meetnet)

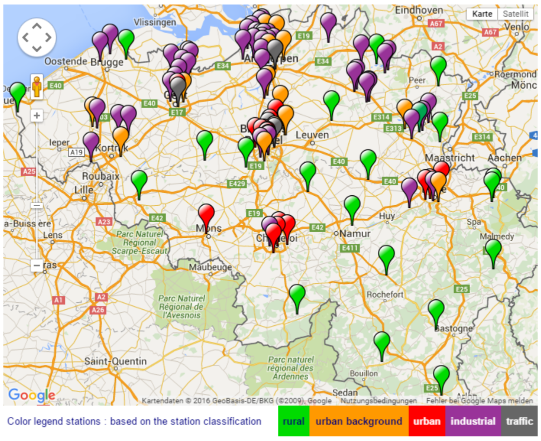


Figure 4.6: Webmap by IRCEL-CELINE showing their air quality sensor network (`http://www.irceline.be/en/air-quality/measurements/monitoring-stations/`)

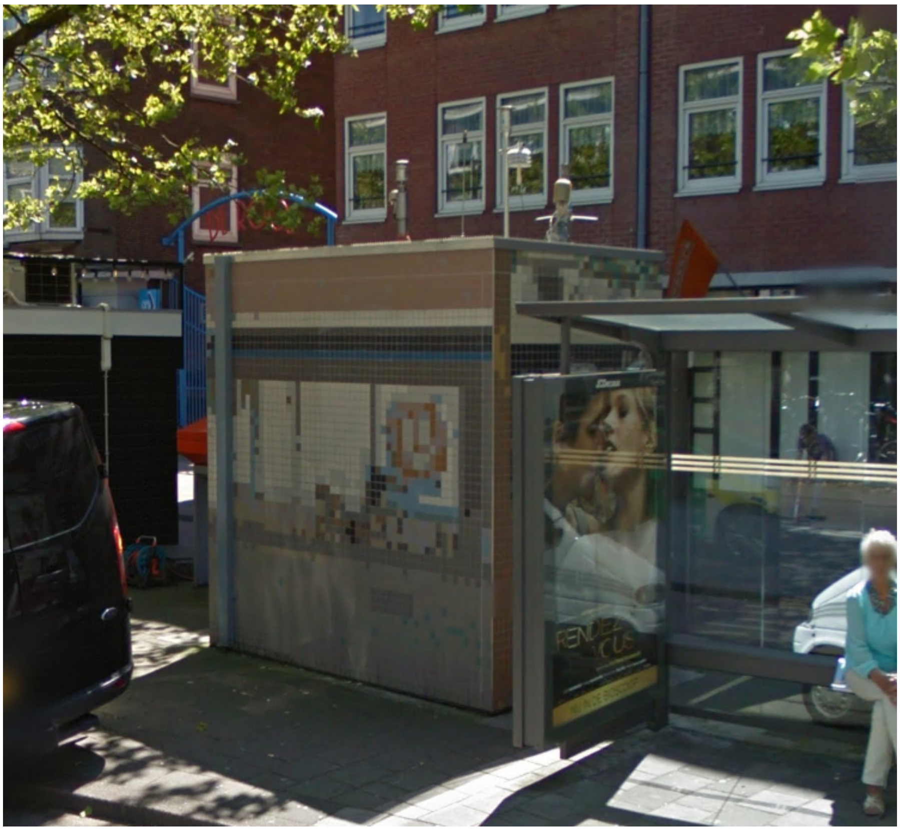


Figure 4.7: Google Streetview image of RIVM sensor location in Amsterdam in 2015

5 IMPLEMENTATION

Implementation of the methods are described in this chapter.

5.1 PREPARING LINKED DATA

First linked data is prepared that will be used to retrieve and process sensor data on the semantic web (Figure 5.1).

