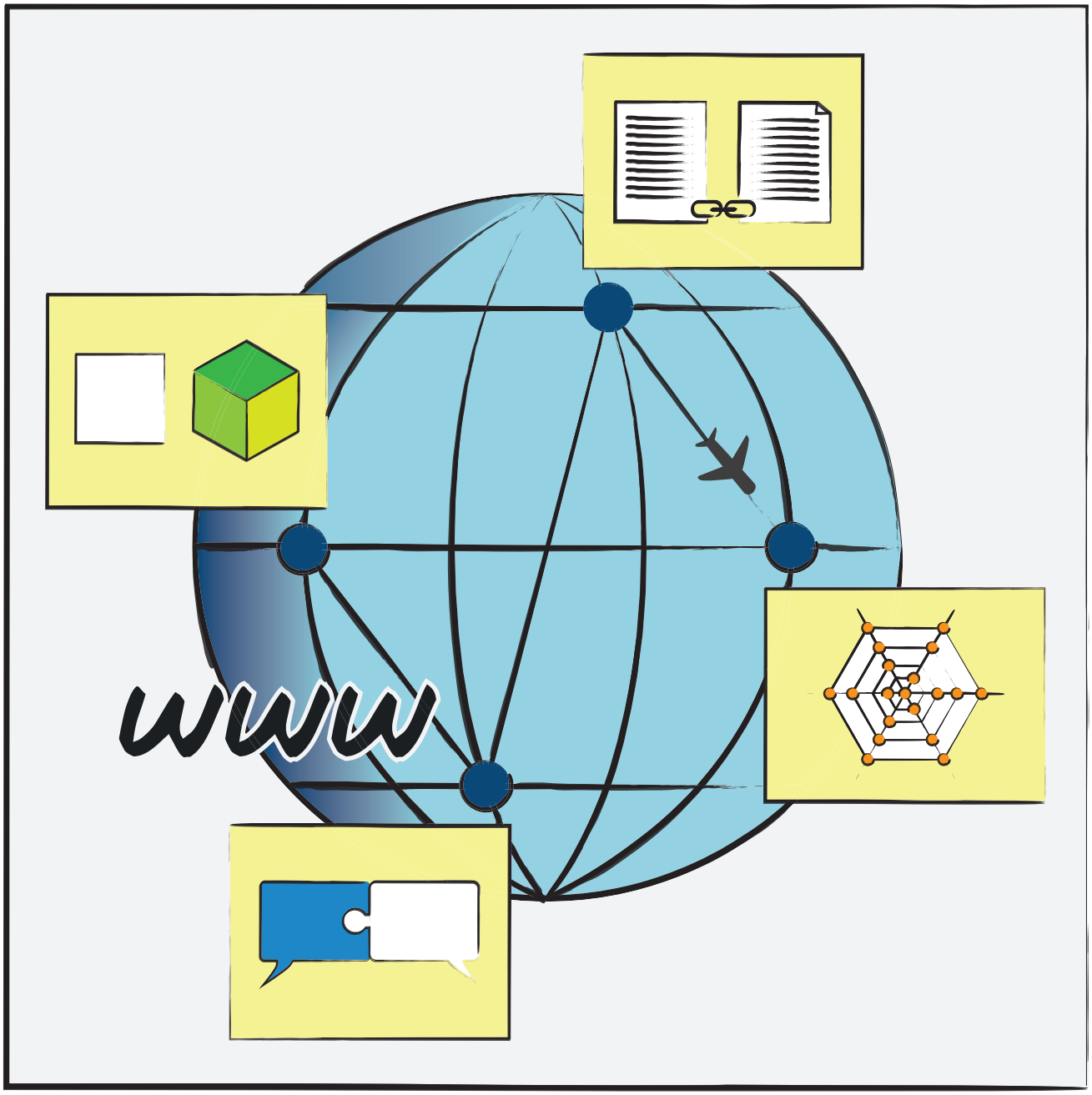


Geospatial Data on the Web

Linda van den Brink

Geospatial Data on the Web

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology

by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on

Tuesday 4 December 2018 at 10:00 o'clock

by

Linda Elisabeth VAN DEN BRINK

Doctorandus in de Algemene Letteren, Utrecht University, the Netherlands

born in Heerhugowaard, the Netherlands

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof.dr. J.E. Stoter	Delft University of Technology, promotor

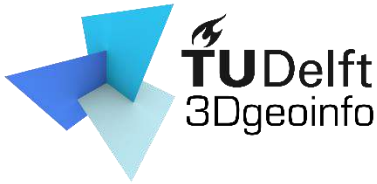
Independent members:

Prof.dr. W.M de Jong	Delft University of Technology
Prof.dr.ir. M.F.W.H.A Janssen	Delft University of Technology
Prof.dr. T.H. Kolbe	Technische Universität München
Dr. H. Ledoux	Delft University of Technology
Dr.ir. E.J.A Folmer	University of Twente

Other member:

Dr. A. Perego	European Commission, Joint Research Centre

The creation of this dissertation was partly supported by Geonovum.



Geospatial Data on the Web

Linda van den Brink

October 2018

Contents

Acknowledgements

Getting your PhD is not easy. A lot of people helped me in my quest. A special thank you ....

First of all to my parents. My mother, who convinced me I could be anything I aspired to. My father, who kindled and stimulated my interest in computers, being an early adopter of home computers himself in the '80s.

To my brothers, who both got their PhDs years ago. I always found motivation in their achievements.

To my elementary school, the Van Nassauschool in Bergen, which featured an experimental learning system that allowed me to gain knowledge at my own pace and that taught me to work independently at an early age.

To Hans Voorbij, who taught Computer & Letteren in Utrecht when I graduated there and who was one of my supervisors. He introduced to me the joy of working at the intersection of language, ICT, and humans.

To my ex-colleague Marcel Reuvers, who gave me a role in his innovative projects, which provided the feeding ground for my research.

To my employer Geonovum in general, and to Rob van de Velde and Ruby Beltman specifically, who gave me the opportunity to combine my job with getting my PhD and who supported me throughout the years it took me to get here. Special thanks to my colleagues Paul Janssen and Wilko Quak who were always ready and willing to co-author a publication with me - and Paul once more for always being willing to help my thoughts get unstuck.

To my husband Barry, for keeping me sane by frequently asking me out on ridiculously long trailruns and for accompanying me on several of them.

To all my publications' co-authors for the excellent cooperation.

To Sisi Zlatanova and Jantien Stoter, who co-authored my first two publications, chapters 2 and 3 in this dissertation, and who inspired me to do this: "If you publish two more articles you can get your PhD!" And again to Jantien Stoter for being my patient and ever helpful supervisor.

Linda van den Brink
Amersfoort, October 2018

Part I

Introduction, research questions and methodology

Chapter 1

Introduction

1.1 Reuse of geospatial data via the web across communities

Geospatial data is an increasingly important information asset for decision-making, from simple every day decisions like where to park your car, to national and international policy on topics like infrastructure and environment. The term *'geospatial'* refers to a location on earth. While *'spatial'* is often used as a synonym for 'geospatial', 'spatial' is strictly speaking a broader term: it could refer to another planet, an imaginary world, a section of a person's body, a location on a computer screen, or any other space. But both geospatial and other spatial data are about the location of things, i.e. about where things are. And locational data is important. A lot of geospatial data is created, for example, as part of governmental processes and nowadays, also disseminated as open data. Examples are data on addresses, buildings, zoning plans, and topographic objects.

Because of the location aspect, geospatial data is often the linking pin between different datasets and therefore important for data integration (Auer et al., 2009). The integration of geospatial data from different sources offers possibilities to infer and gain new information. Therefore, worldwide, governments at different levels have put a lot of effort into disseminating geospatial data via the web for wide reuse. Figure 1.1 shows an example of geospatial data which is published through a data portal.

Within the geospatial community, work on standards and infrastructure for geospatial data dissemination via the web, leading to the so-called *"Spatial Data Infrastructure"* (SDI), has been ongoing since the 1980s (Crompvoets et al., 2004; Maguire and Longley, 2005). SDIs facilitate the access and use of spatial data, often on a national scale, by providing a technical access network

Figure 1.1: The Dutch geoportal "PDOK" showing large-scale topographic data.

and complementary services. They are based on a service-oriented architecture (SOA), in which existing resources are documented using dataset-level metadata, published in catalogs, which are accessible through web services. Other web services within an SDI provide, for example, online viewing of spatial data as web maps and downloading spatial data. All these web services are based on mature standards of the Open Geospatial Consortium (OGC).

Note that the SDI can be said to disseminate both geospatial data and *geospatial information*; information being "data presented in readily comprehensible form to which meaning has been attributed within the context of its use." (Reitz, 2004). An SDI could serve, for example, both raw measurements of soil and soil maps, the latter being interpretations of the raw soil data. The boundary between data and information is often fuzzy. In my thesis I use the term 'geospatial data' unless 'information' is meant explicitly.

The goal of an SDI is to publish spatial data for reuse. Once governmental geospatial data starts to become available as open data, the possibility for reuse is there in theory. However, a government announcing that a specific geospatial data set is "open" is not sufficient to make people actually reuse the data. The first step is making the data, not just the metadata, available on the internet, instead of potential users having to call a specific person in order to actually get the data. Such a cumbersome process required to obtain the data might be a big hurdle. In addition, publishing the data somewhere on the internet is not sufficient either: it is not likely that the data will be

reused if people have no way of knowing where the data can be found or that it even exists, or if the provided data format is largely unknown. In short, there are a lot of aspects of data dissemination that have to be addressed before open data is actually in a good position for getting reused.

Wilkinson et al. (2016) describe these aspects in their *FAIR principles* as findability, accessibility, interoperability, and reusability. The FAIR principles were defined with scientific data in mind, but can be applied more broadly to open data. In the geospatial domain, the SDI has addressed these aspects (before these FAIR principles had been defined) by using information models for describing data structure and semantics for interoperability; dataset descriptions (also known as metadata) for findability and reusability; and web services for accessibility (Crompvoets et al., 2004). The result is that geospatial data has become much more accessible over the years, as national and international SDIs have been implemented. An example is the European environmental SDI established by the INSPIRE directive (INSPIRE, 2007), which was approved by the Council of Ministers and the European Parliament in November 2006 (Masser et al., 2008) with the aim of sharing geospatial data throughout the European Union to support environmental policies.

However, several challenges still need to be addressed for the FAIR use of geospatial data outside the domains for which it was created—ultimately also outside the traditional geospatial sector, where a lot of new uses of geospatial data are potentially possible. This statement applies to the Netherlands—most of my research was embedded in the national context of the Netherlands, where the problems I studied surfaced. However, spatial data—including more and more 3D data—is created and used all over the world, and a lot of other countries worldwide have an SDI in place and may experience the same problems. From reactions to and reuse of my research this seems to be the case. Notwithstanding the national context of much of my work, I emphasised the application of international standards to the solution of interoperability problems; and in some cases my work is a part of the international development of standards.

1.2 Problem description

First problem: *Lack of standard hinders reuse of geospatial data: example of 3D data.*

A general foundation of my work is the common knowledge in the geospatial domain that interoperability between systems is required to make reuse

of data possible, and standards are able to realise this interoperability. Otherwise, there would only be closed software systems, which only support their own proprietary formats, making data exchange and reuse impossible. However, geospatial data is often initially created for a specific use case which makes it hard to reuse the same data in another context, even if it is shared in an interoperable format. A lot of countries have, for example, a well-established process for creating and updating topographic data. Historically, this data is two-dimensional and created for the use case of topographic maps (see Figure 1.1), i.e. navigation and orientation. However, because of its rich semantics it could also be reused for other purposes, such as maintaining public space. This is why many countries have started to define data models underlying the topographic maps at different scales such as the Dutch domain model for large-scale topography, *Informatiemodel Geografie* (IMGeo) (Geonovum, 2013b,a), and the German AFIS-ALKIS-ATKIS-Modell (AdV, 2009). These topographic data models describe semantics as well as geometric representations of objects that occur in the real world, like buildings, water and roads. The fact that the data is described in this way, instead of it just being lines, symbols and colours on paper or in a computer system, has made its reuse possible.

For three-dimensional data (3D) such a national standard did not exist when I started this study, while there was a growing need for 3D data. There was a standard topographic data model, but it only supported 2D geometries. Consequently, 3D data was collected in an ad hoc manner, for specific projects and by specific organisations, and there was hardly any reuse of 3D data. This was an unwanted situation since the acquisition, i.e. creation of spatial, topographic data is expensive. It involves creating and then interpreting raw data from sources like surveying measurements, laser images or aerial photography. Instead of different (governmental) organisations doing this multiple times for the same land area and kinds of objects, it is much more efficient to do this only once, collect the data and make it available in an interoperable way so that it can be reused many times, at the same time also centralising the process for updating this data.

Second problem: *Independently developed domain models model similar concepts in different ways, which makes reuse of data in other domains difficult.*

A second important hurdle for interoperability of geospatial data is the fact that different data models, and datasets, developed for specific purposes, often overlap or offer complementary information about the same objects, while the data cannot easily be combined because the objects and concepts

are defined in slightly different ways. In the Netherlands, for example, a whole family of data models are based on the Base Model Geo-Information (NEN 3610:2011 (NEN, 2011)). This national standard describes geographic concepts and establishes a standard modelling method based on the ISO 191xx series of standards (specifically: ISO 19103:2015 (ISO, 2015a), ISO 19107:2003 (ISO, 2003), ISO 19109:2015 (ISO, 2015b), ISO 19110:2005 (ISO, 2005), ISO 19131:2007 (ISO, 2007)). It contains a generic semantic Unified Modeling Language (UML) model with definitions of the most common, shared concepts in the geo-domain. NEN 3610 thus forms a common base for domain specific information models. They all follow the modelling rules prescribed by NEN 3610 and are based on the same high-level concepts defined in the NEN 3610 semantic model like terrain, road and water. However, it appeared that this does not assure these domain specific information models are compatible from a semantic standpoint. Each model is developed for a specific sector and use case; NEN 3610 does not prevent them from overlapping or contradicting each other, and this actually does frequently happen in reality. Domain models that are not harmonised severely limit the reuse of the geospatial data outside the original use case for which it was created.

This is a problem observed in the Netherlands, but there is a bigger picture. Since NEN 3610 was established as a standard, extensive international standardisation work was done on geospatial concepts, resulting in semantic standards for geospatial data and information exchange; e.g. the INSPIRE thematic domain models, which model all kinds of aspects of the world around us with a focus on the environment, and CityGML (Gröger et al., 2008, 2012), a standardized data model and XML-based format for the storage and exchange of virtual 3D city models. National data models need to be harmonised with these international standards, otherwise the data cannot be used in international contexts.

Harmonising data models to obtain semantic interoperability is an important solution to the above mentioned problem of independently developed data models that do not align. But harmonisation is not an easy task. Although tooling can help to discover mappings between datasets or data models (Euzenat and Shvaiko, 2013), humans are needed to identify whether a difference between data models should be harmonised or not. In some cases, such a difference serves a specific purpose, while in other cases the difference is unintentional but hampers reuse of data. In the latter case, improving semantic interoperability can truly lead to data integration opportunities: adjusting data definitions can be achieved and work processes can be optimized, e.g. the same data is no longer created twice.

Realising semantic interoperability is an important step. That is, relating the data to well-known semantic domain standards and harmonising the

meaning of data can improve spatial data interoperability across domains or communities within the geospatial sector; but it is not enough to fully enable reuse outside of the geospatial sector.

Third problem: *Geospatial data disseminated via SDI methods cannot be found, accessed and used by non-geospatial experts.*

For FAIR use of data outside the geospatial sector more is needed than just disseminating the (harmonised) data via SDI based methods. A limitation of SDIs is that only geospatial experts are able to find, access and use such data, because 1) knowledge of the existence of specific portals, containing catalogues of geospatial datasets, is necessary: only people who know these portals can go there and browse or search for geospatial data, and 2) a high level of specific technical expertise, e.g. knowledge of geospatial standards (i.e. Open Geospatial Consortium standards), is necessary to access and understand the descriptions of the dataset, access the web services and use the data (Taylor and Parsons, 2015). To achieve reuse of geospatial data outside the geospatial domain, the distribution should not be limited to the methods of the SDI, but instead should be based on general standards and methods for data publication from a much broader community, the World Wide Web.

When looking at these general standards, linked data is an obvious candidate. Linked data 'provides a publishing paradigm in which not only documents, but also data, can be a first class citizen of the Web' (Heath and Bizer, 2011). The linked data paradigm is defined in a set of open, general standards, developed by the World Wide Web Consortium (W3C), which describe how data should be published on the web and interlinked using hyperlink technology. However, if and how linked data standards can be utilised to publish geospatial data, and how this can be done with existing SDI standards and practice as a basis, is not a trivial question (Hart and Dolbear, 2016), as the standards underlying linked data are different from those underlying the SDIs in the geospatial domain, having evolved in different communities. One problem is that the data in an SDI is commonly exchanged using the Geography Markup Language (GML) data format, which is based on Extensible Markup Language (XML), with standardised support of geometry. In linked data, Resource Description Format (RDF), a W3C standard for the publication and interlinking of data on the web, is used for data exchange. In order to publish geospatial data as linked data, a conversion to RDF is necessary. Moreover, RDF should be able to express geometries; however, there is no single standard way of doing this: a geometry expression method has to be selected from several options. Also, RDF

requires Uniform Resource Identifiers (URI) for each resource i.e. each object within a dataset; a URI strategy is needed to ensure these URIs meet several requirements for linking data, for example persistence. In addition, data models in the geospatial community are not described using Web Ontology Language (OWL), as is done in the linked data community. Instead, within SDIs, information models are commonly defined using the Unified Modeling Language (UML), following a method described in a series of ISO standards (ISO 19109 in particular). Both modelling languages have different underlying paradigms, which leads to problems when expressing information models originally described using one of these, using the other (Kiko and Atkinson, 2008; Cox, 2013).

Fourth problem: *Geospatial data disseminated with linked data technologies are not easy to use by users who are not linked data experts.*

Linked data, while broad in its applicability, is somewhat of a niche set of standards. Using the linked data paradigm to disseminate spatial data as the only alternative of the SDI would still keep potential users away from geospatial data, who experience linked data as an impediment to ease of use. A last challenge to improve reuse of geospatial data outside the geospatial sector is therefore to practice web architecture without mandating a specific metamodel such as linked data (Wilde, 2007), thereby reaching a larger potential audience.

An important set of present-day users can be called "data users": web developers, data journalists etc. who use different kinds of data, including geospatial data, directly to create applications or visualisations that supply information to end users (citizens). In order to achieve the wide re-use of geospatial data across communities, data should be easily accessible by these data users. In order for them to be able to further process the data for end users, geospatial data must be published on the web using well-known and widely used web standards, some of which are related to linked data, but not all. This brings another challenge, since the SDI was not designed with the web of data in mind: how to broaden SDIs towards the Web of Data. Furthermore, which web standards are optimal for publication of geospatial data, and how they should be applied, needs further research.

1.3 Research objective

In order to integrate data from different sources, standards are necessary to achieve at least a minimum of interoperability. In the geospatial domain, there are mature standards already and (expert) users are able to find, access

and exchange geospatial data through what is called the SDI. Geonovum—my employer—is responsible for realising this SDI in the Netherlands. Data models and standards play an important role in facilitating the exchange of data and information within sectors, but also between different sectors within the geospatial domain.

As explained in the Problem Description, even though a lot of geospatial data is open and published using the SDI method, there are further impediments to its reuse. The main objective of my research is to overcome these impediments, thereby improving reuse of geospatial data across communities via the web.

1.4 Research questions

The main question of my research can be formulated as:

How to reuse geospatial data, from different, heterogeneous sources, via the web across communities?

I have formulated four questions to cover different aspects of this problem, related to the four problems identified above:

1. How to define a national standard for large-scale topographic objects in 3D for wide re-use of once collected 3D data to solve current ad hoc acquisition and use of such data?

2. How can semantic interoperability between different kinds of geospatial datasets that have been created for different purposes best be achieved?

3. How to apply the Linked Data paradigm to disseminate geospatial data outside the traditional geospatial data sector?

4. How to apply general Web based principles to improve the discoverability and accessibility of spatial data?

I have published five peer-reviewd journal papers, which address these four research questions via several sub questions. These sub questions, divided over the five publications, are introduced in the next section in which the research approach, methodology and main achievements of each publication are summarised.

1.5 Research approach and methodology

The research can be divided in four parts related to the four research questions. The first part studies the definition and establishment of a national 3D standard for reusing 3D data across the whole geospatial information chain, from 3D data acquisition to use. The second part focuses on obtaining semantic interoperability by harmonising information models. The third part studies the use of linked data technologies to achieve better linking mechanisms and higher reusability through the use of general web standards. And finally, the fourth step moves beyond linked data by studying current practice for publishing spatial data according to web architecture principles and compiling a best practice based upon this analysis.

This section describes per part my research approach and methodology and summarises the main achievements. All of the research described in this thesis is qualitative in nature.

1.5.1 Defining a national 3D standard

3D data is often collected ad hoc, in specific projects and by specific organisations. In practice, reuse of once collected 3D geospatial data hardly takes place. One of the main causes for ad hoc collection of 3D data is the lack of 3D standards. The first part of my research focuses on improving interoperability by defining and establishing a national 3D standard for large scale topography called "IMGeo", aligned to the international 3D standard "CityGML". IMGeo is the underlying standard of a national 'basic registry' (*Basisregistratie Grootschalige Topografie*, BGT). The provision and reuse of data within this basic registry is regulated by law.

The Information Model Geography (IMGeo), originally focussing on 2D data (Geonovum, 2007), describes the semantics of topographic objects commonly found on large scale maps, i.e. scale 1:1000 – 1:2000. Data about these objects is collected and maintained by municipalities, water boards, provinces, ProRail (the manager of Dutch railway network infrastructure) and Rijkswaterstaat (Dutch Ministry for infrastructure). The mandatory core model contains object definitions for large-scale representations of roads, water, land use, land cover, bridges, tunnels, etc. Specific governmental organisations must collect this data and other governmental organisations must reuse it. The optional part of IMGeo allows further division of these objects into parts suitable for maintenance, and contains definitions for all kinds of city furniture and other non-mandatory classes. IMGeo prescribes 2D point, curve or surface geometry for all objects. The data providers are required by law to provide their objects that fall under the definitions of the IMGeo 2.0

core to the BGT national basic registry where they are available for reuse.

The IMGeo standard provides a semantic model for 2D large scale topography. In my research I have contributed to the development of a national standard for 2.5D and 3D geospatial data within the IMGeo framework: 2D being simple point, line and polygon representations of objects on a flat surface as on a typical map; 2.5D being points, lines and polygons with a height component added, e.g. such that the polygon of a bridge would be elevated above the flat surface; and 3D being full volumetric shapes in the form of solids, positioned on a surface which models the elevation of the terrain (often called a digital terrain model). For this, IMGeo needed to be aligned with international 3D standards for topographic data in order to preserve the semantics while extending 2D data into the third dimension. The main research question was:

How to define a national standard for large-scale topographic objects in 3D for wide re-use of once collected 3D data to solve current ad hoc acquisition and use of such data?

I researched the question by conducting an exploratory study. The main principle of CityGML-IMGeo is the reuse of CityGML concepts, that is, IMGeo classes are remodelled in accordance with CityGML, as this international 3D standard provides a good basis for a national 3D standard. To comply with this principle, the concepts from IMGeo needed to be mapped to CityGML concepts. I carried out the task of mapping IMGeo to CityGML as a desk study, the result of which was reviewed and approved by a working group of IMGeo stakeholders.

Next, I experimented with implementing this mapping in a UML class model. Based on the mapping, IMGeo classes, as they are called in UML, are defined as a specialisation (subtype) of the relevant CityGML generic class. New classes have been added if they were present in IMGeo but missing in CityGML, also modelled as a CityGML specialisation. Added classes include constructions related to water management, separating objects like walls and fences and other constructions which are not quite buildings, like storage tanks or wind turbines. Pending a proposal to add constructions like this to the CityGML standard in v3.0, these classes have been added with one superclass called OtherConstruction. Some IMGeo classes were remodelled in order to better fit with CityGML. These are Vegetation for modelling any vegetation-related concept (in IMGeo these were divided over several classes) and AuxiliaryTrafficArea for road segments which are not used for traffic.