5.1.1 Creating linked data of administrative units

Data of countries, provinces and municipalities

Using RDFlib triples are added for every administrative unit:

- administrative unit > type > dbpedia country/province/municipality/neighbourhood
- administrative unit > foaf:Name > Name
- administrative unit > hasGeometry > geometryObject
- geometryObject > type > geometry
- geometryObject > dataType > WKTLiteral
- geometryObject > literal > Polygon(....)

5.1.2 Creating linked data of land cover

CORINE 2012 land cover dataset.

Using RDFlib triples are added for every landcover feature:

- landcover feature > type > dbpedia landcover type definition
- landcover feature> foaf:Name > ID
- landcover feature> hasGeometry > geometryObject
- geometryObject > type > geometry
- geometryObject > dataType > WKTLiteral
- geometryObject > literal > Polygon(....)

5.1.3 Creating linked raster data

The European Environment Agency (EEA) reference grid with resolutions of 10km^2 and 100km^2.

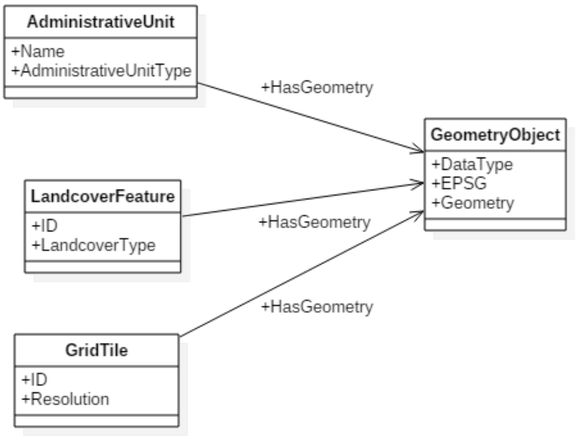
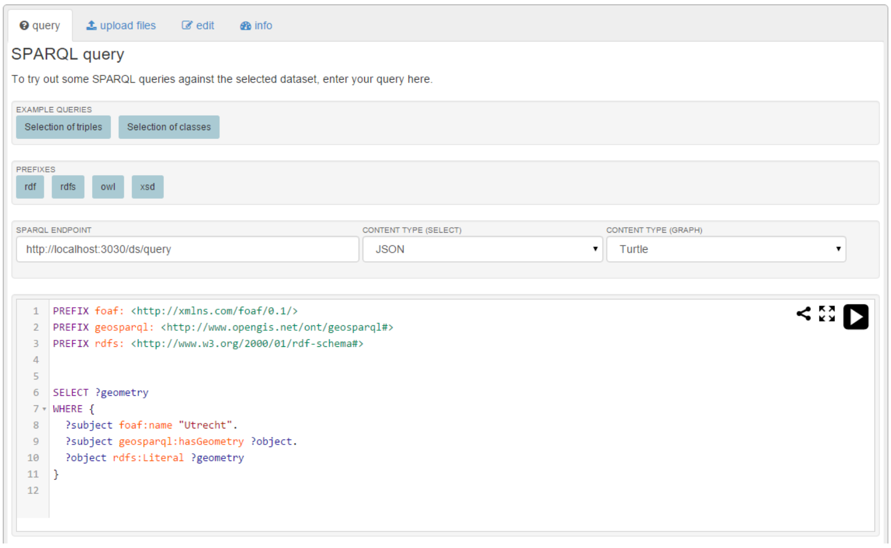


Figure 5.1: Static Data



SPARQL query

To try out some SPARQL queries against the selected dataset, enter your query here.