During this work, I obtained the insight that clearer guidelines for extending CityGML are needed. The CityGML standard describes an “Application

Domain Extension" (ADE) mechanism that should be used to extend the standard with additional concepts; however, this is only described in the context of GML Application Schemas, which is not compatible with UML modelling. I have conducted an analysis of several approaches to model application specific concepts of an ADE in UML and intensively discussed these with CityGML and UML modeling experts within the German Special Interest Group 3D (SIG3D) modelling subgroup and in e-mail discussions with SIG3D members and OGC CityGML working group members. The agreed modeling method was implemented in a software tool, and then tested by automatically deriving GML and CityGML compliant XML Schemas from the model. Based on these steps I have proposed the preferred modelling approach for modelling CityGML ADEs in UML.

The result of our work on the integration of IMGeo and CityGML is a national standard which supports both *2D, 2.5D and 3D* representations of large scale topography objects according to geometric and semantic principles of CityGML. The standard is established as such by the Dutch national government. Based on the experiences of developing this CityGML–IMGeo standard, we defined a framework for extending CityGML for national purposes. In addition, I submitted a number of change requests to the Open Geospatial Consortium (OGC). Some of these have led to revisions in CityGML 2.0, others have been approved for version 3.0. Finally, the modelling approach we proposed for modelling CityGML ADEs in UML has been published as an OGC Best Practice.

1.5.2 Harmonising models

The second part of my research focuses on obtaining semantic interoperability by harmonising information models. This is related to the research field of ontology matching, a good and comprehensive overview of which is given by Euzenat and Shvaiko (2013). Information models fall under their broad definition of 'ontology': 'a set of assertions that are meant to model some particular domain'. In simple terms an ontology provides a vocabulary describing some domain and specifying the meaning of terms from that domain. Ontology matching, which concerns the mapping of concepts from different ontologies based on their meaning (i.e. terms that mean the same thing are mapped), has been studied since the 1980s; most research has focussed on automatic matching, where computer algorithms are used to find correspondences between different ontologies. These algorithms have evolved to an advanced state.

However, this was not the focus of my research; the goal was to obtain semantic interoperability to improve the actual reuse of spatial data, and

therefore I needed to find areas where similarities existed but semantic problems prevented interoperability, and to solve these problems by identifying the best way to adjust the ontologies in question. The task of finding potential harmonization issues was computer-assisted, but human experts were needed both to identify the areas where harmonization would be most beneficial, and to assess how each semantic problem could best be solved.

The main question in this part of my research was:

How can semantic interoperability between different kinds of geospatial datasets best be achieved?

This part of my research addresses a problem with a much wider scope than just one standard: semantic interoperability between Dutch information models at a national level as well as between these models and INSPIRE themes. It requires the integration of information models from different domains and the analysis of their similarities and differences. The method of research was again exploratory. IMGeo was taken as a starting point because of its semantic overlap with a lot of other domain models. We compared it to several other Dutch data models in a desk study. Next, we conducted a series of interviews with the domain experts who were responsible for the domain models. Several semantic problems were identified in this phase. These include: concepts with the same name being used in different domains with different meanings; the same concepts being duplicated in several models, and other forms of overlap. The next step was experimentation with tooling that would aid humans in discovering the semantic problems in more detail. I integrated all Dutch geo-information models and INSPIRE themes into one software environment suited for the analysis of semantic similarity of concepts from different models. I then tagged all concepts in these information models with keywords based on two broad classifications. The tool could then be used to group the concepts based on these keywords. Thus, related concepts from different domains could be found. The groups of related concepts were visualised and used in a brainstorm session with domain experts, selected based on their knowledge of the domains described by the domain models that were part of the study, to discover and analyse overlaps and discrepancies.

This study of IMGeo and its semantic overlap with other information models resulted in several proposed changes to both IMGeo and other Dutch information models, some of which have already been realised, while others will be addressed in future versions of the models. With this step multiple semantic discrepancies that were hindering data reuse have been solved, and the geo-information models are thus becoming more harmonised resulting in

higher interoperability.

1.5.3 Publishing data via Linked Data principles

To address the findability, accessibility and reusability of geospatial data outside the geospatial data sector, in a third step I conducted an exploratory study on linked data as a paradigm that might be employed to disseminate both spatial semantics and geospatial data to new groups of data users. Linked data provide an alternative route for dissemination of spatial information as compared to the traditional Service Oriented Architecture (SOA)-based SDI approach. The linked data approach makes linking to and from any data over the Web possible, and the semantics of the data can be made clear by integrated use of ontologies or vocabularies, which can also be interrelated. Therefore, it is very promising for achieving semantic interoperability of geo-information both within the geo-domain and across other domains.

The main research question was thus:

How to apply the Linked Data paradigm to disseminate geospatial data outside the traditional geospatial data sector?

Since geospatial data is already structured in a standardised and defined way using Geography Markup Language (GML), it is possible to standardise transformation of this data to linked data, in particular the Resource Description Framework (RDF). In the context of the Platform implementation Linked Open Data, an initiative by Geonovum and other organisations, I participated in a pilot project where we experimented with such an automated transformation, leveraging the RDF-like object-property structure of GML. Based on a literature study we evaluated ways of describing geometries in RDF resulting in the selection of OGC GeoSPARQL and designed a URI pattern for the data object identifiers. Then, we experimented with automated transformation of a UML model describing the data model to an RDF Schema (RDFS)/OWL ontology. Because the resulting ontology was not linked to existing OWL ontologies, while in OWL interoperability is achieved via the re-use of vocabularies, we created a method of annotating the UML beforehand with mappings to existing OWL ontologies, and tested this method in our experiment.

Results of this part of the research include a method for semi-automated UML to OWL conversion, which formed the basis of my contribution to the development of a method to derive OWL vocabularies from the INSPIRE UML models, and ultimately for the publication of INSPIRE data as geospatial linked data.

1.5.4 Web of data

In order to reach the wider audience of web "data users", who have not embraced the linked data paradigm, the fourth part of my research focussed on making geospatial data part of the web of data, using widely known and accepted general web standards and principles. The main question related to this was:

How can we apply general Web based principles to improve the discoverability and accessibility of geospatial data?

While the OGC has standardised solutions for publishing geospatial data and services in an interoperable way, the World Wide Web Consortium (W3C) has developed standards for the web of data, with an emphasis on the development of methods that can be rapidly taken up by application software developers without special expertise. These organisations started a joint Working Group on Spatial Data on the Web, of which I am a member, and which did some crucial work in this area. I contributed in particular, as an editor, to the effort to describe the best practices for publishing spatial data on the Web. The best practices were compiled based on use cases and current good practices, both gathered by members of the working group. This was followed by an analysis of the current practices that were found, and by discussions within the working group, to establish which practices could be labelled 'best practice' and in which areas gaps existed. A set of principles that guide the selection of best practices was created based on discussions within the working group. The result describes best practices that are employed to enable publishing, discovery and retrieving (querying) spatial data on the Web, and identifies some areas where a best practice has not yet emerged.

Results of the fourth part include the Spatial Data on the Web Best Practice, which was published as a W3C Note and an OGC Best Practice. When implemented, the guidelines from this document make it easier to discover, interpret and use geospatial data for data users in general – not just geospatial experts. Implementations of the guidelines are in progress, for example at the Dutch Geoportal and in OGC Web Feature Service 3.0 (Portele and Vretanos, 2018).

1.6 Overview of the dissertation

This dissertation consists of the five original peer-reviewed journal papers that have been published during the research and are published unchanged

in this thesis, except for occasional footnotes to supply current information about outdated statements. The dissertation is divided into parts corresponding to the research questions formulated in Section 1.4. Part II, "Definition and establishment of a national 3D standard" contains two articles, in Chapters 2 and 3; the first of these addressing the definition of a national 3D standard as a whole (van den Brink et al., 2013a), while the latter focusses on the UML-based approach we used (van den Brink et al., 2013b). Part III, "Semantic harmonisation", addresses semantic interoperability between different datasets (van den Brink et al., 2017) in Chapter 4. Part IV, "Geospatial Linked Data", contains Chapter 5 which explores automated conversion of geospatial data to linked data (van den Brink et al., 2014). Part V, "Web of Data", describes the best practices for publishing, retrieving and using spatial data on the web (van den Brink et al., 2018), in Chapter 6. Finally, Part VI, "Discussion, Conclusion and Future work" contains Chapter 7, which discusses the developments since the five original articles were published, and Chapter 8, which contains the conclusions and suggestions for future work.

Bibliography

AdV. AFIS/ALKIS/ATKIS Dokumentation zur Modellierung der Geoinformationen des amtlichen Vermessungswesens. Available online: http://www.adv-online.de/AAA-Modell/, 07 2009.

Sören Auer, Jens Lehmann, and Sebastian Hellmann. Linkedgeodata: Adding a spatial dimension to the web of data. In *International Semantic Web Conference*, pages 731–746. Springer, 2009.

Simon Cox. An explicit OWL representation of ISO/OGC Observations and Measurements. In *SSN@ ISWC*, pages 1–18, 2013.

Joep Crompvoets, Arnold Bregt, Abbas Rajabifard, and Ian Williamson. Assessing the worldwide developments of national spatial data clearinghouses. *International Journal of Geographical Information Science*, 18(7): 665–689, 2004.

Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer Science & Business Media, 2013.

Geonovum. Informatiemodel Geografie, 10 2007.

Geonovum. Basisregistratie grootschalige topografie Gegevenscatalogus BGT 1.1.1. Available online: https://docs.geostandaarden.nl/bgt/def-imgcbgt111-20130700/doc.pdf, 07 2013a.

Geonovum. Basisregistratie grootschalige topografie Gegevenscatalogus IMGeo 2.1.1. Available online: https://docs.geostandaarden.nl/bgt/def-imgcimgeo211-20130700/doc.pdf, 07 2013b.

Gerhard Gröger, Thomas H Kolbe, Angela Czerwinski, and Claus Nagel. OpenGIS city geography markup language (CityGML) encoding standard, version 1.0.0. Available online: http://portal.opengeospatial.org/files/?artifact%5Fid=28802 (accessed 11 February 2014), 2008.

Gerhard Gröger, Thomas H Kolbe, Claus Nagel, and Karl-Heinz Häfele. Ogc® city geography markup language (CityGML) encoding standard, Version 2.0, 2012.

Glen Hart and Catherine Dolbear. *Linked data: a geographic perspective*. CRC Press, 2016.

Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

INSPIRE. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Available online: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002 (accessed 2-February-2016), 2007.

ISO. ISO 19107:2003 Geographic information – Spatial schema., 2003.

ISO. ISO 19110:2005 Geographic information – Methodology for feature cataloguing., 2005.

ISO. ISO 19131:2007 Geographic information – Data product specifications., 2007.

ISO. ISO 19103:2015 Geographic information – Conceptual schema language., 2015a.

ISO. ISO 10109:2015 Geographic information – Rules for application schema., 2015b.

Kilian Kiko and Colin Atkinson. A detailed comparison of UML and OWL. *Reihe Informatik*, TR-2008(4), 2008.

David J Maguire and Paul A Longley. The emergence of geoportals and their role in spatial data infrastructures. *Computers, environment and urban systems*, 29(1):3–14, 2005.

Ian Masser, Abbas Rajabifard, and Ian Williamson. Spatially enabling governments through SDI implementation. *International Journal of Geographical Information Science*, 22(1):5–20, 2008.

NEN. NEN3610 Basismodel Geo-informatie - Termen, definities, relaties en algemene regels voor de uitwisseling van informatie over aan de aarde gerelateerd ruimtelijke objecten. Available online: https://www.nen.nl/NEN-Shop/Norm/NEN-36102011-nl.htm, 2011.

Clemens Portele and Panagiotis (Peter) A. Vretanos. OGC® Web Feature Service 3.0 - Part 1: Core. Draft. Available online: https://cdn.rawgit.com/opengeospatial/WFS%5FFES/master/docs/17-069.html (accessed 2018-04-18), 2018.

Joan M Reitz. *Dictionary for library and information science*. Libraries Unlimited, 2004.

Kerry Taylor and Ed Parsons. Where is everywhere: bringing location to the web. *IEEE Internet Computing*, 19(2):83–87, 2015.

L. van den Brink, P. Janssen, W. Quak, and J. Stoter. Linking spatial data: automated conversion of geo-information models and GML data to RDF. *International Journal of Spatial Data Infrastructures Research*, 9:59–85, 2014.

Linda van den Brink, Jantien Stoter, and Sisi Zlatanova. Establishing a national standard for 3D topographic data compliant to CityGML. *International Journal of Geographical Information Science*, 27(1):92–113, 2013a.

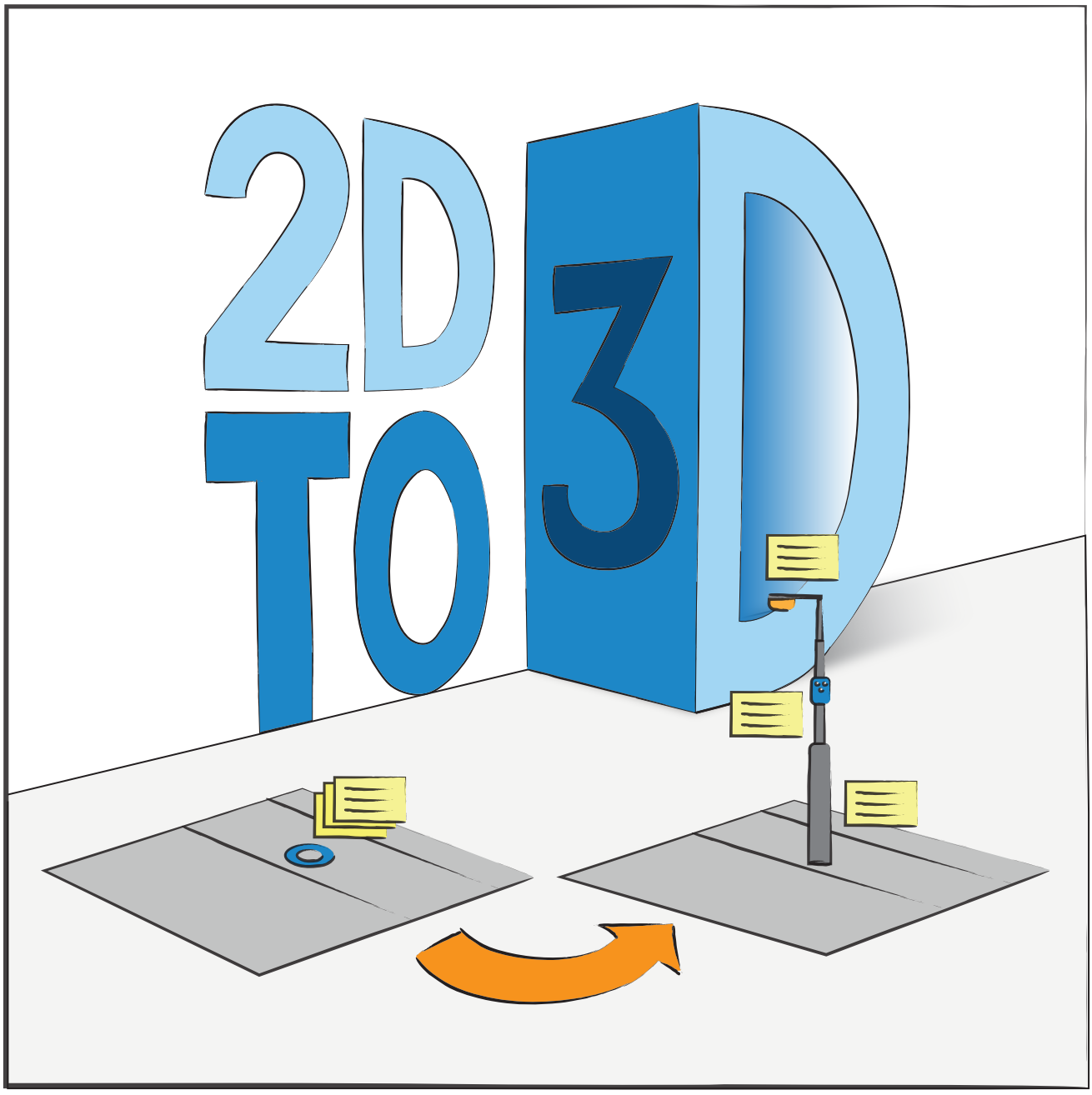
Linda van den Brink, Jantien Stoter, and Sisi Zlatanova. UML-Based Approach to Developing a CityGML Application Domain Extension. *Transactions in GIS*, 17(6):920–942, 2013b.

Linda van den Brink, Paul Janssen, Wilko Quak, and Jantien Stoter. Towards a high level of semantic harmonisation in the geospatial domain. *Computers, Environment and Urban Systems*, 62:233–242, 2017.

Linda van den Brink, Payam Barnaghi, Jeremy Tandy, Ghislain Atemezing, Rob Atkinson, Byron Cochrane, Yasmin Fathy, Raúl Garcia Castro, Armin Haller, Andreas Harth, Krzysztof Janowicz, Sefki Kolozali, Bart van Leeuwen, Maxime Lefrancóis, Josh Lieberman, Andrea Perego, Danh Le-Phuoc, Bill Roberts, Kerry Taylor, and Raphaël Troncy. Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web. *Semantic Web Journal*, Pre-press:1—20, 2018.

Erik Wilde. Declarative Web 2.0. In *Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on*, pages 612–617. IEEE, 2007.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 2016.



Definition and establishment of a national 3D standard

Chapter 2

Establishing a national standard for 3D topographic data compliant to CityGML

Authors: L. van den Brink, J. Stoter and S. Zlatanova. (*International Journal of Geographical Information Science*, 27(1):92-113, 2013).

This paper has been published in a peer-reviewed scientific journal in 2013 and is published unchanged in this chapter, except for footnotes where necessary to supply current information about an outdated statement. The paper describes how a national standard for 3D topographic data was developed as an extension to the international standard CityGML. To assure international interoperability, CityGML was selected as the best suited international standard to serve as the basis for the national standard in question.

Contributions: 1) establishment of a national standard on 3D topography that aligns to an international 3D standard; 2) improvements for IMGeo like the addition of classes for vegetation, auxiliary traffic areas and the addition of 2.5D and 3D geometry; and 3)change requests for CityGML, most importantly for the addition of a class for other constructions, which are not buildings, bridges or tunnels (to be implemented in next version of CityGML).

text of published paper starts after this line

Abstract: This paper describes a research project that realised a national standard for 3D geo-information. The standard was developed as part of a pilot in which more than 65 private, public and scientific organisations

collaborated to analyse and push 3D developments in the Netherlands (run between March 2010 and June 2011). The 3D standard was established through several steps. Firstly a comparison between the existing 3D CAD and Geographic Information Systems (GIS) standards was carried out that selected the OGC standard CityGML as the optimal 3D standard to align to. Secondly, the equivalent concepts in CityGML and the existing national standard for large scale topography (IMGeo) were identified. Thirdly IMGeo was extended to 3D following the principles of CityGML Application Domain Extensions (ADE). The model was tested by applying it to real data. Based on the experiences of this pilot, this paper proposes a framework of guidelines and principles for extending CityGML for national purposes, deduced from the modelling experiences. This is a unique contribution since experiences on extending CityGML are new and not well-described in the OGC CityGML specifications. Finally this paper presents the change requests which have been submitted to OGC to make the CityGML standard more suited for integration with existing 2D topographic information models. The change requests were formulated based on experiences from developing this nation-wide 3D standard.

Keywords: 3D standard, 3D geo-information, 2D/3D integration

2.1 Introduction

Over the past ten years technologies for generating, maintaining and using 3D geo-information have matured. Nowadays, many local governments have 3D models of the city, a large number of companies are providing services for constructing 3D models, and universities and research organisations are investigating 3D technologies (3D re-construction, data management, validation and visualisation). Yet many (governmental) organisations face numerous challenges in introducing 3D applications and technologies in their day-to-day processes. Despite the practical difficulties, it is clear that 3D information is becoming increasingly important in many applications. These developments motivated a pilot in the Netherlands to advance the use of 3D in this country. The pilot was initiated by the Dutch Kadaster, Geonovum, the Netherlands Geodetic Commission (NCG) and the Dutch Ministry of Infrastructure and Environment.

From January 2010 until June 2011 more than 65 private, public and research organisations got together to study the state-of-the-art of 3D developments and applications in the Netherlands and to instigate innovations. The pilot realised a proof of concept for a 3D Spatial Data Infrastructure (SDI) that addresses issues ranging from 3D data acquisition, maintenance

of 3D data and use of the 3D information in specific applications. An important goal was establishing a 3D standard NL with wide support of many stakeholders. For this purpose use cases were defined and executed on a 3D test bed. In addition large amounts of test data were made freely available for all participants. Finally the established Dutch 2D standardisation framework was studied for extension into 3D while aligning to the international standardisation developments driven by experiences of the use cases and the test bed.

The overall pilot goals and results are described in Stoter et al. (2011). This paper describes the development of the national 3D standard and proposes a framework for this that can be used by other countries.

Although other efforts are known for defining agreements on 3D geo-information in formal information models for different domains (Tegtmeier et al., 2009; Emgard and Zlatanova, 2007; Penninga and Van Oosterom, 2008; Stoter and Salzmann, 2003; Van Oosterom and Stoter, 2010), no attempts have been made to create a 3D national standard that is aligned to both the OGC (Open Geospatial Consortium) standard CityGML (Gröger et al., 2008) and the national 2D standardisation framework. The Netherlands has well-established national standards, but as in most countries, they are all 2D. The new 3D standard preserves valuable 2D concepts from the existing national standard for large scale topography (Information Model Geography: IMGeo), and extends them with 3D concepts from CityGML. The 3D standard is therefore not just another standard on geo-information, instead the realised CityGML implementation profile bridges the 2D and 3D standardisation developments.

The pilot experiments showed four technical reasons to preserve information from existing 2D models, while extending to 3D and aligning with international 3D standards. These are:

- Connection to existing datasets means connecting to existing application areas which provides a justification for the 3D information;
- Existing datasets often contain rich semantics, which is difficult to obtain from automated acquisition techniques;
- Existing datasets contain information about objects that often provides possibilities to automatically generate a 3D model;
- The update process (which is well-established in 2D) of existing datasets can still be used for updating the 3D datasets, before full update of 3D data sets is developed.

The result of a nationwide 3D standard extending CityGML and integrating it with 2D topographic information may be seen as a solution limited to one specific country as well as to topographical context. However, the defined 3D standard contains many generalities which are of interest to both different countries and domains. In particular, extension of CityGML to a specific context is not well described in the OGC specifications and experiments on CityGML extensions are new. Therefore the major contribution of this paper is the proposed generic framework for extending CityGML for national purposes that structures the findings of this research. Another contribution are the Change Requests (CR's) for CityGML which were formulated (and submitted to OGC) based on insights obtained during the development of the 3D standard and sequential testing.

The paper is organised as follows. Firstly, Section 2.2 motivates why CityGML was elected as the most promising 3D standard for the Dutch case. Section 2.3 describes how IMGeo was integrated with CityGML, based on principles of CityGML Application Domain Extensions (ADEs). Section 2.4 discusses the resulting information model. Section 2.5 presents the framework for extending CityGML for national purposes. The change requests for CityGML are formulated in Section 2.6 and Section 2.7 presents conclusions and elaborates on future developments.

2.2 Motivation to use CityGML as base for 3D standard NL

3D standards have been developed throughout the years for many different purposes: visualisation (fast and realistic), data management (efficient storage), modelling (validity and topology) or data exchange (platform independent). The parties working on 3D standardisation vary from companies to international standardisation organisations, originating from CAD/BIM (i.e. Computer Aided Design and Building Information Models (BIMs)), GIS or Web domains. Many of the company developed formats have become de facto industry standards (e.g. SHP, DXF) or have been approved as open international standards (e.g. KML). Other international standards have been developed without major company involvement (e.g. CityGML). Being developed with different goals, the information (such as type of geometry, textures, semantics, relationships) varies significantly between standards and makes the integration of data in one 3D environment almost an impossible task. The experiences in the 3D pilot have clearly revealed many problems with converting data from one de facto or international standard to another:

information was lost or was improperly converted, validity of objects was not ensured, relationships were diminished, etc. Therefore it was important to study and analyse the most used international and de facto 3D standards and their characteristics.

The de facto and international standards that were compared are DXF, SHP, VRML, X3D, KML, Collada, IFC (one of the most important standard to model construction objects in the BIM domain), CityGML and 3D PDF. For more details on those standards see Stoter et al. (2011).

The comparison showed that every 3D standard is designed for specific purposes. DXF, VRML, X3D, Collada, and IFC support the largest variety of geometries. VRML, X3D and Collada are the most advanced in supporting realistic textures. All these standards, except IFC, contain poor support for semantics and attributes. Clearly these standards originate from the CAD domain. In contrast, standards such as SHP, IFC, and CityGML have a very good support of semantics, objects, attributes and relationships between the objects. This means that these standards provide the means to keep information that is important for analysis and not only for visualisation. Because of the support for semantics, geo referencing and Web use, the selection of CityGML as generic standard for a 3D SDI envisaged in this study was justified. IFC shows similar support but is characterised by its local and very detailed approach, the limited number of construction models usually available in a city and high precision necessary for reliable construction calculations. For the 3D SDI a standard for geo-information is needed characterised by coverage of large areas (e.g. a complete city) and lower precision. Since BIM (IFC) files may serve geo-information applications and vice versa, it is important to study the alignment of both standards. Further details on the comparison can be found in Stoter et al. (2011).

The OGC standard CityGML (Gröger et al., 2008) originated in academia in Germany (Bonn, TU Berlin) and was originally defined as an exchange format. But it is also—and especially—an information model for representing 3D spatial objects. CityGML distinguishes both at the geometric and semantic level between thematic concepts (buildings, vegetation, water, land use, etc.) (Albert et al., 2003; Gröger et al., 2004; Benner et al., 2005; Gröger et al., 2007). It supports multi-resolution features by means of different levels of detail (LODs). LOD0 represents the surface geometry of objects at terrain level. In addition a building object can vary from a simple block model (LOD1; accuracy 5m), with roof shapes (LOD2; accuracy 2m), with windows, doors and other exterior features (LOD3; accuracy 0.5m) to a fully detailed interior model (LOD4; accuracy 0.2m) with or without texture information (called 'appearance').

The standard CityGML is based on GML3 (Consortium et al., 2007).

Generally volumetric objects are possible as in GML 3, but the validity of closed volumes cannot be enforced in CityGML. In addition 3D topological structures, although available in GML 3, are not utilized. Because of the complexity to build and maintain topology in 3D, objects are modeled with geometrical primitives and not with topological primitives. A simple xlink approach is followed to connect the surfaces of a 3D geometry. The time and scale dimensions are handled in CityGML, but not in an integrated manner. The time dimension is separately handled by adding attributes to geometrical objects, i.e. creationDate and terminationDate. Scale is intended to be linked to the LOD concept, although our experiences have shown that this is not the case, i.e. many 3D models in LOD1 are created from high-accuracy data.

CityGML is intended as generic standard with limited thematic content compared to the national information models and therefore the standard needs further agreements to make the standard suitable for national purposes. This was studied in a next step, i.e. How to use the generic standard CityGML as a standard in a specific (i.e. Dutch) context, i.e. which additional classes, attributes and attribute values are necessary? Which codes should be added to the code lists of CityGML to make the code lists appropriate for a specific context and how can this be done? Which LOD should be used? These questions are studied in the next sections (for the specific Dutch context in Sections 2.3 and 2.4, and in general in Section 2.5). Note that also the European data specifications for buildings (Building, 2011) have an optional CityGML profile.

2.3 Extending CityGML for Dutch context

After the election of CityGML as standard to align with, a CityGML implementation profile has been developed. Because the Dutch information model on large scale topography (IMGeo) resembles CityGML the most, the first focus has been on integrating IMGeo and CityGML into one standard. For this integration we used the available version of CityGML (version 1.0.0). Version 1.1 is in consultation at the moment of writing (OGC, 2011). This section presents the process that established the national 3D standard CityGML-IMGeo (Section 2.3.2). First Section 2.3.1 introduces technical aspects of IMGeo.

2.3.1 IMGeo (BGT)

The Dutch Information model Geography (IMGeo) describes how object-based, large scale (1:1000 and 1:2000) topographic features must be defined

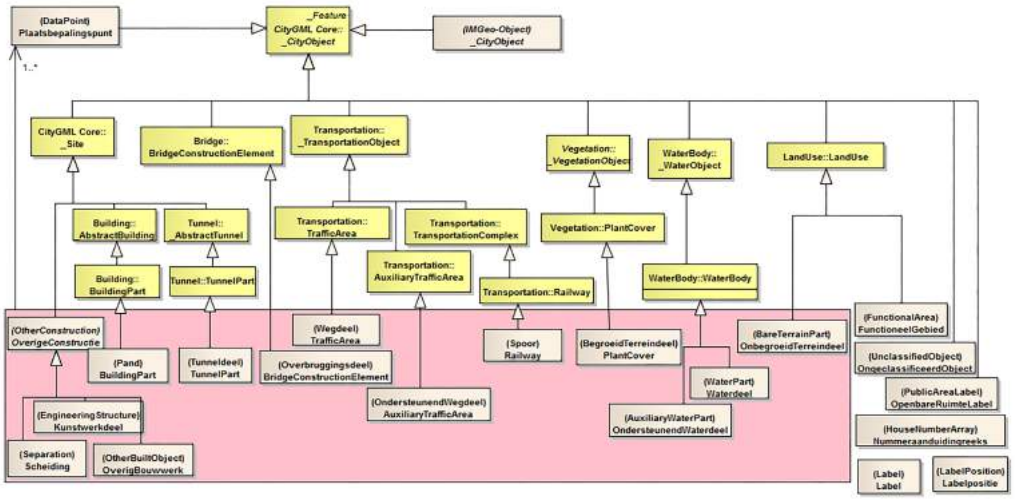


Figure 2.1: Overview of the classes in the Dutch Information Model Geography (IMGeo); classes in light color are part of the mandatory core, the classes in orange are optional. The names in italics are CityGML super classes of the IMGeo classes; English translation of class names is shown in parentheses.

to make the national exchange of this information possible. This large-scale topographic map is created and maintained by the municipalities and financially supported by local governments and private companies. Version 1.0 of IMGeo was published in 2007 (Geonovum, 2007). Version 2.0 is due to be completed end 2011[1]. IMGeo 2.0 has a mandatory core, see Figure 2.1 (and Table 2.2, Section 2.3.2.1 for English translation of the main classes).

Data providers such as municipalities, organisations responsible for the road, water and railway infrastructure etc. will be required by law to provide their objects that fall under the definitions of the IMGeo 2.0 core to a national 'basic registry' (Basisregistratie Grootschalige Topografie, BGT) where they are available for reuse. The mandatory core contains object definitions for large scale representations of roads, water, land use/land cover, bridges, tunnels etc. The optional part of IMGeo allows further division of these objects into parts suitable for maintenance, and contains definitions for all kinds of city furniture and other non-mandatory classes. It should be noted that utilities and geology objects are not part of IMGeo. They are covered in two other domain models, i.e. respectively IMKL and IMBRO (Geonovum, 2012). The terrain as a regular or irregular grid is also covered in the latter (e.g. AHN, Height Model of The Netherlands, `http://www.ahn.nl/`)

The mandatory core of IMGeo prescribes 2D point, curve or surface ge-

[1] [Added 2018] IMGeo 2.0 was published in February 2012.

ometry for all objects, but because the new version (2.0) of the model is completely integrated with CityGML (as result of the study described in this paper , the optional part of IMGeo 2.0 also allows 2.5D (i.e. LOD0) and 3D geometries (i.e. volumetric representations as prescribed at CityGML LOD1, LOD2 and LOD3).

IMGeo (also the 2D part) contains man-made objects above (e.g. viaduct) and below the surface (e.g. tunnels or underground waterways), modelled in 2D with the attribute relatieveHoogteligging (relative height). This attribute indicates whether an object is located at surface level (and is thus part of the planar partition, value=0), above (>0) or below (<0) the surface. The attribute is mainly used to infer that objects do not exist on the same level and which object is above which.

IMGeo classes have a small number of attributes, besides their geometry. Most have one or two attributes to further classify the object or to indicate their function. Code lists are used to provide allowed values for these attributes. In addition, all classes share attributes for identification and versioning, for a reference to the data provider, for the object's status (planned, existing, or historic) and an indication whether a possible error in the data is under investigation. In addition, all measure points of the object's geometry are stored with metadata such as information on the accuracy.

All objects have 2D geometry for which GML 3 geometry types are used. Topological rules are part of the standard, but are not modelled by GML topological types. The most notable rule is that the complete set of polygon-objects at surface level (height level 0) in the mandatory core must together form a complete coverage of the Netherlands without gaps or overlap.

The IMGeo standard also includes rules for visualisation, e.g. the colour of lines and areas, type and thickness of lines, etc. These are based on a Dutch standard for web cartography. Two visualisation themes are provided: one where large scale topography is the main focus, and one where it is used as a background for other themes.

2.3.2 3D IMGeo as extension of CityGML

For the design of 3D IMGeo extending CityGML, the rules for creating Application Domain Extensions (ADE) were applied (Gröger et al., 2008; Wiki, 2012). That is, 3D IMGeo was modelled as an ADE, even though the intention is not to position 3D IMGeo as a formal (i.e. to be approved by OGC) extension module. It is nevertheless useful to follow the rules for creating an ADE because these rules assure a standard extension method, enabling software systems to not only understand the CityGML part of the model, but also the extensions. Since the rules for modelling an ADE in UML(Unified

Table 2.1: Available LOD representation for each CityGML thematic module

CityGML module	LOD0	LOD1	LOD2	LOD3	LOD4
Building		o	o	o	o
CityFurniture		o	o	o	o
LandUse	o	o	o	o	o
Relief	o	o	o	o	o
Transportation	o	o	o	o	o
Vegetation		o	o	o	o
WaterBody	o	o	o	o	o
Generics	o	o	o	o	o
Bridge		o	o	o	o
Tunnel		o	o	o	o

Modelling Language) are not described in the CityGML standard we followed the rules for implementing an ADE in a GML Application Schema, applying those to UML as much as possible. In addition the publication by Portele (2009) was used.

The remainder of this section clarifies the main considerations that led to the CityGML implementation profile for Dutch context realised in IMGeo 2.0.

2.3.2.1 Reuse and extension of CityGML concepts

The main principle of CityGML-IMGeo is the reuse of CityGML concepts, i.e. IMGeo classes are remodelled in accordance with CityGML. This principle made it necessary to map the concepts from IMGeo to CityGML concepts. For this mapping-study both the CityGML core standard and existing (draft) ADEs (bridges, tunnels, noise, and utility networks) were considered. The Bridge and Tunnel models were used (these have become part of CityGML 1.1). However, the other ADEs were not used because the (semantic) details they add to the CityGML standard are not available in IMGeo.

As mentioned above, IMGeo only contains 2D objects, which makes the mapping between IMGeo classes and CityGML LOD0 classes (representing the terrain) the most obvious choice, even though the accuracy of IMGeo (varying between 0.30m and 0.60m) is a far better match for LOD3 . However, only looking at LOD0 classes has limitations because not all CityGML classes have representations at LOD0 (see Table 2.1). Therefore the mapping also took other LODs into account and LOD0 representations of those classes have been added at a later stage (see Section 2.3.2.2). It should be noted that LOD0 for Building is added in CityGML 1.1.

The mapping was the source of the CityGML-IMGeo implementation modelled in UML. Since CityGML 1.0 is only available as xml schema, a

Table 2.2: Mapping between IMGeo and CityGML classes for CityGML-IMGeo implementation profile

IMGeo Class	**CityGML Class (not only LOD0)**	**In CityGML LOD0?**
Kunstwerkdeel (Construction part)	Bridge, Tunnel and other constructions (OtherConstruction class with sub classes added as specialization of _Site)	No
Inrichtingselement (furniture element; generic term for city furniture and other small objects that populate the public space)	Most are CityFurniture; Some are Constructions	No
Pand (building part, premise)	BuildingPart	No
Spoor (railway tracks)	Railway	Yes, as network; surface from LOD1 (TrafficArea)
Terreindeel (terrain part). To match CityGML this class was split in two classes: bare terrain part and covered terrain part.	LandUse (bare terrain parts, with no vegetation) and PlantCover (covered terrain parts, with vegetation)	LandUse: Yes Vegetation: No
Waterdeel (water part)	WaterBody	Yes
Wegdeel (road part)	Traffic Area	Yes, as network; surface from LOD1 (TrafficArea)
Ondersteunend wegdeel (auxiliary road part). Was introduced to better match CityGML.	AuxiliaryTrafficArea	No
Registratief Gebied (administrative area)	LandUse	Yes

UML model was recreated in the modelling tool Enterprise Architect, based on Gröger et al. (2008). IMGeo classes (with Dutch names) were defined as a specialisation of the relevant CityGML generic class (with English names). In cases where only extra properties were defined for an existing CityGML class, the subclass was marked with a stereotype <<ADEElement>>and the English names of CityGML were reused to make the link to the original classes well-visible (which English class refers to which Dutch class) which is reflected in the UML models via the specialisation-relationships.

New classes have been added if they are present in IMGeo but missing in CityGML. These are also modeled as a specialization, but with a stereotype <<featureType>>and a Dutch class name. Added classes include constructions related to water management, separating objects like walls and fences, and other constructions which are not quite buildings, like storage tanks or wind turbines. These classes have been added with one superclass called "OverigeConstructie (OtherConstruction)", specialisation of CityGML _Site.

Attributes are added to all classes for defining the following aspects: 2D geometry (not modelled in CityGML), the LOD0 geometry if not present in CityGML (see Section 2.3.2.2) and the Dutch classification code lists.

Retaining Dutch code lists instead of reusing the CityGML code lists breaks the CityGML rules for extending code lists. However several aspects gave reasons to do this (and this is in line with the findings of Portele (2009)). Firstly the national character of the standard favors Dutch language code lists. Secondly the CityGML standard does not provide definitions for the code list values, which makes it hard for users to decide which value to use. Because IMGeo also contains a obligatory part which prescribes data

providers to provide a complete set of the core objects with a certain accuracy and semantic correctness, it is even more important that every concept is defined precisely. The Dutch code list values do have definitions and therefore these values are retained (and if possible mapped to CityGML code list values, see Table 2.3 in Section 2.4. It should be noted that CityGML 1.1, which was not yet available at the time of writing[2], allows extension and replacement of code lists.

A final adjustment of the model was the use of CityGML classes which were not present in IMGeo 1.0, but which appeared to be more appropriate for handling specific concepts. These are Vegetation for modelling any vegetation-related concept (in IMGeo 1.0 divided over several classes) and AuxiliaryTrafficArea for road segments which are not used for traffic (such as verges).

The resulting standard supports both the 2D representations of the mandatory and optional objects, as well as 2.5D and possibly 3D representations of those objects according to geometric and semantic principles of CityGML. In addition the consequence of remodeling IMGeo 1.0 concepts into IMGeo 2.0 concepts according to the rules described above is that all matches between IMGeo 2.0 and CityGML are 1 to 1.

2.3.2.2 Define further agreements on geometry

CityGML-IMGeo contains further agreements on the use of LOD representations, which solves CityGML's ambiguity about the LOD definitions (geometry- or semantic-based):

- Accuracy: All LODs have the same accuracy according to the positional accuracy required by IMGeo. Consequently the accuracy requirements of CityGML are not adopted, and the LOD only represents the detail of the object in the third dimension, i.e. LOD0 for 2.5D surfaces, LOD1 for block models, LOD2 for semantics for roofs, walls etc and LOD3 for textures and even more details.

- LOD0 representation for all classes: All CityGML-IMGeo classes get a LOD0 representation (either inheriting from CityGML or added conform the IMGeo geometry type) and a 2D geometry in the 2D-LOD. This makes it possible to have one integrated model that supports the full range of spatial representations of real world objects, i.e. from 2D, 2.5D to full 3D, depending on the application.

[2][Added 2018] CityGML 1.1 was never published; instead it was renamed CityGML 2.0 and published in 2012.

- Solid and indoor geometry: Some classes in CityGML-IMGeo have representations at higher LODs, up to and including LOD3 conforming to CityGML, examples are buildings, tunnels, bridges and trees. However LOD4 is excluded from CityGML-IMGeo as indoor information is considered to be information from another domain (i.e. BIM/IFC) than topography (this is in line with Building (2011)).

For some classes other geometry types were used than modelled in CityGML. In our approach Water and Road objects are represented by surface geometry at LOD0 and Railway is represented by the curve geometry at LOD0, while CityGML represents those classes with the GeometricComplex at LOD0 to accommodate networks. The changes were done to be compliant with geometry types in the previous version of IMGeo.

2.3.2.3 Define agreements on topological structure

As mentioned above IMGeo defines a topological structure containing all objects at surface level. This topological principle is extended to LOD0. Consequently the LOD0 representations of all objects at surface level constitute a 2.5D topological structure (without gaps and overlaps). Because this topological structure contains also the footprints (i.e. the geometry of objects at the surface), no specific Terrain Intersection Curves are necessary to locate the LOD1-LOD3 representations in the terrain.

This principle extends the CityGML topological structure for LandUse (page 94 of CityGML specifications (Gröger et al., 2008)) to other classes at surface level which may be additional LOD0 classes compared to CityGML (since LOD0 representations have been added for all classes). At the same time the class LandUse is limited to those objects which do not represent the other classes, such as water, road, railway, vegetation.

Figure 2.2 shows the result of a 2.5D triangulated surface of IMGeo objects. The 2.5D surfaces are a result of a "constrained triangulation" of high density laser points and the 2D geometry of all IMGeo objects at surface level.

2.3.2.4 Model objects above or below the surface

The concept of levels (relative height) in IMGeo is reflected in the integrated model. LOD0 representations of objects above and below the surface are located in space while connected to the topological structure at surface level (based on the assumption that those objects do connect to the surface somewhere). It may be necessary to add extra 2.5D surfaces to the structure to avoid gaps (see Figure 2.3). This is because objects that touch in 2D

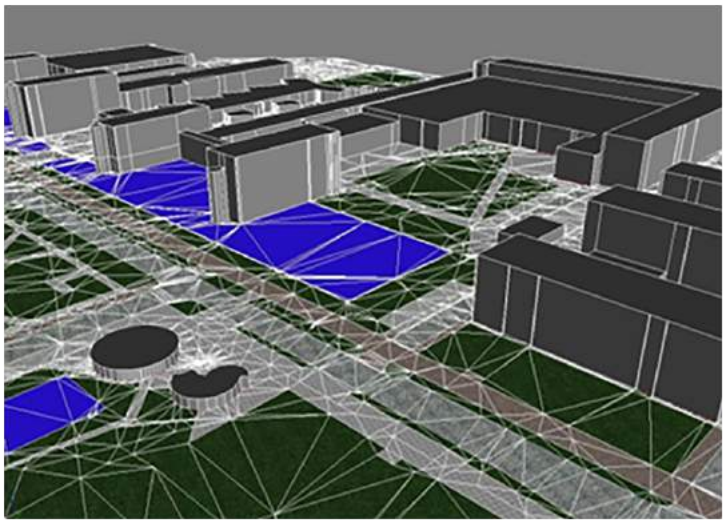


Figure 2.2: Triangulated terrain integrated with IMGeo objects (Emgard and Zlatanova, 2007)

may not touch in 3D, for example a waterbody and a road along the waterbody. This is further explained in Oude Elberink and Vosselman (2009) and Oude Elberink (2010).

Another Dutch dataset, i.e. the 2.5D topographical dataset of Rijkswaterstaat (DTB) already supports a similar 2.5D topological structure of topography. Examples are shown in Figure 2.4 (Rijkswaterstaat, 2011). Although this dataset shows that such 2.5D data structures are already in existence and therefore feasible, it does not yet contain height information within the polygon-surfaces, in contrast to the LOD0 representations of CityGML-IMGeo.

2.3.2.5 Define the reference system

CityGML-IMGeo uses the Spatial Reference System (SRS) EPSG:7415. This is a combined SRS of x, y coordinates in the national reference system (called Rijksdriehoeksmeting) (EPSG:28992) and z coordinates in the national height system (called Nieuw Amsterdams Peil) (EPSG:5709). The experiences in the pilot showed that this reference system should be explicitly stated in the CityGML files.

3D model of 3D Pilot test area (Oude Elberink, 2010), based on high resolution laser data and the 2D topographic dataset at scale 1:10k

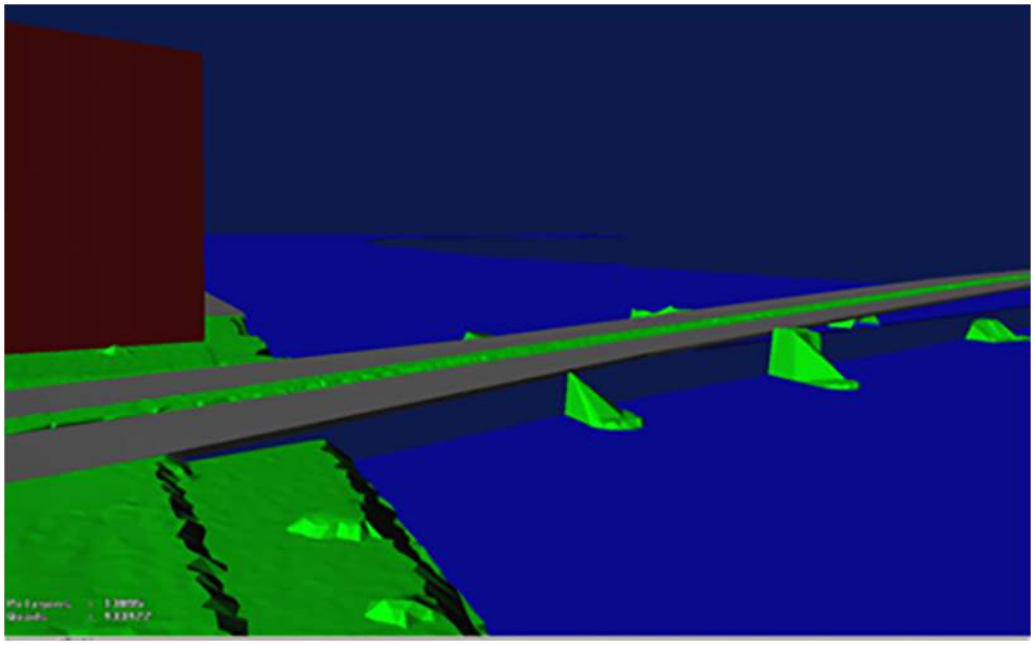
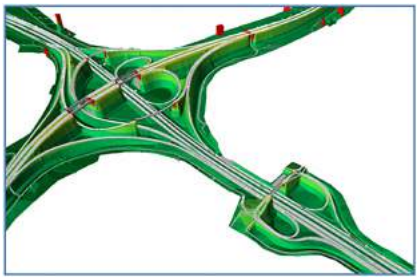


Figure 2.3: Example of the employed 2.5D topological principle

(a) a.



(b) b.

Figure 2.4: Examples of 2.5D topographic data provided to the 3D Pilot by and Province of Noord Brabant (a) and Rijkswaterstaat (b)

2.4 Experiments with the model

This section illustrates how the model works in practice by showing how the IMGeo class 'TrafficARea (Wegdee)l' (i.e. Road Part; road segments that constitute a road) is modelled in CityGML-IMGeo and by applying the model to real data at the end of this section.

The IMGeo TrafficArea (Wegdeel) class is modelled as a specialisation of the CityGML class TrafficArea, with a stereotype <<ADEElement>>to reflect that only properties are added. The UML diagram for class TrafficArea is shown in Figure 2.5.

IMGeo-Object is the collection of the generic properties that all IMGeo classes share. In IMGeo 1.0 all IMGeo classes were specialisations of this abstract feature type. To avoid multiple inheritance in the new model the IMGeo-Object class is no longer a <<featureType>>but an <<ADEElement>>. When a GML Application Schema is created from this UML model, the specialisation relation between IMGeo-Object and Wegdeel is not seen as inheritance but as an addition of properties to CityGML class _CityObject. These properties are realised in the GML application schema by attaching them to the _CityObject type using the CityGML 'hook' mechanism described in OGC (2011).