SPARQL ENDPOINT: http://localhost:3030/ds/query

CONTENT TYPE (SELECT): JSON

CONTENT TYPE (GRAPH): Turtle

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX geosparql: <http://www.opengis.net/ont/geosparql#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>


SELECT ?geometry
WHERE {
  ?subject foaf:name "Utrecht".
  ?subject geosparql:hasGeometry ?object.
  ?object rdfs:Literal ?geometry
}

```

Figure 5.2: Apache Jena Fuseki endpoint

5.1.4 Publishing Resource Description Framework data

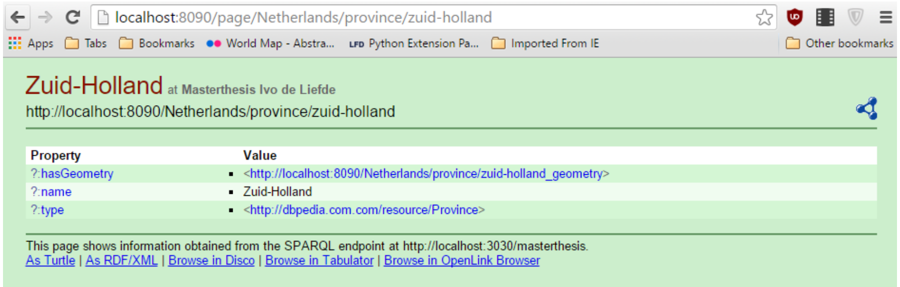
Setting up the Apache Jena Fuseki (Figure 5.2), Apache Tomcat and Pubby software (Figure 5.3).

5.2 RETRIEVING METADATA FROM THE SENSOR OBSERVATION SERVICE

The metadata is automatically retrieved from the SOS.

problems:

- Observation are the central concept in a SOS. This makes it hard to retrieve the locations of the sensors from the capabilities document. A workaround is to retrieve small amounts of observations for each feature of interest.



Zuid-Holland at Masterthesis Ivo de Liefde
http://localhost:8090/Netherlands/province/zuid-holland

Property	Value
?:hasGeometry	▪ <http://localhost:8090/Netherlands/province/zuid-holland_geometry>
?:name	▪ Zuid-Holland
?:type	▪ <http://dbpedia.com.com/resource/Province>

This page shows information obtained from the SPARQL endpoint at http://localhost:3030/masterthesis.
As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

Figure 5.3: Pubby interface for RDF

- Although standardised, the SOS contains information that can be implemented in different ways. This makes it hard to automatically retrieve data. An example is the feature of interest in the GetObservation document.

5.3 MODELLING WITH THE OM-LITE AND SAM-LITE ONTOLOGIES

Using the ontologies to create a model that suits the purpose of this thesis.

5.4 ESTABLISHING INWARD LINKS FROM DBPEDIA

Sending triples to DBPedia the project.

5.5 SETTING UP THE WEB PROCESSING SERVICES

Creating two WPS using PyWPS.

5.6 PROTOTYPE IMPLEMENTATION

Crawling the semantic web to find specific sensor data.

BIBLIOGRAPHY

Atkinson, R. A., Taylor, P., Squire, G., Car, N. J., Smith, D., and Menzel, M. (2015). Joining the Dots: Using Linked Data to Navigate between Features and Observational Data. In *Environmental Software Systems. Infrastructures, Services and Applications*, pages 121–130. Springer.

Atzori, L., Iera, A., and Morabito, G. (2010). The internet of things: A survey. *Computer networks*, 54(15):2787–2805.

Beckett, D., Berners-Lee, T., Prud'hommeaux, E., and Carothers, G. (2014). W3C RDF 1.1 Turtle. [online] http://www.w3.org/TR/turtle/ [accessed on December 9th, 2015].

Berners-Lee, T. and Connolly, D. (2011). W3C Notation3 (N3): A readable RDF syntax. [online] http://www.w3.org/TeamSubmission/n3/ [accessed on December 9th, 2015].

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5):28–37.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227.

Botts, M., Percivall, G., Reed, C., and Davidson, J. (2007). OGC Sensor Web Enablement: Overview And High Level Architecture. OGC document 06-021r1.

Botts, M., Percivall, G., Reed, C., and Davidson, J. (2008). OGC sensor web enablement: Overview and high level architecture. In *GeoSensor networks*, pages 175–190. Springer.

Bröring, A., Stasch, C., and Echterhoff, J. (2012). OGC Sensor observation service interface standard.

Cambridge Semantics (2015). Introduction to the Semantic Web. [online] https://www.cambridgesemantics.com/semantic-university/introduction-semantic-web [accessed on December 8th, 2015].