The stereotype <<BGT>>indicates the parts that are in the mandatory core of IMGeo. As can be seen, the only part which is not mandatory is lod0SurfaceWegdeel.

The attribute Wegdeel.fysiekVoorkomen is equivalent to TrafficArea.surfaceMaterial. The attribute Wegdeel.functieWeg is equivalent to TrafficArea-.function, while TrafficArea.usage gives sometimes significant additional information to establish a better matching, e.g. for the values "Ruiterpad" and "Voetgangersgebied". In a 3D IMGeo dataset both English and Dutch attribute names must be provided. This is also true for the code list values, because the Dutch code lists are retained (see Section 2.3.2.1). However, the mapping between attributes, code lists and code list values is not modelled explicitly in the UML model. Although it would have been possible to use constraints for this, those mappings are described in a document added to the standard. This is less formal but more accessible by users. The format that was used for the mapping is illustrated in Table 2.3 for the attribute Wegdeel.functieWeg. This attribute has two equivalent CityGML attributes, as was explained above.

As can be concluded from Table 2.3, the mapping between IMGeo and CityGML naming conventions is not perfect, because of several reasons:

1. In IMGeo there is no equivalent class for TransportationComplex or

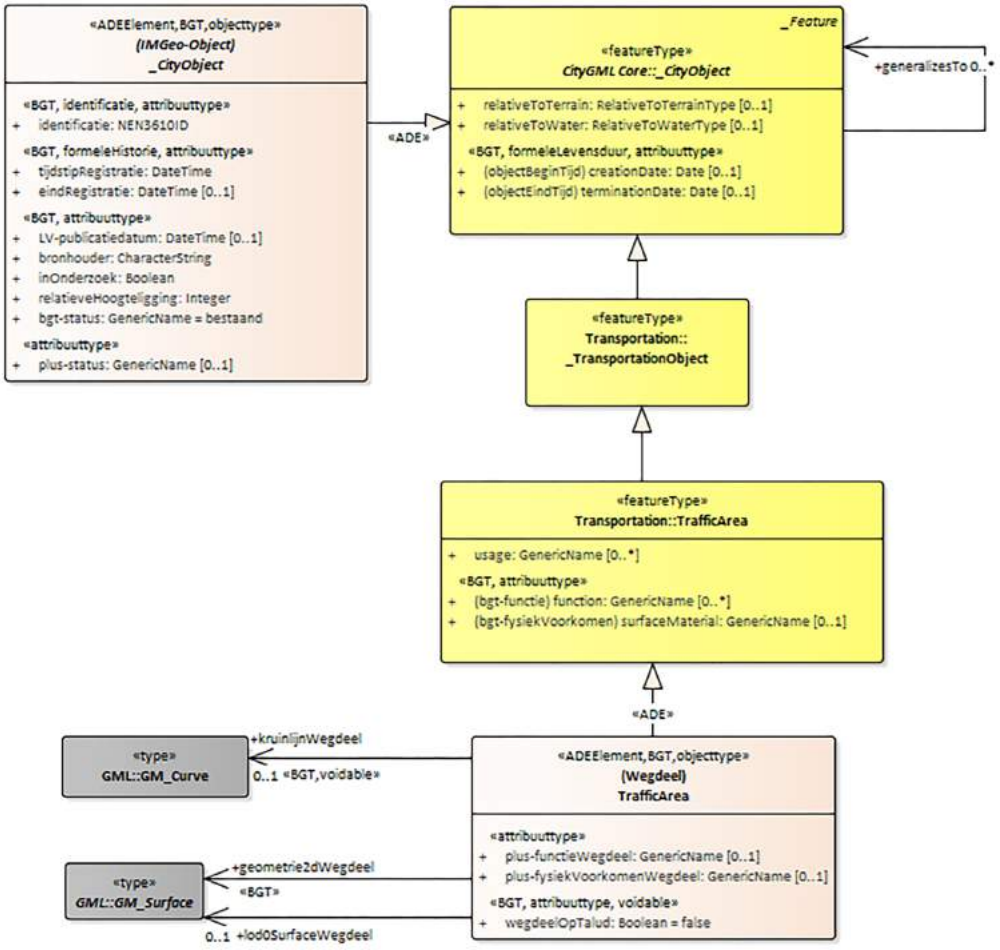


Figure 2.5: Example – IMGeo modelling of Wegdeel as subclass of CityGML-TrafficArea

Table 2.3: Example - Mapping of code values for Wegdeel-TrafficArea

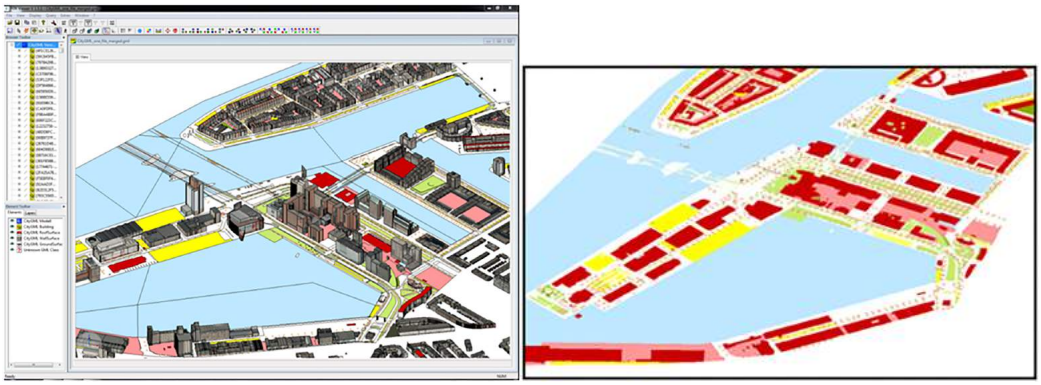
IMGeo	**CityGML**	**CityGML**
Attribute	***Attribute***	***Attribute***
FunctieWeg	*TrafficAreaFunctionType*	*TrafficAreaUsageType*
Code list value	**Code list value**	**Code list value**
OV-baan (public transport lane)	Rail or (missing: bus lane)	Bus, taxi Train Tram, streetcar
Overweg (railway crossing)	(missing: level crossing / railway crossing)	Car Truck Bus, taxi Train etc.
Baan voor vliegverkeer (lane for aeroplanes)	airport_runway, airport_taxiway, or airport_apron	Aeroplane
Rijbaan: autosnelweg (driving lane: motorway)	Motorway	Car Truck Motorcycle
Rijbaan: autoweg (driving lane: main through road)	driving_lane	Car Truck Motorcycle
Rijbaan: regionale weg (driving lane: district road)	driving_lane	Car Truck bicycle etc
Rijbaan: lokale weg (driving lane: road)	driving_lane	Car Truck bicycle etc
Fietspad (cycle path)	Cyclepath	Bicycle
Voetpad (footpath)	Footpath	Pedestrian
Ruiterpad (bridle path)	unknown	(missing: path) Horse
Parkeervlak (parking area)	parking_lay_by or car_park	Car
Voetgangersgebied (pedestrian area)	unknown (missing: pedestrian area)	Pedestrian
Inrit (road to private property)	private_area	Car Truck bicycle etc
Woonerf (pedestrian priority area)	unknown (missing: pedestrian priority area)	Car Truck bicycle etc
Attribute	***Attribute***	
FysiekVoorkomenWeg	TrafficSurfaceMaterialType	
Gesloten verharding (closed hardening)	asphalt or concrete	
Open verharding (open hardening)	pavement or cobblestone	
Half verhard (semi hardened surface)	Gravel	
Onverhard (unhardened)	soil or sand	

Road. Only road parts (Wegdeel) are modelled. However, the code list for Wegdeel function in some cases gives the function of the road part as well as the function of the whole road it is part of. This is the case with the values starting with 'rijbaan' (English: driving lane).

2. Some values are difficult to map because there is no good equivalent in the CityGML code list. For example, there are values available in TrafficAreaFunctionType for cycle path and footpath, but not for paths designated for use by equestrians. A generic term 'path' in CityGML would have better suited, in combination with the TrafficAreaUsageType code list to indicate the modes of transportation allowed on the path.

3. The corresponding code list for traffic surface material in IMGeo starts at a more abstract level than TrafficSurfaceMaterialType in CityGML. These issues of mapping code list value should be addressed in a future version of the model, since currently insufficient experiences are available to show which modelling approach (i.e. CityGML or IMGeo) is best.

To test the CityGML-IMGeo implementation profile, the UML model was converted into an xsd file. In addition data modelled according to IMGeo 2.0



(a) (b)

Figure 2.6: Visualisation of CityGML-IMGeo encoded data: CityGML LOD2 (a) and 2D-LOD (b)

and laser point data were combined to generate CityGML-IMGeo encoded data. Figure 2.6 shows the result. Since the model supports 2D as well as 2.5D and 3D representations, both 2D IMGeo data (Figure 2.6b) and 3D IMGeo data (Figure 2.6a) can be generated according to the same model.

2.5 Framework for extending CityGML for national purposes

Based on the experiences of developing the CityGML-IMGeo standard, this section defines a framework for extending CityGML for national purposes. The framework defines how a 2D information model can be integrated with CityGML to support the full range of dimensional geometry types for geographical features in an integrated manner: i.e. 2D, 2.5D and 3D. The current research focused on the integration of CityGML and 2D large scale topographical information. Future research should identify how such an integration works for mid- and small scale topography as well as for geographical information captured in information models from other domains. The framework covers the following aspects:

- Integration of 2D information model and CityGML
- Geometry types and LOD
- Topology
- Use of code lists

- Use of CityGML properties
- Reference system

2.5.1 Integration of 2D information model and CityGML

The main principle of the framework is that the integration of the national information model on 2D topography and CityGML should adhere to the CityGML standard as much as possible while keeping the CityGML standard unchanged. This requires the following steps:

1. Conceptual mapping is needed between CityGML and the 2D information model to identify equivalent concepts, i.e. classes, attributes, code lists and code list values. In these mappings one should compare at semantic level, i.e. considering the concepts in any CityGML-LOD.
2. All classes of the national information model (in most cases with class names in national language) must be modelled as subclasses of CityGML classes (with English class names).
3. For classes for which no equivalent CityGML class can be found, two solutions are possible of which the first one is preferred:
   (a) Remodel the 2D concept so that an equivalent CityGML class can be found (as was done for Vegetation and AuxiliaryTrafficAra in IMGeo)
   (b) If a. is not possible, CityGML needs to be extended with a new class (as was done for Construction in CityGML-IMGeo).

NB1 : For extending CityGML with our own classes and additional class specific attributes we did not make use of the extension possibilities of GenericCityObject and _genericAttribute. The reason was that the extension would not be formally defined in a schema with (Dutch language) names and definitions, and it would therefore not be possible to validate data that uses such extensions.

NB2 : Also for extending the generic _CityObject with own attributes (that apply to all classes) we did not make use of _genericAttribute. Instead the stereotype <<interface>>was used to enable that a subclass inherits attributes both from the _CityObject (via a generalisation-specialisation relationship) and the root class of the 2D information model (via the <<interface>> stereotype). The reason for this is that some of the 2D IMGeo

attributes have datatypes (complex, Boolean, code list) that cannot be described with _genericAttribute.

2.5.2 Geometry types and LOD

For the use of geometry types and the LOD concept, the following guidelines apply:

1. Which LOD is used should depend on the required level of detail in the third dimension rather than the positional accuracy as mentioned in CityGML specifications. Therefore the integrated model defines the same positional x, y accuracy for all LODs as prescribed by the 2D information model.

2. Extra attributes defining geometry types are needed to support the full range of geometries for every class: 2D geometry (not modelled in CityGML) and the LOD0 geometry if not present in CityGML (as footprints). With those LOD0 representations the exact 3D location of objects is known and 3D representations of those objects (LOD1 and higher) can easily be placed on the terrain assuring that they do not float in the air or disappear in the ground. In addition this approach supports 2D, 2.5D and 3D representations of objects in an integrated manner. Other deviations of CityGML geometries may be necessary if the 2D information model supports different geometry types than CityGML. For example surface for Water and Road at LOD0 instead of network geometry as in the CityGML-IMGeo case.

3. Special attention should be given to the link between buildings and terrain. Two approaches can be followed:

   (a) The buildings have always horizontal foundation in the terrain and can sink, but the above surface and below surface geometries are modelled separately. The 2D representations of the buildings represent the building geometries at surface level, which can straightforwardly be extended into 2.5D. This modelling makes the Terrain Intersection Curves redundant (TIC) . This approach was used in CityGML-IMGeo.

   (b) The buildings have always horizontal foundation and may have underground parts (Zlatanova et al., 1996). This means that the building can sink under the surface and the TIC for buildings has to be maintained to ensure the consistency with the terrain surface.

4. Terrain object as given in CityGML should be carefully considered. Three different approaches can be followed:

   (a) Terrain is represented by a surface of a regular or irregular grid. The topographic objects are not integrated in the terrain (the current concept of CityGML)

   (b) Terrain is represented by constrained TIN, which ensures the consistency between objects and terrain surface. The terrain is still present in the model.

   (c) Terrain is not represented as a separate object, i.e. all the object on the ground (such as roads, land use, etc.) incorporate the terrain curvature in their representation (see Figure 2.2. This approach is used in CityGML-IMGeo.

2.5.3 Topology

The following guidelines apply for the topological structure:

1. If the 2D model contains a topological structure, the 2.5D surface (LOD0) should support this as well: all objects at surface level have a representation at LOD0 (= 2.5D surface) which together form a 2.5D topological structure for those objects located at surface level.

2. Objects that are located above and below the surface can also be placed in the third dimensional space with their LOD0 2.5D representation. An important requirement here is the connection to the 2.5D DTM that represents the surface. This may require adding new 2D boundaries for adding more variance in 3D or extra 2.5D surfaces to the structure at the surface to avoid gaps.

2.5.4 Use of code lists

To make use of national classification code lists (which will be so specific in most cases that those national lists are preferred over the CityGML code lists) mapping tables between the code lists/ code list values should be provided (see Table 2.3 as example).

Code list validation can be done using standard XML techniques such as Schematron constraints (ISO, 2006). CityGML 1.1, which was not yet available at the time of writing, allows extension and replacement of codelists. Further work is needed to assess how code lists can be used, maintained and validated in the new approach.

2.5.5 Use of CityGML properties

Mapping tables should prescribe which CityGML properties must, may or must not be filled when exchanging data according to the national 3D standard. For example, in the case of IMGeo, 2D geometry must always be provided, LOD0 – LOD3 may be provided, and LOD4 must not be provided. Another example is that the 'function' and 'surfaceMaterial' properties of TrafficArea must be present and filled exactly once for each TrafficArea object. For the IMGeo standard more work is needed to describe these rules and mappings.

2.5.6 Reference system

For a national 3D standard it is required to identify the reference system (x, y, z) to be used in all CityGML files.

2.6 Change requests for OGC CityGML

The development of the model and sequential testing revealed some deficiencies both for IMGeo and CityGML. The deficiencies for IMGeo have been solved in the modelling process (for example adding classes for Vegetation and AuxiliaryTrafficArea). For CityGML a number of change requests (CRs) have been formulated and submitted to OGC if not yet covered by already submitted CRs (which are being considered for the next version of CityGML, see OGC (2011)). The CRs that were formulated as part of this research are summarized and justified in this section.

Firstly, using CityGML as implementation for a national standard requires clearer definitions. The advantage of not having definitions for classes, properties and code list values is that the model can be applied in different contexts with greater flexibility. However lack of definitions makes it difficult to understand what is meant by a concept in CityGML. This issue confirms the CityGML change request (Portele, 2010) resulting from the OWS 6 test bed (Portele, 2009). This change request has been accepted with the decision to collect such definitions in a registry which has just been set up.

Another change request relates to the use and extension of code lists. The CityGML codelists are difficult to use because the values may not be ignored. However they are not easy to reuse since definitions are missing, concepts are overlapping or missing, etc. Therefore a guideline is needed for how to extend code lists or replace them. This issue is already addressed by the CityGML change request mentioned above (Portele, 2010) and is solved in CityGML 1.1.

For an implementation of CityGML further clarification is needed for the CityGML accuracy requirements: are these rules or guidelines? Page 10 of the CityGML standard states that "accuracy requirements in this standard are debatable and should be considered as discussion proposals". Further on, the requirements are formulated as formal requirements ("The positional and height accuracy of LOD2 must be 2m or better"...). This issue is also addressed in the mentioned change request (Portele, 2010) and is solved in CityGML 1.1.

Clearer guidelines for extending CityGML are needed for CityGML implementation profiles that support a specific context. Currently it is not clear whether the extension mechanisms of CityGML (generic objects and attributes and the ADE mechanism) are guidelines or rules, nor is it clear when to use which method for extension. Also, the ADE mechanism is only described in the context of GML application schemas. How to use the method in UML modelling is not described. The guidelines (or rules) for extending CityGML should be described more fully and clearly so that a working extension of CityGML (i.e. working in CityGML software) can be created by relying on these rules and/or guidelines. This issue is partly addressed by the CityGML CR09-039 (Portele, 2010). A separate change request has therefore been submitted to the OGC to cover this issue. At the time of writing a decision on this CR has not yet been made, but it is likely it will be postponed to a later version of CityGML to allow more discussion on the topic[3].

LOD0 footprints for all CityGML classes are required in order to integrate the full range of possible geometries of semantic objects in one model, i.e. 2D, 2.5D and volumetric geometries. This enables to use 2D topographic data with a DTM and to upgrade 2D data to 2.5D and 3D. All geometries at LOD0 including these footprints must ideally have a topological structure. This issue is partly addressed by an earlier change requests, which asks for planimetric building footprints in LOD0 (Roensdorf, 2010b). A separate change request has been submitted to the OGC to cover this issue. At the time of writing a decision on this CR has not yet been made. As a result of having footprints of all classes available at LOD0, the LandUse functions that are not related to terrain but represent other classes that are at the moment missing at LOD0, should be removed from the LandUse class. A change request has been submitted to cover this issue, combined in one change request with the previous, related issue.

Enriching LandUse class with land cover information would improve the semantics for this class significantly. Information on the usage of land is sufficiently addressed in CityGML, but land cover is not. The code list

[3][Added 2018] The guidelines were published as an OGC Best Practice in 2014.

values for the usage often refer to land cover, i.e. the physical appearance. This is often different from the way the land is used, analogue to traffic areas which do have function, usage, as well as physical appearance properties in CityGML. In the INSPIRE domain models Land Cover and Land Use are modelled as separate entities, which might be even better. A change request has been submitted to cover this issue. It has been addressed in CityGML 1.1 by adding a paragraph stating that the LandUse class can also be used for land cover information.

The semantics in CityGML could be enriched with an additional class for constructions other than buildings. In the 3D IMGeo model such a class was added, i.e. Construction as a specialisation of _Site. These other constructions are human-built objects, that are not 'buildings' in the normal meaning: e.g. they have no roof, doors, windows etc, may not be able to be closed off, etc. The INSPIRE Building data specification also has a class OtherConstruction (Building, 2011). This issue is partly addressed by earlier change requests:

- Harmonisation with Inspire Themes (Roensdorf, 2010a)
- Thematic module for bridges (Nagel, 2010b)
- Thematic model for man-made subsurface structures (Nagel, 2010a)
- Thematic model for walls in cities (Nagel, 2010c)

Since those thematic models do not cover all types of Constructions represented in IMGeo, a change request has been submitted to cover this issue in line with these earlier requests. At the time of writing a decision on this CR has not yet been made.

2.7 Conclusions and further research

This paper presents the development of a national standard for 3D geo-information as one of the main results of an extensive collaboration pilot in the Netherlands. The standard aligns both to the existing 2D standardisation framework and the OGC standard CityGML. Aligning to existing 2D information models is considered as important to reuse the valuable concepts as defined in those domain models. In May 2011 the full integration with CityGML was approved as modelling approach for large scale topography in the Netherlands. The resulting model was formally approved in February 2012. From this moment on it will be the standard for both 2D and 3D large scale topography in the Netherlands. This close integration between an

existing information model for 2D geo-information and CityGML is a major step for 3D use and a unique achievement for standardisation in 3D. The involvement of many stakeholders in the development of the standard appeared to be essential to obtain the necessary support for the national 3D standard.

The main contributions of this paper can be characterised as a nationwide 3D standard on geo-information, covering all topographic classes, extending OGC CityGML standard, while integrating the different CityGML LODs with a 2D LOD and preserving the concepts of the 2D information model. Other main contributions are the generic findings deduced from our experiences: the proposed framework for extending CityGML for national purposes (Section 2.5) and change requests submitted to OGC to make the CityGML standard a better fit for integration with 2D topographic information models (Section 2.6).

In the development of CityGML-IMGeo a number of aspects were identified for further research.

Firstly, follow-up research is needed to test the 3D IMGeo model on its usability in practice including the consequences of this new modelling method for IMGeo when used for both 2D and 3D datasets, e.g. how to preserve the links between the different LODs and how to upgrade 2D LOD to higher LODs (see for example Van Oosterom and Stoter (2010)). Secondly, technical issues must be tested such as the ability to use 3D IMGeo data in CityGML-aware software. The new insights will lead to further refinements of the model.

Also the other domain information models need to be studied for extension into 3D when appropriate. Besides specific IMGeo aspects, CityGML-IMGeo contains generic aspects that can be reused for extending other domain information models such as for 3D cadastre (Stoter and Salzmann, 2003; Stoter and Ploeger, 2003; Döner et al., 2010, 2011) and 3D noise mapping (Stoter et al., 2008).To get a feeling of the possibilities a study was carried out in the context of the 3D pilot on how to match the concepts of the other Dutch domain models to CityGML. For this matching also several ADEs were considered, such as the extensions for Tunnel and Bridge. In the future the planned ADE for cables and pipelines may be relevant (Becker et al., 2010; Hijazi et al., 2011). A proposal for support of voxels in CityGML is already in preparation. This proposal builds on Tegtmeier et al. (2009) and Zobl and Marschallinger (2008).

Other aspects that warrant further research are:

- use of the CityGML Relief module to represent specific aspects of elevation (not used in current approach);

- use of the Tunnel and Bridge ADEs for more detailed modelling of these types of constructions;
- classify certain types of LandUse using the Vegetation class;
- define constraints to exclude CityGML concepts that are not used by IMGeo (or other external model), such as LOD4 geometry;
- define constraints to make optional properties of CityGML classes mandatory and to relate them to class properties of IMGeo (or of another external model);
- describe use, maintenance and validation of code lists;
- define constraints to link attributes that inherit from a CityGML class to added attributes with specific codelists of IMGeo (or of another external model).

Finally, more research is needed concerning the creation and management of CityGML-IMGeo data. Which methods can be used to generate CityGML-IMGeo data? How should this data be maintained? How can 2.5D topology be created, validated and maintained? Best practice documents could be a useful result of such studies. Therefore a new project has been started[4] (more than 140 participants have subscribed) in which the results as presented in this paper are being made ready for use in practice by addressing the open issues as described above.

Bibliography

J. Albert, M. Bachmann, and A. Hellmeier. Zielgruppen und Anwendungen für Digitale Stadtmodelle und Digitale Geländemodelle., 2003.

Thomas Becker, Claus Nagel, and Thomas H Kolbe. Integrated 3D modeling of multi-utility networks and their interdependencies for critical infrastructure analysis. In *Advances in 3D Geo-Information Sciences*. Springer, 2010.

J Benner, A Geiger, and K Leinemann. Flexible generation of semantic 3D building models. In *Proceedings of the 1st international workshop on next generation 3D city models, Bonn*, pages 17–22, 2005.

[4][Added 2018] This phase of the 3D pilot would take place in 2011-2012 and focus on creation, management and validation of CityGML-IMGeo data.

INSPIRE Thematic Working Group Building. D2.8.iii.2 Data Specification on Building – Draft Guidelines. Available online: http://inspire.ec.europa.eu/documents/Data%5FSpecifications/ INSPIRE%5FDataSpecification%5FBU%5Fv2.0.pdf (accessed January 2012), 06 2011.

Open Geospatial Consortium et al. Geography Markup Language (GML) Encoding Standard. Version 3.2.1, doc nr OGC 07-036 [online]. Available online: http://portal.opengeospatial.org/files/?artifact%5Fid=20509, 2007.

Fatih Döner, Rod Thompson, Jantien Stoter, Christiaan Lemmen, Hendrik Ploeger, Peter van Oosterom, and Sisi Zlatanova. 4D cadastres: First analysis of legal, organizational, and technical impact—With a case study on utility networks. *Land Use Policy*, 27(4):1068–1081, 2010.

Fatih Döner, Rod Thompson, Jantien Stoter, Christiaan Lemmen, Hendrik Ploeger, Peter van Oosterom, and Sisi Zlatanova. Solutions for 4D cadastre–with a case study on utility networks. *International journal of geographical information science*, 25(7):1173–1189, 2011.

KL Emgard and S Zlatanova. Design of an integrated 3D information model. *Urban and regional data management: UDMS annual*, pages 143–156, 2007.

Geonovum. Informatiemodel Geografie, 10 2007.

Geonovum. Geonovum Standaarden. Available online: www.geonovum.nl/wegwijzer/standaarden (accessed January 2012), 2012.

Gerhard Gröger, Thomas H Kolbe, and Lutz Plümer. Mehrskalige, multifunktionale 3D-Stadt-und Regionalmodelle. *Journal of photogrammetry, remote sensing and geoinformation processing*, 2004(2), 2004.

Gerhard Gröger, Thomas H Kolbe, and A. Czerwinski. Candidate OpenGIS CityGML Implementation Specification (CityGeography Markup Language), 2007.

Gerhard Gröger, Thomas H Kolbe, Angela Czerwinski, and Claus Nagel. OpenGIS city geography markup language (CityGML) encoding standard, version 1.0.0. Available online: http://portal.opengeospatial.org/files/?artifact%5Fid=28802 (accessed 11 February 2014), 2008.

Ihab Hijazi, Manfred Ehlers, Sisi Zlatanova, Thomas Becker, and Léon van Berlo. Initial investigations for modeling interior Utilities within 3D Geo Context: Transforming IFC-interior utility to CityGML/UtilityNetworkADE. In *Advances in 3D Geo-information sciences*, pages 95–113. Springer, 2011.

ISO. Information technology–document schema definition language (DSDL)–Part 3: Rule-based validation – Schematron, 2006.

Claus Nagel. Change request nr 10-048 [online]. Available online: www.opengeospatial.org/standards/cr, 2010a.

Claus Nagel. Change request nr 10-051 [online]. Available online: www.opengeospatial.org/standards/cr, 2010b.

Claus Nagel. Change request nr 10-053 [online]. Available online: www.opengeospatial.org/standards/cr, 2010c.

OGC. The OGC seeks comment on City Geography Markup Language (CityGML) V1.1 [online]. Available online: http://www.opengeospatial.org/standards/requests/82, 2011.

Sander Oude Elberink. *Acquisition of 3D topography: automated 3D road and building reconstruction using airborne laser scanner data and topographic maps*. University of Twente ITC, 2010.

Sander J Oude Elberink and George Vosselman. 3D information extraction from laser point clouds covering complex road junctions. *The Photogrammetric Record*, 24(125):23–36, 2009.

Friso Penninga and PJM Van Oosterom. A simplicial complex-based DBMS approach to 3D topographic data modelling. *International Journal of Geographical Information Science*, 22(7):751–779, 2008.

Clemens Portele. Ogc® OWS-6 UTDS-CityGML Implementation Profile version 0.3.0, OGC 09-037r1, 2009.

Clemens Portele. Change request nr 09-039 [online]. Available online: www.opengeospatial.org/standards/cr, 2010.

Rijkswaterstaat. Rijkswaterstaat Eisen digitaal topografisch bestand. Available online: https://www.rijkswaterstaat.nl/kenniscentrum/contracten/data%5Feisen/digitaal%5Ftopografisch%5Fbestand (accessed January 2012), 2011.

Carsten Roensdorf. Change request nr 10-062 [online]. Available online: www.opengeospatial.org/standards/cr, 2010a.

Carsten Roensdorf. Change request nr 10-007 [online]. Available online: www.opengeospatial.org/standards/cr, 2010b.

Jantien Stoter and Martin Salzmann. Towards a 3D cadastre: where do cadastral needs and technical possibilities meet? *Computers, environment and urban systems*, 27(4):395–410, 2003.

Jantien Stoter, Henk De Kluijver, and Vinaykumar Kurakula. 3D noise mapping in urban areas. *International Journal of Geographical Information Science*, 22(8):907–924, 2008.

Jantien Stoter, George Vosselman, Joris Goos, Sisi Zlatanova, Edward Verbree, Rick Klooster, and Marcel Reuvers. Towards a national 3D Spatial Data Infrastructure: case of the Netherlands. *Journal of photogrammetry, remote sensing and geoinformation processing*, 2011(5), 2011.

Jantien E Stoter and Hendrik D Ploeger. Property in 3D—registration of multiple use of space: current practice in Holland and the need for a 3D cadastre. *Computers, Environment and Urban Systems*, 27(6):553–570, 2003.

W Tegtmeier, S Zlatanova, PJM Van Oosterom, and HRGK Hack. Information management in civil engineering infrastructural development: with focus on geological and geotechnical information. In *Proceedings of the ISPRS workshop*, volume XXXVIII-3-4/C3 Commission III/4, IV/8 and IV/5: Academic track of GeoWeb 2009 conference: Cityscapes International archives of photogrammetry, remote sensing and spatial information sciences, pages 1–6, 2009.

Peter Van Oosterom and Jantien Stoter. 5D Data Modelling: Full Integration of 2D/3D Space, Time and Scale Dimensions. In *International Conference on Geographic Information Science, Lecture Notes in Computer Science*, volume 6292, pages 310–324. Springer-Verlag Berlin Heidelberg, 2010. ISBN 978-3-642-15299-3.

CityGML Wiki. Application Domain Extensions. [maintained online], Available online: www.citygmlwiki.org/index.php/CityGML-ADEs (accessed July 2012), 2012.

Siyka Zlatanova, Michael Gruber, and M Kofler. Merging DTM and CAD data for 3D Modeling purposes for Urban Areas. In *Proceedings of ISPRS*, volume XXXI, pages 311–15, 1996.

Fritz Zobl and Robert Marschallinger. Subsurface geobuilding information modelling GeoBIM. *GEOinformatics*, 8(11):40–43, 2008.

Chapter 3

UML-Based Approach to Developing a CityGML Application Domain Extension

Authors: L. van den Brink, J. Stoter & S.Zlatanova. (*Transactions in GIS*, 17(6):920-942, 2013).

This paper has been published in a peer-reviewed scientific journal in 2013 and is published unchanged in this chapter, except for footnotes where necessary to supply current information about an outdated statement. The paper extends the research of the first paper and focuses specifically on the modelling of the national base topography standard as an extension of the UML class diagrams of CityGML. It focusses on how a national 3D standard can be best described as an extension of the international standard CityGML. Specifically, the Application Domain Extension (ADE) mechanism is applied to Model Driven Architecture (MDA) UML modelling of CityGML extensions. The paper studies several alternatives for a model-driven framework to model CityGML ADEs and selects the most optimal one.

Contributions: The UML-based ADE mechanism was adopted as OGC best practice and is currently used for other ADEs, such as the CityGML Utility Network Ade (Kutzner and Kolbe, 2016), the CityGML Energy ADE (Nouvel et al., 2015), The Land Administration Domain Model (LADM) CityGML ADE (Rönsdorff et al., 2014), and an ADE for immovable property taxation (Çağdaş, 2013).

text of published paper starts after this line

Abstract Recently[1] a national 3D standard has been established in the Netherlands as a CityGML Application Domain Extension (called IMGeo). In line with the Dutch practice of modelling geo-information, the ADE is developed using a model driven approach. The classes are designed in UML and automatically mapped to GML schema. The current OGC CityGML specification does not provide rules or guidance on correctly modelling an ADE in UML. This paper fills this gap by studying how CityGML can be extended for specific applications starting from the UML diagrams. Six alternatives for modelling ADEs in UML are introduced and compared. The optimal alternative is selected and applied to obtain the national 3D standard. The approach was extensively discussed with international experts, who were members of both SIG3D and other working groups. As a consequence the approach was adopted by the SIG3D, the Special Interest Group 3D which, among other things, work on the 3D standard CityGML in cooperation with OGC. Therefore the approach contains many issues that can be generalized and reused by future domain extensions of CityGML. To further support this, this paper formulates a model-driven framework to model CityGML ADEs. Open issues are described in the conclusions.

Keywords: 3D, Open GIS consortium, Unified Modeling Language, three-dimensional, geo-information, standard, CityGML

3.1 Introduction

In the last several years a broad discussion has started in the Netherlands on the need to develop a national standard for a 3D city and landscape model. The discussion is summarised in a number of papers (Stoter et al., 2010, 2011, 2013a; Verbree et al., 2010). The main question was whether to develop a separate 3D model from scratch or to re-use and extend already existing 3D standards. The study on international and vendor specific 3D standards and formats has clearly revealed that CityGML is a good candidate for a national 3D standard (Stoter et al., 2013a; Zlatanova et al., 2012). The OGC standard CityGML (Gröger et al., 2008, 2012) is an application independent information model and exchange format for 3D city and landscape models. It maintains semantics, geometry, topology and appearance of objects (Stadler and Kolbe, 2007). Furthermore CityGML is supported by an increasing number of vendors by providing import/export functionalities and viewers (Lapierre and Cote, 2007; Rumor and Roccatello, 2009; Groneman and Zlatanova, 2009). CityGML database implementations are also available

[1][Added 2018] IMGeo 2.0, in which the 3D standard was included, was published in February 2012.

(de Vries and Zlatanova, 2011). Considering all these aspects, it was decided to not generate a separate 3D model but to align the existing national 2D information model to CityGML. The CityGML standard is meant as a generic standard for modelling topographic features. Domain specific information can be modelled in CityGML either by generic classes or by the definition of an extra formal schema based on the CityGML schema definitions. Such a schema is called a CityGML Application Domain Extension (ADE). The second approach allows definition of classes, their relationships and attributes and is recommended for applications that require a large number of new features to be defined. Therefore the recently developed national 3D standard in The Netherlands is also based on the ADE mechanism (Van den Brink et al., 2012; Stoter et al., 2011). This ADE extends CityGML with the existing 2D national Information Model for large-scale Geo-information (called 'IMGeo').

IMGeo is modelled in UML (Unified Modelling Language) and contains object definitions for large-scale representations of roads, water, land use/land cover, bridges, tunnels etc. and their properties. It prescribes 2D point, curve or surface geometry for all objects. The original IMGeo classes were compatible with CityGML but there were also quite some differences. The new version of IMGeo (version 2.0) is developed as specialization of CityGML. Using this approach 2D IMGeo data can be extended into 2.5D (i.e. as height surface representation) and 3D (as volumetric representation) according to geometric and semantic principles of CityGML.

The development of the ADE encountered two major challenges: designing ADE in UML, and automatically deriving the GML schema from the UML diagram. According to the Dutch practice of modelling geo-information, all information models must be represented with UML. However, the CityGML specification does not provide rules or guidance on modelling an ADE in UML. It describes how an ADE must be modelled in the XML schemas, which is not compatible with UML modelling. A complete description of the CityGML-IMGeo ADE (i.e. how the model was established) can be found in Van den Brink et al. (2012).

This paper presents the generic technical modelling principles of the ADE that were encountered during the establishment of the national 3D standard, i.e. how the UML models of CityGML can be extended to support concepts defined in a specific domain (e.g. noise (Stoter et al., 2008) and 3D cadastre (Stoter and Van Oosterom, 2005; Stoter et al., 2013b)), and how a GML application schema can be automatically generated from the UML model. We have followed strict rules and reached the solution proposed in this paper together with international teams, which are involved in GML and CityGML modelling. Furthermore the proposed modelling approach is adopted by the

SIG3D and our experiences are accepted as OGC best practice to standardise the developments of domain specific CityGML ADEs (Van den Brink et al., 2012). Therefore the approach described in this paper may serve as generic approach. Compared to the OGC best practice paper, this paper contains more detailed descriptions of the alternative approaches. In addition the application to derive new ADEs (by giving the example of the new Dutch IMGeo standard) has been added.

The paper is organized as follows. Section 3.2 describes the background of the study, i.e. the Dutch context in which the ADE is modelled, including the UML-based approach for modelling geo-information. Section 3.3 compares several alternatives for modelling an ADE in UML and selects the optimal modelling approach. Section 3.4 explains how the selected modelling approach has been applied to model the CityGML ADE 'IMGeo 2.0'. Although IMGeo is country-specific, our approach to model a CityGML ADE in UML and to automatically generate a GML schema accordingly can be seen as a standard approach for designing a CityGML ADE. In order to make our approach reusable for future domain extensions of CityGML, in Section 3.5 we propose a model-driven framework for modelling CityGML ADEs in UML. Section 3.6 concludes on findings and topics for further research.

3.2 Explanation of the Dutch context

The Dutch standardisation organization for geo-information (Geonovum) follows a formal approach for modelling and implementing Information models. Such an approach is currently not provided by the SIG3D and OGC CityGML SWG, which is responsible for the development of the CityGML standard. Therefore, in the following section we further explain the generic UML modelling approach for geo-information in the Netherlands. Then the Information Model IMGeo is briefly presented and the most important features are highlighted. The last section describes the implications of our UML modelling approach for the IMGeo ADE.

3.2.1 Model driven approach

Formal representation of conceptual models for geo-information applying UML is seen as an important prerequisite of the Dutch Spatial Data Infrastructure (SDI). UML is one of the most used modelling languages by standardisation bodies dealing with geo-information. Using UML class diagram, geo-information objects can be formally described with their properties, relationships and semantic meaning. A good understanding of the

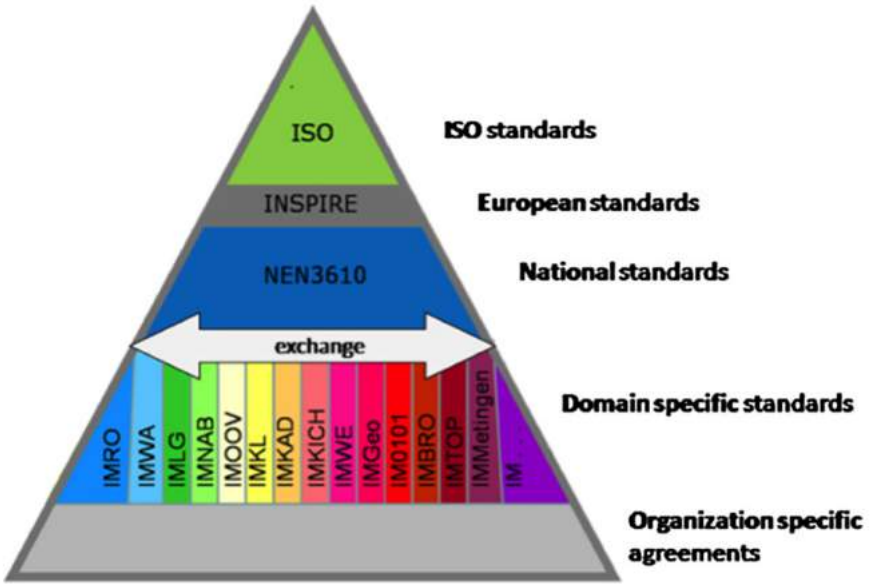


Figure 3.1: Pyramid of information models. The abbreviations are mnemonic names for Dutch standards e.g. IMRO = Information model ruimtelijke ordening (spatial planning); IMWA=Information model water; IMGeo=Information model geography.

meaning of objects is especially important when different organisations reuse each other's information. Although not as elaborated as some modelling languages focused on semantics (such as OWL, RDF, etc.,), UML provides sufficient means to record the meaning of objects.

In the Netherlands, the Base Model Geo-Information (NEN, 2011) forms a common base for domain specific information models. This national standard establishes a standard modelling method based on ISO 19101 and contains a generic semantic UML model with definitions of the most common, shared concepts in the geo-domain such as Road, Water, etc. Since NEN 3610 was first published in 2005, many domain specific information models have been developed. These domain models, such as the information model for spatial planning (IMRO) or the information model water (IMWA) extend the classes defined in NEN 3610 with properties and more specific classes. IMGeo is also one of these information models. The semantic geo-standards in the Netherlands can be viewed as a pyramid of information models (see Figure 3.1), which forms a common language across organisations containing definitions of concepts from different domains. Because these are all based on the shared definitions in NEN 3610, exchange of information across organizations, based on different information models in the pyramid is possible.

In the Netherlands, a Model Driven Approach (MDA) is applied for modelling concepts and their implementation in different domains (Gaševic et al., 2006). A key point of this approach is either that the conceptual information models are independent of their technical implementation(s) or they are platform-independent (OMG, 2003; Hespanha et al., 2008). This means that the technical implementations (for data storage or data exchange) are automatically created from the UML schemas of the domain models. For data exchange based on these models, Geography Markup Language (GML) is used. The technical implementations (in this case GML application schemas) are not designed and maintained separately, but are automatically derived from the UML models using the Java tool ShapeChange (Portele, 2008). ShapeChange implements the UML to GML encoding rules described in ISO 19136:2007, ISO/TS 19139:2007, ISO 19118 rev 1, ISO/TS 19103:2005, and ISO 19109:2005.

3.2.2 Information Model Geography (IMGeo)

The Dutch Information model Geography (IMGeo) describes how object-based, large-scale (between scale 1:1000 and 1:2000) topographic features must be defined to make the national exchange of this information possible. This large-scale topographic map is created and maintained by the municipalities and several other public organizations, and financially supported by local governments and private companies. Version 1.0 of IMGeo was published in 2007 (Geonovum, 2007). Version 2.0 was completed end 2011 and published in February 2012. IMGeo 2.0 has a mandatory core, see Figure 3.2 and Appendix A. The mandatory core model contains object definitions for large-scale representations of roads, water, land use, land cover, bridges, tunnels etc. The optional part of IMGeo allows further division of these objects into parts suitable for maintenance, and contains definitions for all kinds of city furniture and other non-mandatory classes.

Data providers such as municipalities, organisations responsible for the road, water and railway infrastructure etc. are required by law to provide their objects that fall under the definitions of the IMGeo 2.0 core to a national 'basic registry' (Basisregistratie Grootschalige Topografie, BGT) where they are available for reuse.

All IMGeo object types have 2D geometry for which ISO 19107 geometry types are used. Topological rules are part of the standard, but are not modelled by ISO topological types. The most notable rule is that the complete set of polygon-objects at surface level (height level 0) in the mandatory core must together form a complete coverage of the Netherlands without gaps or overlap.

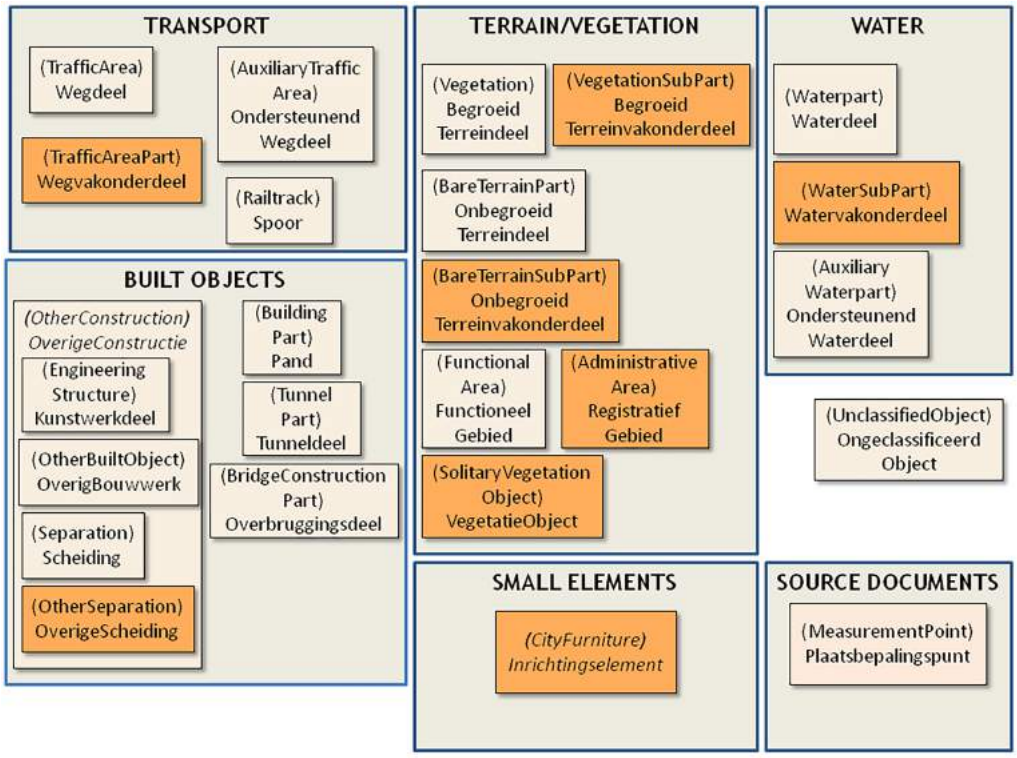


Figure 3.2: Conceptual overview of the classes in the Dutch Information Model Geography (IMGeo); classes in light colour are part of the mandatory core; the classes in orange are optional. English translation of class names is shown in parentheses (i.e. as alias).

IMGeo classes have a limited number of attributes. Most of the classes have one or two attributes to further classify the object or to indicate their function. Code lists are used to provide allowed values. In addition, all classes share attributes for identification and versioning, for a reference to the data provider, for the object's status (planned, existing, or historic) and an indication whether a possible error in the data is under investigation. In addition, all measurements are stored with metadata such as information on the accuracy. The IMGeo model also includes rules for visualization, e.g. the colour of lines and areas, type and thickness of lines, etc.. These are based on a Dutch standard for web cartography. Three visualization themes are provided: one vector-based representation where large-scale topography is leading, and two where it is used as a backdrop map for other themes.

3.2.3 Implications of the Dutch UML approach for the CityGML ADE

As mentioned above, it was unclear how exactly the CityGML UML diagrams can be extended. Although it is not explicitly specified in the CityGML specifications, there are examples of CityGML ADE modelling in UML (Wiki, 2012). Most of these ADEs extend CityGML with one or more new feature types, while our IMGeo ADE should (also) extend existing CityGML feature types with domain specific properties. In the CityGML standard, the informative Annex G shows one example of an ADE using UML diagrams: the Noise ADE. In this example, new classes are added, but also properties are added directly to the CityGML Building class. These extra properties are visually distinguishable from the Building properties defined in CityGML by using a different colour and addition of a namespace prefix 'noise:' to the property names.

Although this results in a visually understandable diagram, from a modelling perspective these properties are in fact part of the CityGML Building class, and not part of the ADE. In UML there is no concept that allows a class having a number of properties in one package, while other properties of the same class reside in another package. When generating a GML application schema from these diagrams using the ISO 19118 rules, the properties would be added to the CityGML namespace and not the ADE namespace. Therefore we had to look for another option, which is studied further in the remainder of this paper.

3.3 Extending CityGML UML diagrams with application specific concepts

To understand how the problem of modelling ADEs in UML can be addressed, we first further explain the technical issues. Then we present and evaluate different alternatives for extending CityGML UML diagrams and finally select the best alternative. For clarity some basic UML terms used throughout the text are introduced first. See also Figure 3.3.

A set of objects with shared characteristics (attributes, associations and behaviour) is modelled as a class. A class has a name, attributes, relations with other classes and operations. Classes can be abstract or concrete. Abstract classes (represented in UML diagrams with their class name in italics) are used to model super categories with characteristics shared by several subclasses (specialization). The subclasses inherit all characteristics of the superclass. Individual objects (object instances) are always members of a concrete class (NEN, 2011).

The properties of a class and its relations to other classes are modelled as attributes or associations. Attributes have a name and a value, and their cardinality indicates how many times they may occur (zero, one or many times). Both ends of an association have a cardinality and a name.