Compton, M., Barnaghi, P., Bermudez, L., GarcíA-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., et al. (2012). The SSN ontology of the W3C semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:25–32.

Corcho, O. and Garcia-Castro, R. (2010). Five challenges for the Semantic Sensor Web. *Semantic Web-Interoperability, Usability, Applicability*, 1.1(2):121–125.

Cox, S. J. D. (2015a). Observations and Sampling. [online] https://www.seegrid.csiro.au/wiki/AppSchemas/ObservationsAndSampling [accessed on December 1st, 2015].

Cox, S. J. D. (2015b). Ontology for observations and sampling features, with alignments to existing models.

Cox, S. J. D. (2015c). OWL for Observations. [online] http://def.seegrid.csiro.au/ontology/om/om-lite [accessed on November 24th, 2015].

Cox, S. J. D. (2015d). OWL for Sampling Features. [online] http://def.seegrid.csiro.au/ontology/om/sam-lite [accessed on November 24th, 2015].

Gandon, F. and Schreiber, G. (2014). W3C RDF 1.1 XML Syntax. [online] http://www.w3.org/TR/rdf-syntax-grammar/ [accessed on December 9th, 2015].

Ganesan, D., Ratnasamy, S., Wang, H., and Estrin, D. (2004). Coping with irregular spatio-temporal sampling in sensor networks. *ACM SIGCOMM Computer Communication Review*, 34(1):125–130.

Henson, C., Pschorr, J. K., Sheth, A. P., Thirunarayan, K., et al. (2009). SemSOS: Semantic sensor observation service. In *Collaborative Technologies and Systems, 2009. CTS'09. International Symposium on*, pages 44–53. IEEE.

Hu, C., Guan, Q., Chen, N., Li, J., Zhong, X., and Han, Y. (2014). An Observation Capability Metadata Model for EO Sensor Discovery in Sensor Web Enablement Environments. *Remote Sensing*, 6(11):10546–10570.

INSPIRE (2014). Guidelines for the use of Observations & Measurements and Sensor Web Enablement-related standards in INSPIRE Annex II and III data specification development.

INSPIRE (2015). INSPIRE Roadmap. [online] http://inspire.ec.europa.eu/index.cfm/pageid/44 [accessed on December 2nd, 2015].

ISO (2011). ISO 19156:2011; Geographic information – Observations and measurements. [online] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32574 [accessed on December 2 nd, 2015].

Janowicz, K., Broring, A., Stasch, C., Schad, S., Everding, T., and Llaves, A. (2013). A RESTful Proxy and Data Model for Linked Sensor Data. *International Journal of Digital Earth*, 6(3):233–254.

Ji, C., Liu, J., and Wang, X. (2014). A Review for Semantic Sensor Web Research and Applications. *Advanced Science and Technology Letters*, 48:31–36.

Jirka, S. and Bröring, A. (2009). OGC Sensor Observable Registry Discussion Paper. Reference number: OGC 09-112.

Jirka, S. and Nüst, D. (2010). OGC Sensor Instance Registry Discussion Paper. Reference number: OGC 10-171.

Korteweg, P., Marchetti-Spaccamela, A., Stougie, L., and Vitaletti, A. (2007). *Data aggregation in sensor networks: Balancing communication and delay costs*. Springer.

Lassila, O. and Swick, R. R. (1999). Resource Description Framework (RDF) Model and Syntax Specification. [online] http://www.w3.org/TR/PR-rdf-syntax/ [accessed on December 8th, 2015].

Lebo, T., Sahoo, S., and McGuinness, D. (2013). PROV-O: The PROV Ontology. [online] http://www.w3.org/TR/prov-o/ [accessed on December 11th, 2015].

Manola, F., Miller, E., and McBride, B. (2014). W3C RDF Primer. [online] http://www.w3.org/TR/rdf11-primer/ [accessed on December 9th, 2015].