Stereotypes are meta properties of classes, attributes and associations, used to semantically extend UML for a specific domain. In diagrams they are shown between angled brackets at the element they apply to. Another way to extend UML is by using tagged values that can be used to add properties to classes, attributes and associations. Often tagged values are used to add properties relevant for automatic code generation (ISO19103). Tagged values are usually not shown in diagrams.

Code lists and enumerations are special types of classes, which both give a list of allowed values for a specific class attribute. In the case of an enumeration the list may not be extended, while for a code list this is allowed.

Finally, hook elements are not a standard UML concept, but a concept from CityGML and XML modelling. Basically this is an abstract, extensible class property, which can be re-used for many classes with different definitions. Hook elements are explained further in the next section.

3.3.1 Detailed technical explanation of the problem

The main challenge that we had to face was that 'hook elements', the CityGML approach for adding properties in XML Schema as described in Gröger et al. (2008): 10.13.1, do not have an equivalent in UML modelling. Consequently,

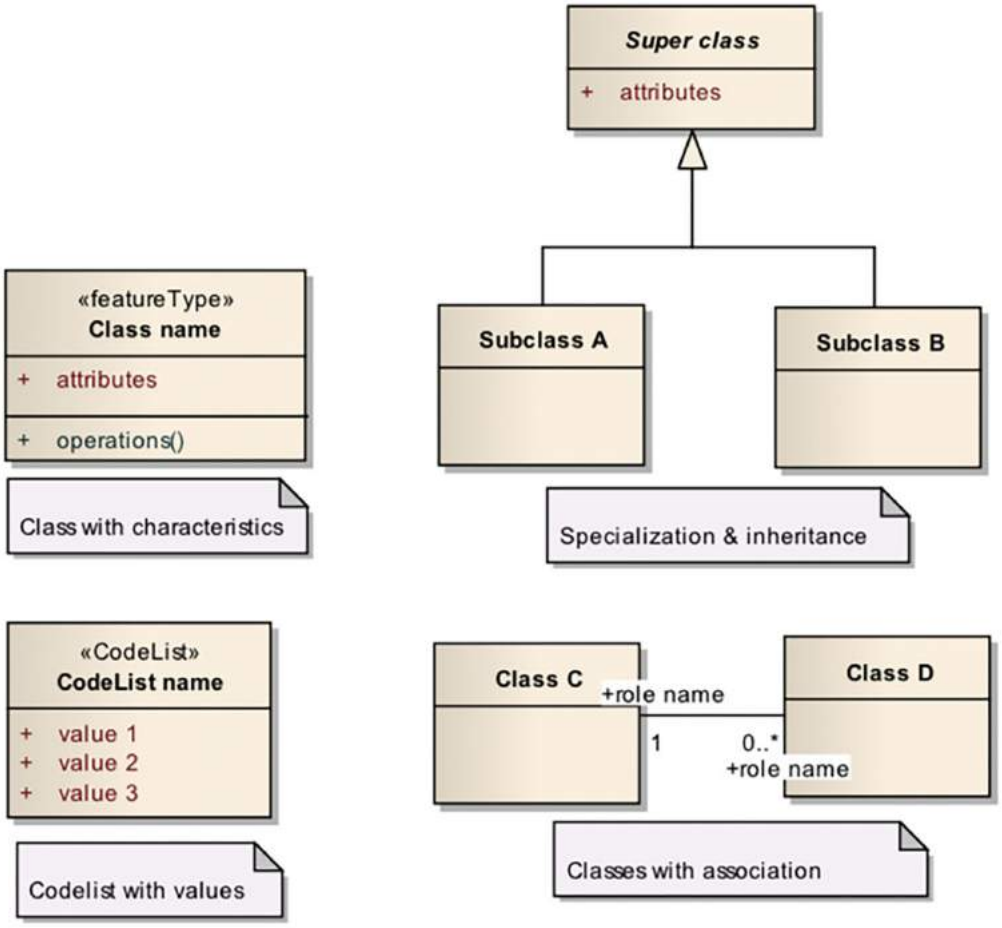


Figure 3.3: UML overview

we also had to design an encoding rule that establishes a mapping between the method for adding properties to CityGML classes in a UML ADE and the way this is encoded in an XML Schema according to the CityGML standard. It should be noted that CityGML 1.0 (Gröger et al., 2008) was the valid CityGML version at the moment of modelling our CityGML ADE. However, developments on this issue were still ongoing when CityGML 2.0 was established (2012) and therefore the problems of modelling an ADE in UML remain in the new version of CityGML.

The reason the CityGML approach for adding properties in XML Schema does not have an equivalent in UML modelling can be explained as follows. For creating XML schemata based on UML models we follow ISO 19118 encoding rules. According to these rules, UML classes are represented by a global element in the XML schema and UML class attributes are represented by local elements in the global element's complex type. UML class specialisation is represented by the complex type of a subclass extending the complex type of its super class. In addition, the corresponding global XML element of a UML subclass is added to the substitution group of the element representing the UML super class. In this way it may appear anywhere in the content model where the super class element is allowed. As noted in Gröger et al. (2008); 10.11.1, a disadvantage of this approach is that two different subclasses, each with their own extra attributes with regard to their common super class, cannot be used together. Either one or the other must always be chosen, and thus attributes from both subclasses cannot be combined. Also, applications must have knowledge of the XML Schema to recognise the subclass as a more specific form of a super class from the CityGML standard.

Because of this, CityGML has a different approach for extending classes with additional properties. Basically, the approach chosen in CityGML is a form of an attribute substitution. Each CityGML class has a 'hook' in the form of an abstract, global XML child element which is referenced from each CityGML class complex type and which can be replaced an arbitrary number of times with ADE property definitions. In the ADE XML Schema, such a property is defined by extending the 'hook' element's complex type and by placing the property element in the 'hook' element's substitution group (Gröger et al., 2008); 10.11.1. This is the same technique as is normally used for mapping UML class specialization to XML Schema, as described above.

While this is perfectly normal in XML Schema, UML has no concept of sub-classing properties and there is nothing in UML or in the ISO 19100 series that is similar to global properties and/or substitution of properties. Therefore, a way of representing this in UML had to be found during the modelling of the IMGeo ADE.

Several approaches to model application specific concepts of an ADE

in UML have been considered and intensively discussed within the SIG3D modelling subgroup and in email discussions by the authors, SIG3D members and other participants. The alternatives for modelling ADEs in UML class diagrams, which featured in these discussions, are introduced and evaluated in the next sections.

3.3.2 Alternatives for modelling ADEs in UML

Alternative 1: Use the extension possibilities of GenericCityObject and _genericAttribute (Gröger et al., 2008); 10.10

Advantage: No extra modelling work is required.

Disadvantage: As noted in the CityGML standard (Gröger et al., 2008): 10.11, the extension would not be formally defined with names, definitions, and types, and it would therefore be impossible to validate data that uses such extensions.

Alternative 2: Add properties in the CityGML classes directly in the CityGML package instead of in an own ADE package (discussed in the SIG3D). The added properties would be marked as ADE extensions in the UML model using a stereotype or tagged value.

Advantage: No subclass definition is necessary.

Disadvantage: There are two disadvantages: a) It is in conflict with UML. Packages are namespaces and reflect governance. The CityGML packages are controlled by the CityGML SWG, an ADE package by some other authority. That authority cannot edit the CityGML packages. b) It does not meet the requirements of the modular specification standard of OGC (OGC, 2009) for very similar reasons. In particular, see section 7.2.2.

Alternative 3: Add properties in a subclass in the ADE package but suppress this subclass from the generated XML Schema

Advantage: This approach does not violate UML, ISO 19100 series, and OGC rules.

Disadvantage: It is confusing to introduce an ADE subtype of a CityGML type although the ADE hooks provide a means to avoid subtyping; and also because in UML a subclass inherits all methods and attributes from its super class, but in this case this is not intended. The class would be marked as ADE extension in the UML model using a stereotype or tagged value. A stereotype is preferred because it makes clear from the UML diagrams that the ADE subtype is not mapped to an XML Schema component. Tagged values are not always (and usually not) shown in the graphical notation. However, this

could be viewed as violating the GML encoding rule that stereotypes are used for conceptual aspects and tagged values for encoding-related aspects.

Alternative 4: Define the ADE hook '_GenericApplicationPropertyOf...' as a class associated to the CityGML class and add properties as subclass of the ADE hook class.

Advantage: this way of modelling provides a clear distinction between the concept of sub typing a CityGML class and extending a CityGML class with properties; any confusion is avoided.

Disadvantage: It is less clear than sub typing the CityGML class, which is a well-known way of modelling. Furthermore it is not in line with the ISO 19109 General Feature Model, where there is no concept of attribute substitution.

Alternative 5: Define the ADE hook '_GenericApplicationPropertyOf...' as a class associated to the CityGML class and add properties in an abstract superclass.

*Advantage:*This avoids the problem of using a subclass while not intending inheritance of properties.

*Disadvantage:*The generalization relationship would have to be added to the CityGML class which violates basic UML and XML namespace governance rules, i.e. an ADE cannot make changes to the CityGML package which is controlled by the CityGML SWG at OGC.

Alternative 6: Define a general type ADEPropertyType with name, definition, type, and extendsType and maintain the added types outside the UML in a registry.

Advantage: This avoids violating UML rules, the General Feature Model (ISO 19109) and other rules from the ISO19100 series.

Disadvantage: It is necessity to maintain added feature properties outside the UML. The ADE extension cannot be completely modelled in UML. Extension with extra classes would be part of the UML model, but extension with properties would not. Also, the feature properties that are modelled outside UML would still have to be included somehow in the generated XML Schema.

3.3.3 Conclusion on the alternatives: best approach

After comparing the advantages and disadvantages of the above alternatives, alternative 3 has been selected as the best option for the IMGeo ADE. Both

the authors and the SIG3D decided to implement this option. This approach defines the to be added properties in subclasses in the ADE package but suppresses these subclasses from the generated XML Schema. There are several reasons why we have chosen this approach.

Firstly, conceptually IMGeo is an extension of CityGML and therefore defining the IMGeo classes as subclasses of CityGML classes and adding the extra properties to these subclasses is appropriate. Another aspect in favour of this alternative is that the use of sub classing is understandable for people with basic knowledge of UML class diagrams. This is an important requirement of the IMGeo UML model. In addition, this approach conforms to relevant rules of UML, the ISO 19100 series and OGC unlike most of the alternatives described previously. Finally this approach is the most in line with the current geo-information modelling approach in the Netherlands.

The fact that in the XML Schema implementation the subclasses are omitted, is seen as a technical implementation choice to allow the combining of properties from different ADEs. While this is a valid reason on the technical level, it is not taken to mean that in the conceptual UML model sub classing should also be avoided.

3.4 Modelling IMGeo as CityGML ADE

This section presents the procedure that was followed to model IMGeo as an ADE of CityGML in UML using the selected approach. In the following sections we will elaborate on the modelling of classes, subclasses, code lists, geometry and topology. Finally we explain how the XML Schema can be automatically generated.

3.4.1 Modelling IMGeo classes as subclasses of CityGML classes

Since CityGML 1.0 (and also 2.0) is only available as XML Schema, the first step was to recreate the UML model in the modelling tool Enterprise Architect, based on Gröger et al. (2008). In the next step all IMGeo classes were modelled as subclasses of CityGML classes. Using the selected modelling approach, these subclasses get the same class name as the CityGML class they are extending. The stereotype <<ADEElement>> is assigned to the subclasses and the specialization relation is marked with a stereotype <<ADE>>. Having a stereotype also on the specialization relation, as well as the names of these stereotypes were proposed by members of the SIG3D during the discussion on this issue. The stereotypes mark these classes as

special subtypes that only add properties to the CityGML class, and accordingly no XML component for these classes will be created in the XML Schema. For documentation purposes, a Dutch translation of the class name is added as an alias. For all CityGML classes, relevant for IMGeo, a subclass is created, adding at least a 2D geometry property to all classes.

Figure 3.4 shows an example in the IMGeo ADE of a subclass TunnelPart that contains additional properties compared to the equivalent CityGML class (2D geometry and LOD0 geometry properties). The yellow classes are classes from the CityGML Tunnel package. The <<ADEElement>> TunnelPart is a class defined in the IMGeo ADE package as a subclass of CityGML TunnelPart class. The Dutch alias is shown between brackets on the class diagram.

By applying this inheritance structure the domain specific information model gets the same structure as defined by the CityGML model, see Appendix A.

To identify equivalent concepts that can be modelled via this sub classing method, first a conceptual mapping was made between CityGML and IMGeo classes. These mappings compared the concepts at semantic level, i.e. independent of at which LOD the concept appears in CityGML. The CityGML Levels of Detail (LOD) concept is used to model objects with different accuracy for different purposes between LOD0 (terrain) to LOD4 (interior); where LOD1-LOD3 also represent volumetric properties. Some classes only get spatial representations at higher LODs in CityGML and because we also wanted to take these concepts into account, we compared the IMGeo and CityGML classes independent of the LOD it appears.

Obviously, not for every IMGeo class a 1-to-1 mapping to an equivalent CityGML class could be found. For these classes, two solutions are possible. The first option, which is preferred and therefore applied as much as possible, remodels the IMGeo concept so that an equivalent CityGML class can be found. For IMGeo this is for example done for Vegetation that models any vegetation-related concept (in IMGeo 1.0 divided over several classes) and AuxiliaryTrafficArea meant for road segments that are not used for traffic, such as verges (in IMGeo modelled under the classes Road or Land Use).

If it is not possible to remodel the concept into a CityGML class, CityGML is extended with a new class, as a subclass of one of the CityGML classes. In this case, because a whole new class is added and not just properties to an existing CityGML class, the hook mechanism is not used; instead the class is modelled as subclass of a CityGML class with stereotype <<featureType>> and not suppressed from the XML Schema. Figure 3.5 shows an example in the IMGeo ADE of a class, which is not available in CityGML but needed in IMGeo. The class 'OverigeConstructie' (OtherConstruction) is a class to

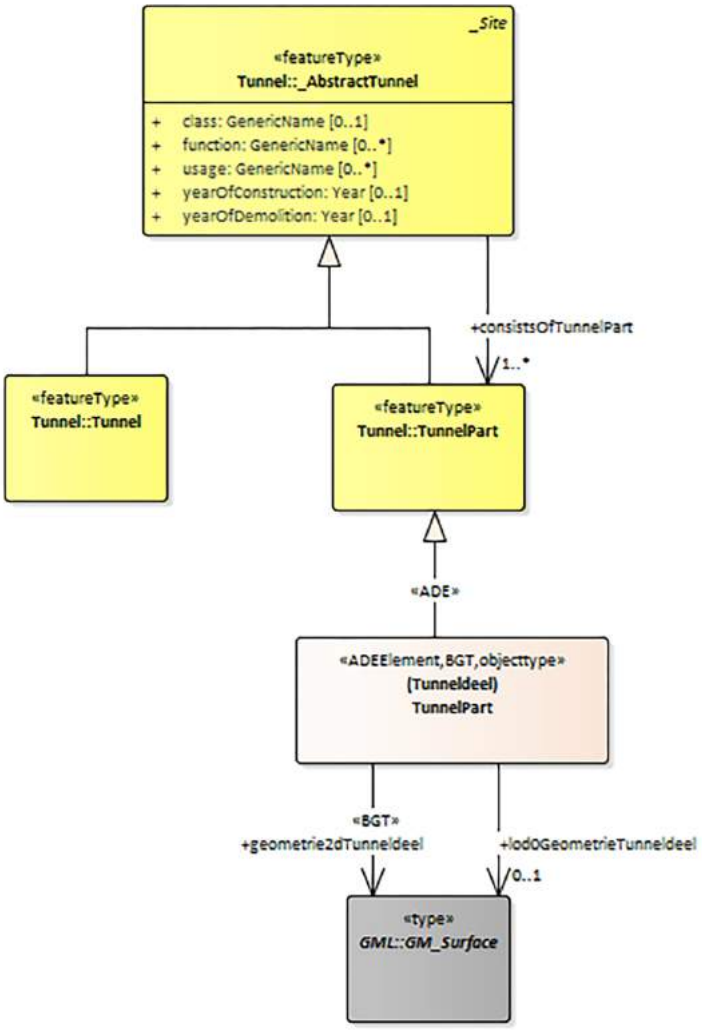


Figure 3.4: TunnelPart AD Element with 2D geometry

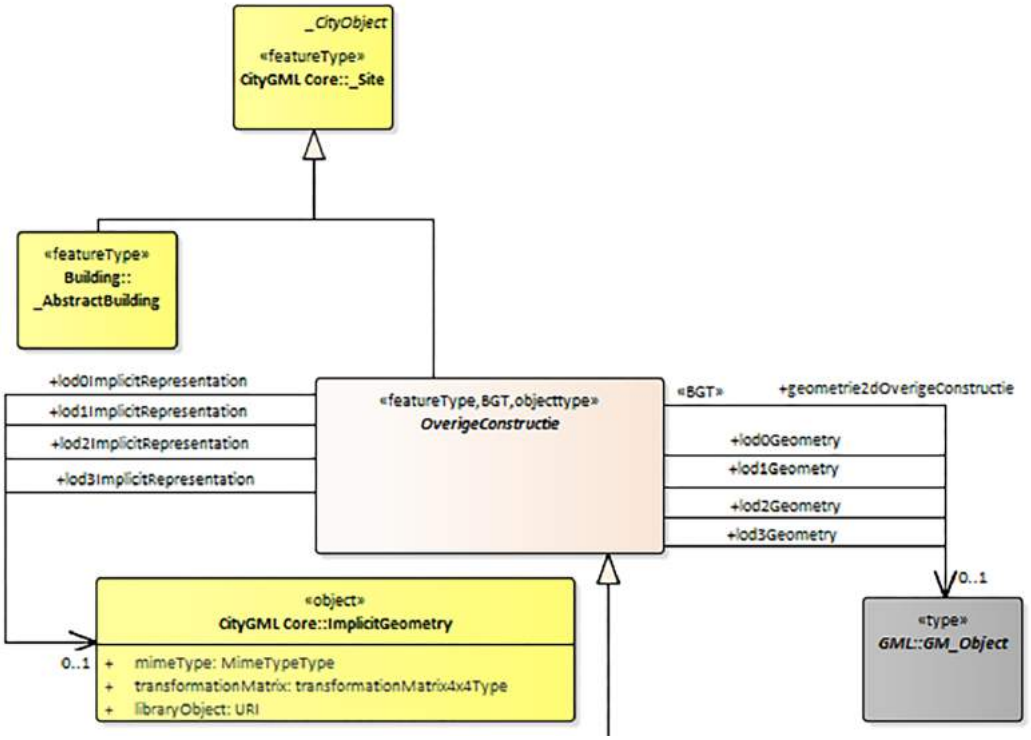


Figure 3.5: OverigeConstructie (OtherConstruction), new class added, derived from _Site

represent man-made constructions other than buildings, bridges and tunnels. Examples are water management constructs such as pumping plants, locks, and weirs but also wharfs, fences, loose-standing walls, high-tension line towers, wind turbines, and so on. It is modelled as a <<featureType>> subclass of the CityGML class _Site (with a Dutch class name) that is not suppressed from the XML Schema. Note that this class does not get the <<ADEElement>> stereotype; this is only used for classes that model extension properties on an existing CityGML class and that need to be suppressed from the XML Schema. The class has its own properties that are modelled similar to CityGML classes, such as implicit geometry on different LODs as well as 2D and 3D geometry up to LOD3.

In 2D IMGeo, the parts of larger built objects like buildings, tunnels and bridges are modelled, but the whole object is not. For the 2D application of IMGeo this is not necessary. When mapping these classes (e.g. TunnelPart) to CityGML, we selected the concept in CityGML that was the closest match semantically, not to the abstract superclass (e.g. Tunnel), even though this could make the model more flexible. When creating 3D IMGeo data from a

2D dataset, a problem occurs because in CityGML the whole object must be modelled as well as the parts, while in the 2D IMGeo data the geometry of the whole objects (buildings, bridges) is not present. This is a gap in IMGeo which remains to be addressed.

Additional properties in an ADE must have a globally unique name. To cope with this restriction in the ADE approach, all additional properties in the IMGeo ADE have a property name containing the class name. In our approach this was done manually in the UML model, but it is feasible to generate globally unique names for properties automatically when the XML Schema is inferred from the UML model, for example by appending the class name.

3.4.2 Code lists in the ADE

CityGML provides code lists to allow predefined values for the CityGML attributes. However, the CityGML-IMGeo ADE makes use of national classification code lists instead of the CityGML code lists, because the national lists are specifically suited to the Dutch context and contain a definition for each concept, approved by the Dutch organizations involved. Other reasons for not using the CityGML code lists are that IMGeo favours Dutch language code lists and that the CityGML standard does not provide definitions for the code list values, which makes it hard to decide which value to use.

The current version of CityGML (2.0), which was not yet published at the time of establishing IMGeo, does allow extension and replacement of code lists. However for the IMGeo ADE there is no need to map the Dutch code lists to the CityGML code lists, as these are non-normative and software does not check on code list values nor process them in specific ways.

Both CityGML and GML do not provide a normative way to structure code lists. Prominent choices are GML dictionary and SKOS (Simple Knowledge Organisation System (Miles and Bechhofer, 2009)).

GML dictionaries can be used to collect sets of definitions or references to definitions (Consortium et al., 2007). In GML, these can be used, for example, to define coordinate reference systems or units of measure. The GML dictionary model implements a simple nested hierarchy of definitions, but is not intended to represent complex interrelating sets of definitions such as taxonomies, thesauri or ontologies. A definition in a GML dictionary has an identifier, and possibly one or more names and inline descriptions or links to descriptions.

GML 3.3 states that “Definition and Dictionary encoding is part of the GML schema as a stop-gap, pending the availability of a suitable general purpose dictionary model” (OGC, 2012). Now that new standards have

matured (in particular RDF, OWL and SKOS) the use of GML dictionaries is deprecated in GML 3.3 for generic definitions and code lists. The best practice, emerged from the semantic web community, is now to use URIs for referring to items in vocabularies. The Resource Description Framework model (RDF; (Klyne and Carroll, 2004)) is in line with this best practice.

Simple Knowledge Organization System (SKOS; (Miles and Bechhofer, 2009)) is based on RDF and contains a common model for vocabularies, thesauri, and taxonomies. It is more lightweight than Web Ontology Language (OWL; (W3C, 2009)), which is a formal knowledge representation language. Items in a SKOS vocabulary are called 'concepts' and can have several labels as well as broader, narrower, and non-hierarchical, associative relations with other concepts. Concepts can be part of concept schemes (vocabularies) and can be grouped in collections.

GML dictionary was considered but not selected, because these are deprecated in GML 3.3, while SKOS adoption is growing in the geo community. SKOS was therefore selected. The code lists are maintained in the UML model and XML structured code lists can be generated from the UML using a ShapeChange customization which allows generation of SKOS-encoded code lists from UML classes with a <<code list>> stereotype. The disadvantage of maintaining the code lists in the UML model, is that the UML model needs to be updated in case the code lists need an update. For IMGeo the code lists are considered as part of the standard and allowed to change only when a new IMGeo version is published.

The SKOS code lists will be published in a national, publicly available registry, which also contains the IMGeo XML Schema. Each code list and code list value is accessible via its own URL. Code list validation can be done using standard XML techniques such as Schematron constraints (ISO, 2006). Further work is needed to assess how the IMGeo code lists are best represented and structured in SKOS. Open questions are whether each code list is encoded as a SKOS concept scheme or a collection, how to construct valid URIs for all code list values and whether the code values are stored in one SKOS file for all code lists, one per code list, or one per code list value.

Figure 3.6 shows the IMGeo code list for classification of WaterBody in UML (attribute class) and a fragment of how this could be encoded in SKOS as a concept scheme, including each code list value's definition. The UML class has a stereotype <<code list>>. The first five values are marked with a stereotype <<BGT>>, which means these are part of the mandatory core of IMGeo. The others are optional further classifications. In the SKOS fragment the first value from the UML code list is shown. Its unique URI is declared in the rdf:about attribute. The code list value itself can be found as a SKOS prefLabel, and its definition as SKOS definition. The inScheme


```
<Concept rdf:about="http://www.geostandaarden.nl/imgeo/2.0#TypeWater_greppel, droge sloot">
  <prefLabel xml:lang="nl">greppel, droge sloot</prefLabel>
  <definition xml:lang="nl">Een ten behoeve van de waterbeheersing gegraven geul die al dan niet met
water bedekt is.</definition>
  <inScheme rdf:resource="http://www.geostandaarden.nl/imgeo/2.0#TypeWater"/>
 </Concept>
```

Figure 3.6: UML code list for Waterbody, attribute typeWater

property declares the concept as a member of the concept scheme TypeWater for WaterBody.class.

3.4.3 Geometry and topology in the IMGeo ADE

For the use of geometry types and the LOD concept in the IMGeo ADE, we formulated the following guidelines:

- The higher LODs will always be derived from IMGeo data. Therefore all LODs have the same x, y accuracy as the 2D IMGeo data. This assures consistency between all LODs starting from the 2D geometry. Which LOD is used depends entirely on the required level of detail in the third dimension rather than the positional accuracy as mentioned in CityGML specifications.

- Extra attributes defining geometry types are added to the subclasses to support the full range of geometries: 2D geometry for all subclasses (not modelled in CityGML) and the LOD0 geometry if not present in the

equivalent CityGML super class. With the LOD0 representations the footprint and the exact location of the footprint in the terrain is known. Consequently 3D representations of those objects (LOD1 and higher) can easily be placed on the terrain assuring that they do not float in the air or disappear in the ground. In addition this approach supports 2D, 2.5D and 3D representations of objects in an integrated manner. For some classes other geometry types were used than modelled in CityGML. In our approach Water and Road objects are represented by surface geometry at LOD0 and Railway is represented by the curve geometry at LOD0, while CityGML represents those classes with the GeometricComplex at LOD to accommodate networks. The changes were done to be compliant with geometry types in the previous version of IMGeo.

- Special attention was given to the link between buildings and terrain. Two approaches can be followed:
  - The buildings always have horizontal foundation in the terrain and can sink, but the above surface and below surface geometries are modelled separately. The 2D representations of the buildings represent the building geometries at surface level, which can straightforwardly be extended into 2.5D. This modelling makes the Terrain Intersection Curves (TIC) as solution proposed in CityGML redundant. This approach was used in CityGML-IMGeo.
  - The buildings always have horizontal foundation and may have underground parts (Zlatanova et al., 1996). This means that the building can sink under the surface and the TIC for buildings has to be maintained to ensure the consistency with the terrain surface.
- For applying the CityGML Digital Terrain Model three different approaches can be followed:
  - Terrain is represented by a surface by regular or irregular grid. The topographic objects are not integrated in the terrain (the current concept of CityGML)
  - Terrain is represented by constrained TIN in which the boundaries of objects form the constraints. This approach ensures the consistency between objects and terrain surface. The terrain is still present in the model.

– Terrain is not represented as a separate object, i.e. all objects on surface level (such as roads, land use, etc.) incorporate the terrain curvature in their representation (as in Emgard and Zlatanova (2007)). This approach is used in CityGML-IMGeo and supported in CityGML.

For topology the following principles were copied from IMGeo 1.0. Although not expressed in the formal model, the IMGeo standard contains a general rule that the 2D objects at surface level must form a topological structure of the surface of the Netherlands without gaps or overlaps. Since the 2D IMGeo is a topologically correct model, the 2.5D surface (LOD0) should support this as well: all objects at surface level have a representation at LOD0 (= 2.5D surface) which together form a 2.5D topological structure for those objects located at surface level. Objects that are located above and below the surface can also be placed in the third dimensional space with their LOD0 2.5D representation. An important requirement here is the connection to the 2.5D DTM that represents the surface. This may require adding new 2D boundaries for adding more variance in 3D or extra 2.5D surfaces to the structure at the surface to avoid gaps (see Stoter et al. (2011) for an example).

3.4.4 Generating XML Schema from the UML ADE

The Java tool ShapeChange is used to generate an XML Schema (GML application schema) from the ADE defined in UML. As mentioned before, ShapeChange implements the UML to GML encoding rules described in ISO 19136, ISO 10118, and ISO 19109. ShapeChange was only used to generate the XML Schema for the IMGeo ADE, not to generate the XML Schemata for CityGML. These schemata are already publicly available and the generated ADE schema only needs to correctly import the CityGML Schemata, which ShapeChange does based on dependencies between UML packages. The CityGML 2.0 XML Schema as published by OGC is not generated from the UML model and this would not be possible without changing the current UML model, because it does not adhere to several aspects of the relevant encoding rules. It is, however, beyond the scope of the paper to discuss this.

ShapeChange was modified to add a custom encoding rule for classes with the <<ADEElement>> stereotype. These classes are suppressed from the GML Application schema, while their properties are added to the ADE namespace as substitutes for the CityGML "_GenericApplicationPropertyOf-<Featuretypename>" hooks as described in CityGML 2.0, section 10.13.1.

The generated schema fragment below shows that the IMGeo ADE extension class of TunnelPart (see Figure 3.4) is suppressed from the XML Schema, while the extra properties (only one example shown below) are implemented according to the CityGML extension hook mechanism. The IMGeo class OverigeConstructie (see Figure 3.5) on the other hand, is a newly added class, and not suppressed from the XML Schema.

```
<element name="geometrie2dTunneldeel"
  substitutionGroup="tun:\_GenericApplicationPropertyOfTunnelPart"
  type="gml:SurfacePropertyType"/>
  ...

<element abstract="true" name="OverigeConstructie"
  substitutionGroup="cit:\_Site"
  type="imgeo:OverigeConstructieType"/>

<complexType abstract="true" name="OverigeConstructieType">
  <complexContent>
    <extension base="cit:AbstractSiteType">
      <sequence>
        <element maxOccurs="unbounded"
          name="plaatsbepalingspuntOverigeConstructie"
          type="imgeo:PlaatsbepalingspuntPropertyType"/>
          ...
      </sequence>
    </extension>
  </complexContent>
</complexType>
```

Fragment of the generated GML application schema

3.4.5 Creation of IMGeo 2.0 Data

The organizations responsible for IMGeo data are now in the process of creating IMGeo 2,0 compliant data based on existing large scale data. IMGeo 2.0 compliant test data has been generated to show how the model works when applied to data, see Figure 3.7. The viewer used is FZK Viewer.

3.5 Model-driven Framework for developing CityGML ADE

The approach to model an ADE of CityGML in UML as presented in this paper contains many points that can be generalized and reused by future do-

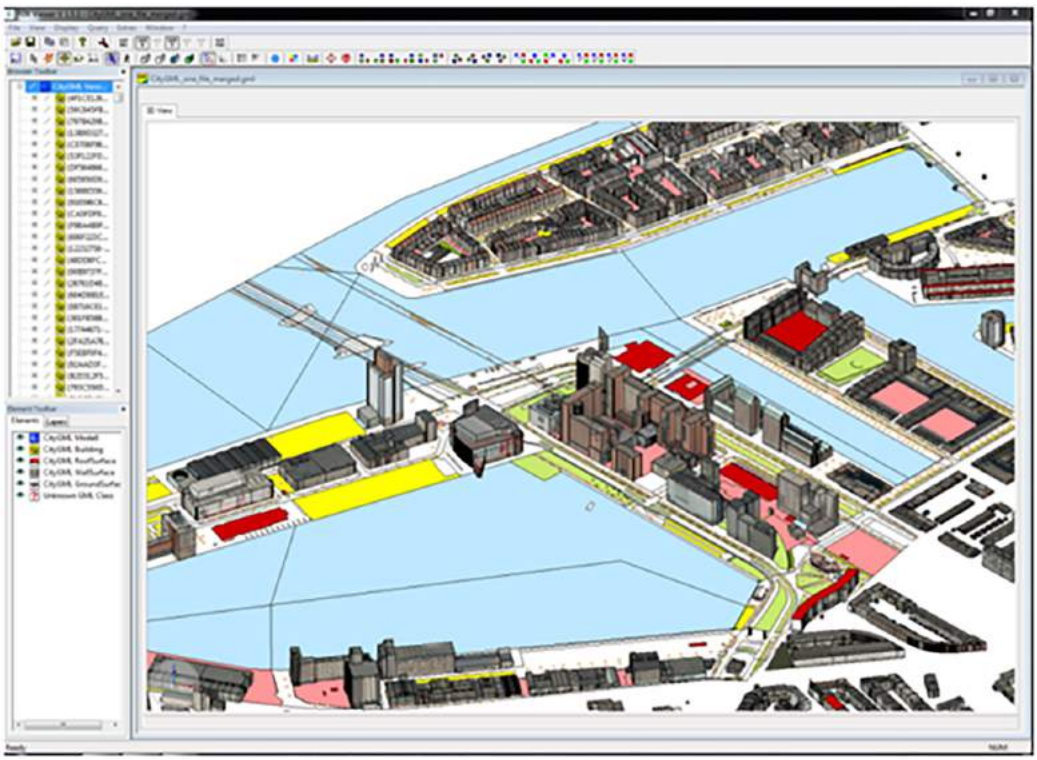


Figure 3.7: Visualisation of CityGML-IMGeo encoded data: CityGML LOD2

main extensions of CityGML. Therefore, in this section we propose a generic model-driven framework for developing a CityGML ADE in UML based on our experiences. It should be noted that this approach is not only CityGML specific. For example the German standard XPlanGML also supports the ADE mechanism. Therefore XPlanGML ADEs can be developed in the same way (Benner et al., 2012).

The steps to be followed when developing an ADE can be summarised as follows:

1. *Select a formal modelling language, e.g. UML to represent all the classes.* In the Netherlands UML is widely used for modelling in the geo-domain. The Basic Schema for Geo-Information (NEN, 2011) is modelled in UML since 2005, and many other information models in the geo domain have followed. Reasons for selecting UML include:

   (a) its visual modelling approach makes the model suitable for communication with stakeholders (users, software developers)

   (b) it is possible to generate XML schema and documentation directly and automatically from the model, i.e. it supports the MDA approach

   (c) it is a formal language which can unambiguously express the structure and rules of an information model

   (d) it is an international standard

   (e) it is used extensively in ISO 19xxx standards which are relevant to the GIS domain

   In this step, all object types are created as UML classes in UML class diagrams with the appropriate properties and relations. A UML model of CityGML is also needed. During our work on the CityGML ADE IMGeo, the CityGML classes also had to be created, because a UML model in which these classes were already fully defined, and which could be imported in the UML tool of choice, was not available. The SIG3D group is currently completing the CityGML part of our UML model and plans to make it publicly available so that future ADEs can build on this work.

2. *Define correspondences between semantic classes of the application and CityGML generic classes.* In this step, a first study is done on the semantic correspondence between the classes from the application domain and CityGML. The mapping is not yet formally defined, but only intended to list all the correspondences, mainly on the class level. In

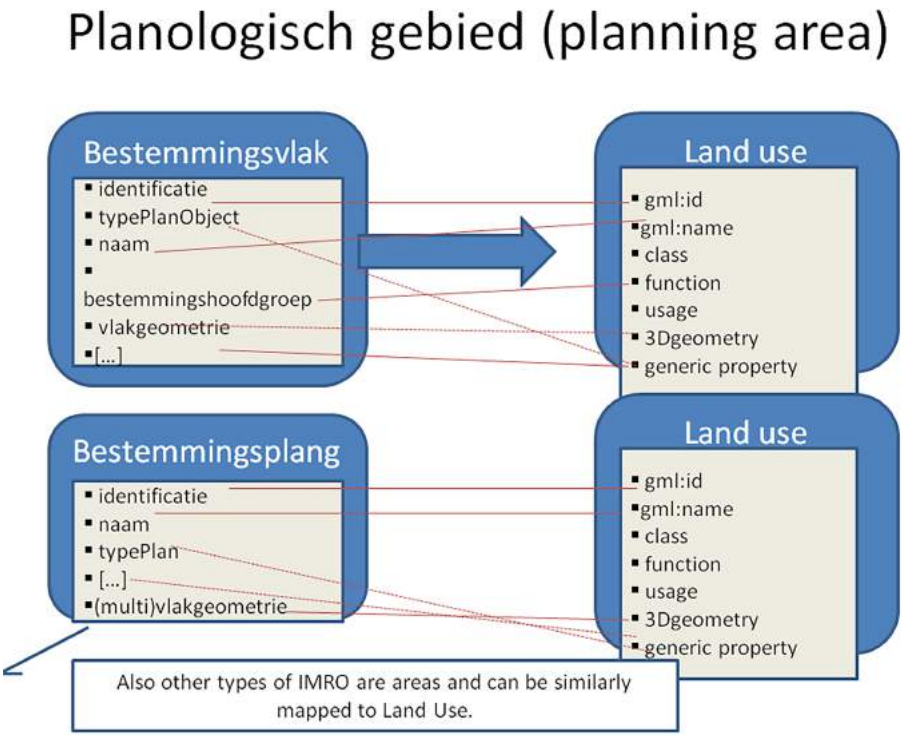


Figure 3.8: Information model for spatial planning

our approach we have created only informal drawings and text in this step. For all classes in the application model, an equivalent class in CityGML was searched based on its name, description, geometry representations and properties. More details on schema mapping can be found in Lehto (2007). Based on this exercise the relevant CityGML classes were identified. Figure 8 below shows the mapping of two other Dutch information models to CityGML.

3. *Decide which subclasses should be extended.* In the current approach we use specialisation relations to define this correspondence. We distinguish two types of correspondence. Either the class is semantically the same as the corresponding CityGML class, and only adds properties, or the class is semantically a subclass of the corresponding CityGML class. In both cases specialization is used, but in the first case the specialization relation is marked with a stereotype <<ADE>>, the application specific class is marked with a stereotype <<ADEElement>>, and its class name is the same as the corresponding CityGML class name. The UML class in this case is only a placeholder for addi-

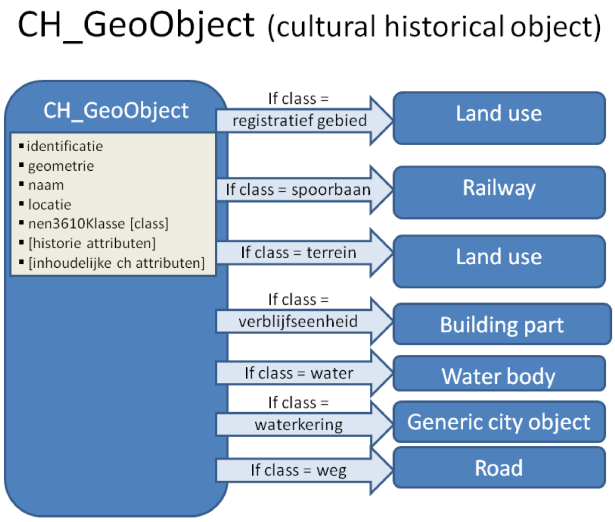


Figure 3.9: Information model for cultural and historic objects

tional properties that are attached to the 'hooks' CityGML provides for extension of its classes. In the second case, the specialization relation receives no stereotype, the application specific class receives the stereotype <<FeatureType>> (ISO19109) and the class name is different from the corresponding CityGML class name. When classes in the source domain model do not correspond exactly with CityGML classes, e.g. some instances of the class correspond with a CityGML class but other instances do not, then the classes in the domain model should be harmonized with CityGML.

4. *Define application code lists if necessary.* Code lists provide additional semantic richness in CityGML. CityGML has non-normative code lists for most properties that add semantic detail. As of CityGML 2.0 it is possible to replace these code lists with others. Reasons for using own code lists include the need to provide non-English code lists and a need for clear definitions of each code list value (not provided in CityGML). To make use of national classification code lists (which will be so specific in most cases that those national lists are preferred over the CityGML code lists) mapping tables between the code lists/ code list values should be provided.

5. Optional: define a geometry representation for each class for each applicable level of detail. This step is optional because often only non-geometric additional properties are defined in ADE-classes, the geomet-

ric representations provided by the CityGML base class being sufficient. In the case of IMGeo, additional geometry representations are required. It was already defined which 2D geometry types must be used for each class. Which levels of detail apply for which classes, and which geometry representations they have in those LODs, should be decided based on user requirements, rather than the positional accuracy as mentioned in CityGML specifications. That is, for each class it should be decided whether it should be represented with full solids or if a 2.5D representation is sufficient. Extra attributes defining geometry types could be needed to support the full range of geometries for every class: 2D geometry (not modelled in CityGML) and the LOD0 geometry if not present in CityGML (as footprints). Special attention should be given to the link between buildings and terrain. Finally, terrain object as given in CityGML should be carefully considered (Van den Brink et al., 2012).

6. *Make a decision on which LOD should be topologically correct.* The most notable rule in IMGeo is that the complete set of polygon-objects at surface level (height level 0) in the mandatory core must together form a complete coverage of the Netherlands without gaps or overlap in 2D. If the 2D model contains a topological structure, the 2.5D surface (LOD0) should support this as well: all objects at surface level have a representation at LOD0 (= 2.5D surface). Objects that are located above and below the surface can also be placed in the third dimensional space with their LOD0 2.5D representation. An important requirement here is the connection to the 2.5D DTM that represents the surface. This may require adding new 2D boundaries.

3.6 Conclusion and further research

This paper presents a model driven approach to generate a CityGML ADE starting from UML schema. Several alternatives were proposed, investigated and discussed within a group of international experts (SIG3D members and others). The most beneficial approach to model a CityGML ADE in UML was applied for the CityGML ADE IMGeo (i.e. the Dutch national 3D standard on large-scale geo-information). The main principle of the selected approach is that all classes of IMGeo are modelled as subclasses of CityGML classes. If these subclasses only add properties to existing CityGML classes, they get the same class name as the CityGML class they are extending. The stereotype <<ADEElement>> marks these classes as subtypes that only add properties to the CityGML class, and accordingly no XML component

for these classes will be created in the XML Schema. Following the design approach for CityGML ADE IMGeo, we have established a model-driven framework for developing CityGML ADE.

In the development of the CityGML ADE IMGeo 2.0 a number of topics are identified that require further research. Firstly, more research is needed to understand how this model works in practice including the consequences of this new modelling method for IMGeo when used for both 2D and 3D datasets, e.g. how to preserve the links between the different LODs and how to upgrade 2D LOD to higher LODs. Secondly, knowledge is required on the ability to use the CityGML ADE IMGeo UML model to generate working database schemes. Thirdly, more research is required on how to handle every type of correspondence between CityGML and domain classes, for example how to deal with overlapping classes that are neither super- or subclass of a CityGML class. Finally, more research is needed concerning the creation and management of CityGML-IMGeo data. Which methods can be used to generate CityGML-IMGeo data? How should this data be validated and maintained? How can 2.5D topology be created and maintained?

These open issues are currently being studied in a follow-up project of the 3D pilot NL. The first phase finished in June, 2011 and has been reported in Stoter et al. (2011). Since October 2011 over 100 organisations (over 300 persons) are contributing to the six activities of the 3D Pilot NL. The activities related to learn more about the UML modelling approach for ADEs are the generation of 3D IMGeo example data (several levels of detail and several classes) and the design and implementation of a 3D validator that tests whether both the semantics and the geometry of the data are compliant with the standard.

In conclusion, this is the first study on extending the UML diagrams of CityGML for specific domains. Since the OGC CityGML specification does not provide rules or guidance on correctly modelling an ADE in UML, and this paper reflects the combined efforts of international experts, we believe that this paper may serve as best practice for future ADEs to be modelled in UML.

Acknowledgments The authors express their sincere gratitude to Thomas Kolbe (TU Berlin); Carsten Roensdorf (Ordnance Survey and chair of OGC CityGML working group), the SIG3D modelling subgroup, and Clemens Portele (interactive instruments GmbH), who contributed to discussions.

This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO) and partly funded by the Ministry of Economic Affairs, Agriculture and Innovation (Project codes: 11300).

Bibliography

J. Benner, K.-H. Häfele, and A. Geiger. Integration raumbezogener Daten in einer CityGML Application Domain Extension (ADE) zur Unterstützung des digitalen Bauantragsverfahrens. In Marc-Oliver Löwner, Florian Hillen, and Ralf Wohlfahrt, editors, *Geoinformatik 2012 'Mobilität und Umwelt'*, page 163 – 170, 2012.

Volkan Çağdaş. An Application Domain Extension to CityGML for immovable property taxation: A Turkish case study. *International Journal of Applied Earth Observation and Geoinformation*, 21:545–555, 2013.

Open Geospatial Consortium et al. Geography Markup Language (GML) Encoding Standard. Version 3.2.1, doc nr OGC 07-036 [online]. Available online: http://portal.opengeospatial.org/files/?artifact%5Fid=20509, 2007.

Ton de Vries and Sisi Zlatanova. 3D intelligent cities. *GEO Informatics*, 14 (3):6–8, 2011.

KL Emgard and S Zlatanova. Design of an integrated 3D information model. *Urban and regional data management: UDMS annual*, pages 143–156, 2007.

Dragan Gaševic, Dragan Djuric, and Vladan Devedžic. *Model driven architecture and ontology development*. Springer Science & Business Media, 2006.

Geonovum. Informatiemodel Geografie, 10 2007.

Gerhard Gröger, Thomas H Kolbe, Angela Czerwinski, and Claus Nagel. OpenGIS city geography markup language (CityGML) encoding standard, version 1.0.0. Available online: http://portal.opengeospatial.org/files/?artifact%5Fid=28802 (accessed 11 February 2014), 2008.

Gerhard Gröger, Thomas H Kolbe, Claus Nagel, and Karl-Heinz Häfele. Ogc® city geography markup language (CityGML) encoding standard, Version 2.0, 2012.

Annet Groneman and Sisi Zlatanova. TOPOSCOPY: a modelling tool for CITYGML. In Onsrud and van de Velde, editors, *Proceedings of GSDI Association*. Citeseer, 2009.

Joao Hespanha, Jan van Bennekom-Minnema, Peter van Oosterom, and Christiaan Lemmen. The model driven architecture approach applied to the land administration domain model version 1.1-with focus on constraints specified in the object constraint language. In *fig working week 2008, integrating generations*, page 19, 07 2008.

ISO. Information technology–document schema definition language (DSDL)–Part 3: Rule-based validation – Schematron, 2006.

Graham Klyne and Jeremy J Carroll. Resource description framework (RDF): Concepts and abstract syntax, version 2004-02-10. Available online: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, 2004.

Tatjana Kutzner and Thomas H Kolbe. Extending semantic 3D city models by supply and disposal networks for analysing the urban supply situation. In T. P. Kersten, editor, *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e.V. Wissenschaftlich-Technische Jahrestagung der DGPF*, pages 382–394, 2016.

A Lapierre and P Cote. Using Open Web Services for urban data management: A testbed resulting from an OGC initiative for offering standard CAD/GIS/BIM services. In *Urban and Regional Data Management. Annual Symposium of the Urban Data Management Society*, pages 381–393. Taylor & Francis, 2007.

Lassi Lehto. Schema translations in a web service based SDI. In *Proceedings of the 10th AGILE International Conference on Geographic Information Science*, page 15. Citeseer, 2007.

Alistair Miles and Sean Bechhofer. SKOS simple knowledge organization system reference. Available online: https://www.w3.org/TR/2009/REC-skos-reference-20090818/, 2009.

NEN. NEN3610 Basismodel Geo-informatie - Termen, definities, relaties en algemene regels voor de uitwisseling van informatie over aan de aarde gerelateerd ruimtelijke objecten. Available online: https://www.nen.nl/NEN-Shop/Norm/NEN-36102011-nl.htm, 2011.

Romain Nouvel, Robert Kaden, Jean-Marie Bahu, Jerome Kaempf, Piergiorgio Cipriano, Moritz Lauster, Joachim Benner, Esteban Munoz, Olivier Tournaire, and Egbert Casper. Genesis of the citygml energy ADE. In *Proceedings of International Conference CISBAT 2015 Future Buildings and Districts Sustainability from Nano to Urban Scale*, volume EPFL-CONF-213436, pages 931–936, 2015.

OGC. OGC® Geography Markup Language (GML) - Extended schemas and encoding rules, version 3.3.0. Available online: http://www.opengeospatial.org/standards/gml, 2012.