Missier, G. A. (2015). Towards a Web application for viewing Spatial Linked Open Data of Rotterdam. Master's thesis, Delft University of Technology.

Moir, E., Moonen, T., and Clark, G. (2014). What are Future Cities: Origins, Meanings and Uses.

Nebert, D., Whiteside, A., and Vretanos, P. (2007). Opengis catalogue services specification.

OWL working group (2012). Web Ontology Language (OWL). [online] http://www.w3.org/2001/sw/wiki/OWL [accessed on December 18th, 2015].

Percivall, G. (2015). OGC Smart Cities Spatial Information Framework. OGC Internal reference number: 14-115.

Perry, M. and Herring, J. (2011). GeoSPARQL - A Geographic Query Language for RDF Data. [online] http://www.opengeospatial.org/standards/geosparql [accessed on December 9th, 2015].

Price Waterhouse Coopers (2014). Sensing the future of the Internet of Things. [online] https://www.pwc.com/us/en/increasing-it-effectiveness/assets/future-of-the-internet-of-things.pdf [accessed on December 18th, 2015].

Pschorr, J. K. (2013). SemSOS: an Architecture for Query, Insertion, and Discovery for Semantic Sensor Networks. Master's thesis, Wright State University.

Sheth, A., Henson, C., and Sahoo, S. S. (2008). Semantic Sensor Web. *IEEE Internet Computing*, 12(4):78–83.

Stasch, C., Autermann, C., Foerster, T., and Pebesma, E. (2011a). Towards a spatiotemporal aggregation service in the sensor web. Poster presentation. In *The 14th AGILE International Conference on Geographic Information Science*.

Stasch, C., Schade, S., Llaves, A., Janowicz, K., and Bröring, A. (2011b). Aggregating linked sensor data. In Taylor, K., Ayyagari, A., and de Roure, D., editors, *Proceedings of the 4th International Workshop on Semantic Sensor Networks*, page 46.

Stasch, C., Scheider, S., Pebesma, E., and Kuhn, W. (2014). Meaningful spatial prediction and aggregation. *Environmental Modelling & Software*, 51:149–165.

Theunisse, I. A. H. (2015). The Visualization of Urban Heat Island Indoor Temperatures. Master's thesis, TU Delft, Delft University of Technology.

van der Hoeven, F., Wandl, A., Demir, B., Dikmans, S., Hagoort, J., Moretto, M., Sefkatli, P., Snijder, F., Songsri, S., Stijger, P., et al. (2014). Sensing Hotterdam: Crowd sensing the Rotterdam urban heat island. *SPOOL*, 1(2):43–58.

Van der Hoeven, F. D. and Wandl, A. (2015). Hotterdam: How space is making Rotterdam warmer, how this affects the health of its inhabitants, and what can be done about it. Technical report, TU Delft, Faculty of Architecture and the Built Environment.

W3C Semantic Sensor Network Incubator Group (2011). Semantic Sensor Network Ontology. [online] http://www.w3.org/2005/Incubator/ssn/ssnx/ssn [accessed on December 9th, 2015].

Wang, M., Perera, C., Jayaraman, P. P., Zhang, M., Strazdins, P., and Ranjan, R. (2015a). City Data Fusion: Sensor Data Fusion in the Internet of Things.

Wang, X., Zhang, X., and Li, M. (2015b). A Review of Studies on Semantic Sensor Web. *Advanced Science and Technology Letters*, 83:94–97.

Xiang, L., Luo, J., and Rosenberg, C. (2013). Compressed data aggregation: Energy-efficient and high-fidelity data collection. *Networking, IEEE/ACM Transactions on*, 21(6):1722–1735.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., and Zorzi, M. (2014). Internet of things for smart cities. *Internet of Things Journal, IEEE*, 1(1):22–32.

COLOPHON

This document was typeset using LaTeX. The document layout was generated using the `arsclassica` package by Lorenzo Pantieri, which is an adaption of the original `classicthesis` package from André Miede.