OMG. Model Driven Architecture, Guide Version 1.0.1. Available online: http://www.omg.org/news/meetings/workshops/UML%5F2003%5FManual/00-2%5FMDA%5FGuide%5Fv1.0.1.pdf (last accessed June 2012), 2003.

Clemens Portele. Mapping UML to GML Application Schemas; ShapeChange - Architecture and Description, version 1.0rc. Available online: www.interactive-instruments.de/fileadmin/gdi/docs/ugas/ShapeChange-1.0.pdf, 2008.

C Rönsdorff, D Wilson, and JE Stoter. Integration of land administration domain model with CityGML for 3D cadastre. In *Proceedings 4th International Workshop on 3D Cadastres, 9-11 November 2014, Dubai, United Arab Emirates*. International Federation of Surveyors (FIG), 2014.

M Rumor and E Roccatello. Design and development of a visualization tool for 3D geospatial data in CityGML format. *Urban and regional data management: UDMS annual*, pages 31–37, 2009.

Alexandra Stadler and Thomas H Kolbe. Spatio-semantic coherence in the integration of 3D city models. In *Proceedings of the 5th International ISPRS Symposium on Spatial Data Quality ISSDQ 2007 in Enschede, The Netherlands, 13-15 June 2007*, 2007.

J Stoter, M Reuvers, G Vosselman, J Goos, L. van Berlo, S. Zlatanova, E. Verbree, and R. Kloosters. Towards a 3D geo-information standard in the Netherlands. In Kolbe, Köning, and Nagel, editors, *International Archives of the Photogrametry, Remote sensing and Spatial Information Sciences*, volume XXXVIII-4/W15, pages 63–67, 2010.

Jantien Stoter, Henk De Kluijver, and Vinaykumar Kurakula. 3D noise mapping in urban areas. *International Journal of Geographical Information Science*, 22(8):907–924, 2008.

Jantien Stoter, George Vosselman, Joris Goos, Sisi Zlatanova, Edward Verbree, Rick Klooster, and Marcel Reuvers. Towards a national 3D Spatial Data Infrastructure: case of the Netherlands. *Journal of photogrammetry, remote sensing and geoinformation processing*, 2011(5), 2011.

Jantien Stoter, Jacob Beetz, Hugo Ledoux, Marcel Reuvers, Rick Klooster, Paul Janssen, Friso Penninga, Sisi Zlatanova, and Linda van den Brink. Implementation of a national 3D standard: Case of The Netherlands. In *Progress and New Trends in 3D Geoinformation Sciences, Lecture Notes in Geoinformation and Cartography*, pages 277–298. Springer, 2013a.

Jantien Stoter, Hendrik Ploeger, and Peter van Oosterom. 3D cadastre in the Netherlands: Developments and international applicability. *Computers, Environment and Urban Systems*, 40:56–67, 2013b.

Jantien E Stoter and PJM Van Oosterom. Technological aspects of a full 3D cadastral registration. *International Journal of Geographical Information Science*, 19(6):669–696, 2005.

Linda Van den Brink, JE Stoter, and Sisi Zlatanova. Modeling an Application Domain Extension of CityGML in UML. In J. Pouliot, S. Daniel, F. Hubert, and A. Zamyadi, editors, *7th International Conference on 3D Geoinformation, Quebec, Canada, 16-17 May 2012; International Archives od the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVIII-4/C26*, pages 11–14. International Society of Photogrammetry and Remote Sensing (ISPRS), 2012.

E Verbree, J Stoter, S Zlatanova, G De Haan, M Reuvers, G Vosselman, J Goos, L Van Berlo, and R Klooster. A 3D model for geo-information in the Netherlands. In *Proceedings of A special joint symposium of ISPRS Technical Commission IV and AutoCarto in conjunction with ASPRS/CaGIS 2010 Fall Specialty Conference*. International Society for Photogrammetry and Remote Sensing, 2010.

W3C. OWL 2 web ontology language overview. Available online: http://www.w3.org/TR/2009/REC-owl2-overview-20091027/, 2009.

CityGML Wiki. Application Domain Extensions. [maintained online], Available online: www.citygmlwiki.org/index.php/CityGML-ADEs (accessed July 2012), 2012.

Sisi Zlatanova, Jantien Stoter, and Umit Isikdag. Standards for exchange and storage of 3D information: Challenges and opportunities for emergency response. In T. Bandrova, M. Konecny, and G. Zhelezov, editors, *Proceedings of the 4th International Conference on Cartography & GIS*, volume 2, pages 17–28. International Cartographic Association, 06 2012.

Siyka Zlatanova, Michael Gruber, and M Kofler. Merging DTM and CAD data for 3D Modeling purposes for Urban Areas. In *Proceedings of ISPRS*, volume XXXI, pages 311–15, 1996.

3.7 Appendix: Overview of main classes in CityGML ADE for IMGeo

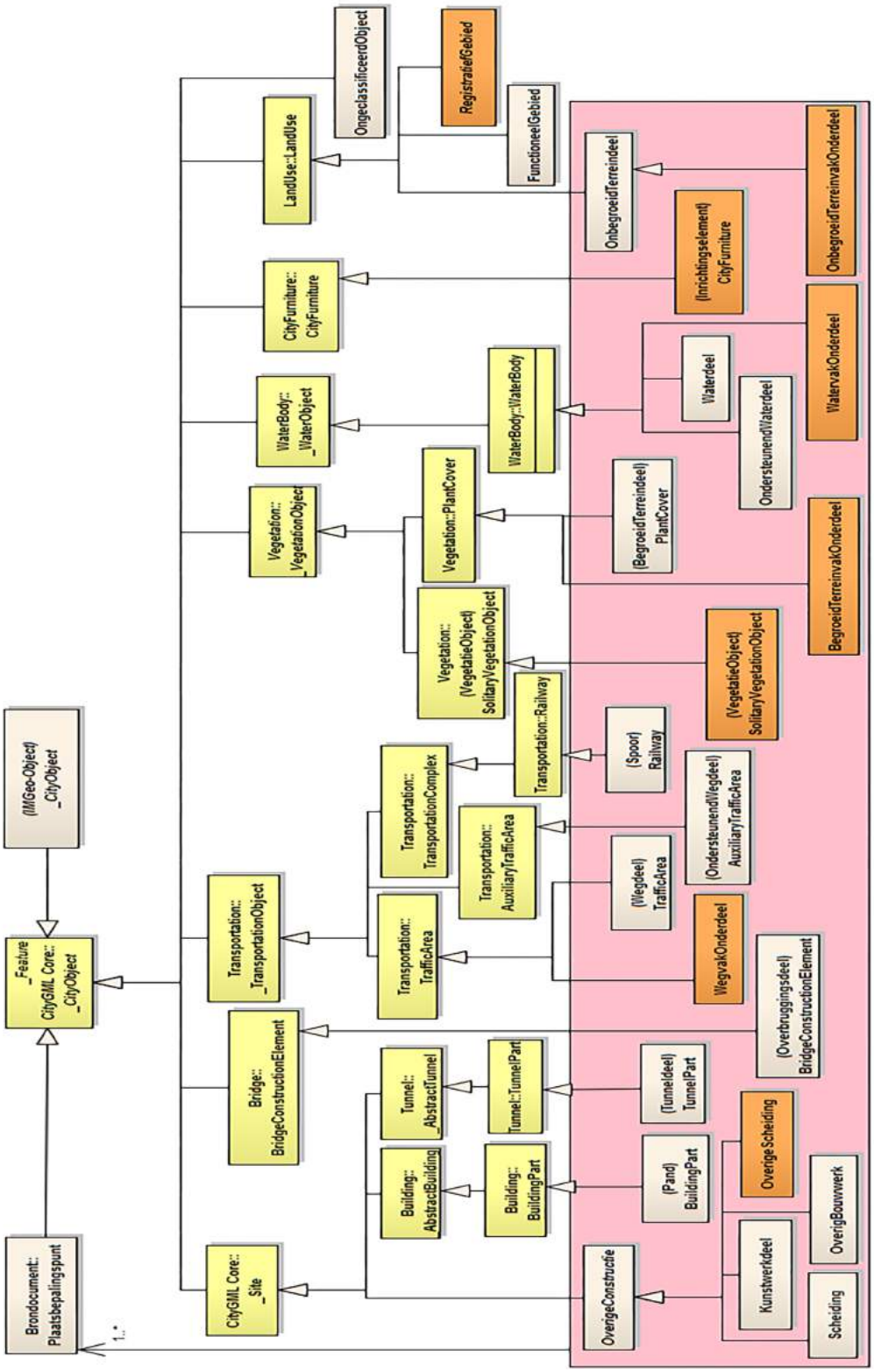


Figure 3.10: Main classes of the IMGeo ADE

Semantic Harmonisation

Chapter 4

Towards a high level of semantic harmonisation in the geospatial domain

Authors: L. van den Brink, P. Janssen, W. Quak & J. Stoter (*Computers, Environment and Urban Systems*, 62:233-242, 2017).

This paper has been published in a peer-reviewed scientific journal in 2016 and is published unchanged in this chapter. In this paper, the problem of semantic non-interoperability of Dutch information models and INSPIRE themes is analysed as a case to describe the problem of lack of semantic harmonisation between existing information models in the geospatial domain.

The main question, how semantic interoperability between different kinds of geospatial datasets can best be achieved, is addressed in this paper by studying how information models from different domains can be integrated and the similarities and differences between them can be analysed and solved. Because of the volume of concepts from information models that needed to be analysed, this is done in a semi-automated way, while human domain experts are involved in solving semantic interoperability problems that were found.

The semantic overlap and differences of the information models are studied. First by taking one information model, IMGeo (Information model for large scale topography), as the basis and comparing it with the others. IMGeo was selected for this because its scope is large scale topography. This makes it a good candidate to identify and solve semantic overlap with other, more specific domain information models that contain similar concepts. In the second part of this study I developed and implemented a methodology in which all the relevant geo-information models can be maintained (and pub-

lished) together. This provides the possibility to identify similar concepts that may be slightly differently defined in different data models in order to improve the semantic interoperability in a next step.

Contributions: 1) several proposed changes to both IMGeo and other Dutch information models, some of which have already been realised, while others will be addressed in future versions of the models; and 2) tool for the discovery and resolution of semi-automated semantic interoperability problems.

text of published paper starts after this line

Abstract Spatial Data Infrastructures (SDIs) aim at making spatial (geographical) data and thus content available for the benefit of the economy and of the society. Agreement and sharing of vocabularies within the SDI are vital for interoperability. But there is a limitation: many vocabularies have been defined within domains while other domains have not been taken into account. Therefore, little harmonisation has been achieved and data sharing between domains within the SDI is problematic. This paper presents a methodology and tools for non-automatic, community driven ontology matching that we developed to harmonise the definition of concepts in domain models that are already being defined and used in operational use cases. Besides the methodology and tools that we developed, we describe our experiences and lessons learned as well as future work.

4.1 Introduction

Spatial Data Infrastructures (SDIs) aim at making spatial (geographical) data and thus content available for the benefit of the economy and of the society. The traditional approach of SDIs is characterised by service-based dissemination of GML data (Geography Markup Language) (Consortium et al., 2007), structured according to agreed information models. In the INSPIRE (INSPIRE, 2007) programme, for example, a lot of effort has been put into establishing information models (i.e. data specifications) to define the vocabulary of a specific domain in a standardised way and to structure spatial data accordingly.

The strong point of this approach is that the purpose of standardisation and harmonisation, being interoperability, can be addressed through agreement and sharing of vocabulary. Once agreed the requirements and rules for communication are set and can be implemented in a verifiable way. But there is a limitation: the vocabularies are defined within domains and thus interoperability is only assured by shared and foreseen concepts. However,

between domains little harmonisation has been realised, and for unforeseen reuse of both concepts and relations across domains the structure of existing information models may be too rigid.

A common problem of the lack of harmonisation between domains is the existence of similar concepts in different domain models. It is often not clear if these concepts are in fact the same in a semantic sense, or subtly different—either unintentionally, or because of different domain specific needs. Linked data and semantic web technology are often expected to solve this problem because they enable data from one domain to be integrated and harmonised with other data and data models. However, re-using or integrating data with similar, but different semantics is often problematic. Consequently, geographical data is often created instead of reusing existing data (a costly process in the geospatial domain). The underlying problem is often one of semantic harmonisation: either the semantics are not clear across domains, or there are subtle semantic differences that limit reuse. Harmonising similar concepts in different domains and related information models is therefore still needed to enable the reuse of data over domain boundaries and to prepare for a linked data approach at a later stage.

This paper presents the methodology that we developed to harmonise the definition of concepts in domain models that are already being used in operational use cases. The starting point of our research is the SDI approach in The Netherlands in which object-oriented information models have been developed in different domains. This resulted in technical harmonisation, but not in semantic harmonisation. The most semantic harmonisation that has been achieved is on an ad hoc basis and depending on the domain model being currently updated or developed; in addition, the outreach of each domain model, for example in the form of public consultations is mainly done within domains. The lack of semantic harmonisation between domains and resulting inefficient data distribution became only apparent after the data distribution within domains was working properly. The research presented in this paper aims at resolving this harmonisation gap.

Since solving harmonisation issues between existing (i.e. currently operating), independent domain models requires an ex post harmonisation repair process, it needs a different approach than harmonisation via establishing new, common data models like INSPIRE. Every domain model is created with a domain-specific world view in mind; classes in the domain models are specialisations of a very generic global ontology that has been standardised in the Netherlands, but their similarity with classes from other domain models has never been considered.

Our study to improve harmonisation between different domain models contained two parts. The first part (I) aimed at obtaining in-depth insight

into semantic differences and overlap in existing domain models and compared semantic concepts defined in existing domain models of a national SDI. The second part (II) aimed at establishing an environment to capture and publish all concept definitions valid in the SDI to make reuse of concept definitions possible. This enables concepts to operate as individual information objects instead of being only related to individual domains or information models.

In this paper, we describe the methodology and tools for non-automatic, community driven ontology matching that we developed in our research. In addition, we describe our experiences and lessons learned as well as future work.

Section 4.2 describes the background of the SDI approach in the Netherlands and the resulting harmonisation problems between domain models. Section 4.3 presents related work on harmonisation and ontology matching. Section 4.4 presents the overall methodology and tools that we developed to obtain a higher level of harmonisation between domain models. Section 4.5 presents the first part of the research (part I) in which we developed a methodology to provide in-depth insight into semantic overlap and discrepancies between information models of the current SDI. Section 4.6 presents part II of the research in which a further step was taken to resolve semantic discrepancies between information models where possible. Section 4.7 closes with conclusions and future work.

4.2 Background: Model driven approach of the Dutch SDI

As explained in van den Brink et al. (2013), formal representation of conceptual models for geo-information defined with the Unified Modelling Language (UML) is seen as an important prerequisite of the Dutch Spatial Data Infrastructure (SDI). UML is worldwide one of the most used modelling languages by standardisation bodies dealing with geo-information. With UML class diagrams, geo-information objects can be formally described with their properties, relationships and semantics. A good understanding of the meaning of objects is required when different organisations reuse each other's information. Although not as elaborate as some ontology engineering languages focusing on semantics (such as Ontology Web Language (OWL) W3C (2009), UML is not widely different from these languages and provides sufficient means to record the meaning of objects (Kiko and Atkinson, 2008).

In the Netherlands' SDI, a Model Driven Approach (MDA) such as de-

scribed in Gaševic et al. (2006) is applied for modelling concepts and their implementation in different domains. A key point of this approach is that either the conceptual information models are independent of their technical implementation(s) or they are platform-independent (Hespanha et al., 2008; OMG, 2003). As the UML models are conforming to an agreed meta model, i.e. the ISO 19109 citepiso19109 general feature model, the technical implementations for data storage or data exchange can automatically be created from the UML schemas using standardised mapping and encoding rules. For data exchange based on these models, Geography Markup Language (GML) (Consortium et al., 2007) is used. The technical implementations (in this case GML application schemas) are not designed and maintained separately, but are automatically derived from the UML models using the standardised mapping rules described in GML 3.2.1 appendix E. This provides a one to one relation between the conceptual UML environment and the GML implementation specifications.

In the Netherlands, the Base Model Geo-Information (NEN, 2011) forms a common base for domain specific information models. This national standard describes geographic concepts and establishes a standard modelling method based on the ISO 191XX series of standards (specifically: ISO 19103 (ISO, 2015a), ISO 19107 (ISO, 2003), ISO 19109 (ISO, 2015b), ISO 19110 (ISO, 2005), ISO 19131 (ISO, 2007)). It contains a generic semantic UML model with definitions of the most common, shared concepts in the geo-domain such as Road, Water, etc. Therefore it can be considered as a global ontology, although a small one. In 2011 the standard was revised and parts of the INSPIRE Generic Conceptual Model (INSPIRE, 2014) were included. Many domain specific information models have been developed on top of NEN3610. These domain models define specialisations of the base classes defined in the NEN 3610 global ontology with more specific classes and properties. The resulting semantic geo-standards in the Netherlands can be viewed as a pyramid of information models (see Figure 4.1).

The abbreviations in the pyramid of Figure 4.1 are mnemonic names for Dutch standards; from left to right these are (IM=Information Model):

- IMRO = ruimtelijke ordening (spatial planning)
- IMWA= water
- IMLG=landelijk gebied (rural area)
- IMNAB= Natuur Beheer (Nature Management):
- IMOOV= Openbare Orde en Veiligheid (Public Order and Safety)

- IMKL= Kabels en Leidingen (Cables and Pipelines)
- IMKAD= Kadastrale percelen (cadastral parcels)
- IMKICH=Kennisinfrastructuur Cultuurhistorie (cultural heritage)
- IMWE= Welzijn (welfare)
- IMGeo= geography
- IM01010= soil
- IMBRO= Basisregistratie Ondergrond (subsoil)
- IMTOP= Topography
- IMMetingen= Measurements

Via the extension of NEN 3610, vertical harmonisation, i.e. from more generic to more specific concepts, has been realised. However, since every domain model has been established independently of other domain models, horizontal harmonisation, i.e. cross-domain harmonisation is poor and similar concepts with slightly different definitions or different properties may have been defined in two or more different domains. They are specialisations of the same class defined in the NEN 3610 global ontology, but their similarity has not been considered by the domain modellers. Having a global ontology did apparently not ensure harmonisation across domains.

Some attempts have been made to realise better harmonisation across domains, i.e. when a model was updated or newly developed. However, sharing semantics and thus data across domain borders needed a more drastic and systematic approach for semantic alignment.

4.3 Our research in the context of related Work

Information models fall within the domain of "ontology" according to Euzenat's broad definition of this term: 'a set of assertions that are meant to model some particular domain'. There is a lot of research in the field of ontology matching. Euzenat and Shvaiko (2013) give a good overview.

As can be concluded from Euzenat's overview, matching of ontologies in a broad sense, e.g. schema integration and data integration, has been applied to the information integration problem since the 1980s. Ontology

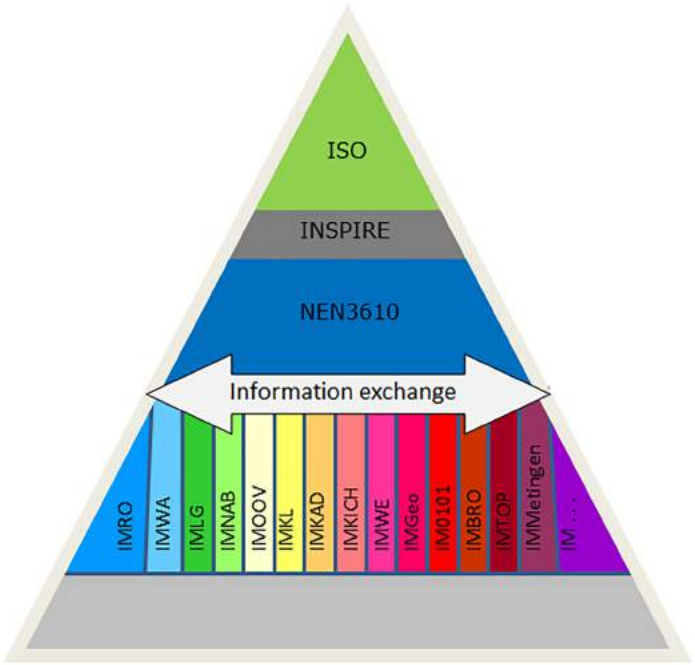


Figure 4.1: The pyramid of domain information models.

matching is the process of finding correspondences (relationships between terms) between different ontologies. The result is ontology alignment. Often, ontology matching is automated as much as possible. Most research focusses on this; again, a good overview can be found in Euzenat and Shvaiko (2013). An example of a matcher that was used in the context of SDI is Vaccari et al. (2012).

Automated matching is difficult and often requires human interpretation. There are several ways of obtaining better matching results.

First, to assure better results of the automated procedure, some initiatives have involved users in the matching process. Human users can provide an initial alignment before automatic matching is carried out, they can tune parameters of the matching system, or they can provide feedback on the alignment. Collective matching is a process where many users work on the matching together using supporting tools. In Euzenat and Shvaiko (2013) this is called 'community-driven ontology matching' and an overview of this is given in chapter 11 of their work. Community-driven ontology matching involves publishing the ontologies and obtaining alignment via tools for social interaction and collaboration. For example, users can record their feedback on alignments by annotating or voting (Correndo et al., 2008). Their input can be collected using online surveys and serious games (Dewaraja, 2010) or

using crowd sourcing (McCann et al., 2008).

Nogueras-Iso et al. (2004) studied the problem of harmonisation of metadata standards for the geographic domain and described transformations between ISO 19115 (ISO, 2014), Dublin Core (Board, 2012) and other metadata standards. This can be seen as an earlier step in harmonisation: metadata is about datasets; harmonising metadata supports the better discovery of datasets across domains. Our aim is one step further: to harmonise data models in order to be able to reuse data across domains.

Another example of ontology matching is the ISO 19146 standard on cross-domain vocabularies (ISO, 2010). This standard defines a methodology for cross-mapping technical vocabularies in the geospatial domain. The methodology proposes to take a reference vocabulary that serves as the target to which different domain vocabularies are mapped. A detailed schema of a vocabulary register is defined including terminology for concept relationships. For our study this approach was only partly implemented and on a very basic level because we were starting from a more experimental point of view. We started with a topographical model as a reference vocabulary and later on positioned all domains at the same semantic level. Our first goal was to simply bring to light related and possibly shared concepts from different domains. The second step was to relate the concepts and harmonise when possible.

Another way for obtaining matching results, is defining common definitions of concepts at an abstract level and providing transformation tools between the concepts in existing information models and in the common model. This way of harmonisation of geospatial data and models has been studied by several researchers. Cruz et al. (2004) created a semi-automatic alignment tool for matching land use classifications. The resulting alignment is used as a mapping between different data models and a global ontology. Our study does not use a global ontology. Although NEN 3610 contains a basic global ontology, this has not resulted in harmonisation between concepts on a more specific level. We therefore concentrate on harmonising more specific concepts directly.

Finally, Reitz and Kuijper (2009) describe an interactive matching process in the geo-information domain, where users are aided by visual tools, implemented in the HUMBOLDT Alignment Editor (HALE). The HUMBOLDT project, started in October 2006, was supported by the European Community through the 6th Framework Programme and had the aim of bringing together a variety of scientific, technical, economic and policy driven points of view with the aim of implementing a Framework for harmonisation of data and services in the geo-information domain (Villa et al., 2008). They state that user interaction is required to overcome heterogeneity on the se-

mantic interoperability level. This is also our observation and premise for choosing mainly a non-automatic method with emphasis on the social aspect of harmonisation. The geo-information domain has some specific matching problems. This becomes apparent from examples where concepts seem the same, but are not applied in the same way in data, due to differences in the geometric representations. For example, the concept of 'flood plane' is used in one case to define designated flooding zones, while in the other the regularly flooded area is meant. This becomes apparent from the data which is visualised in HALE, which thus aids users in judging whether possible matches are actual matches or not.

All this is relevant to our study of semantic harmonisation. However, automatic matching is not the method we chose to apply in our cross-domain harmonisation case. In the first part of our study an important reason for this was that the goal was not only to harmonise concepts but to focus on cases where better harmonisation would lead to more actual data reuse. Automatic matching would probably give us a significant amount of matches, but whether this would actually lead to harmonisation and ultimately more data reuse was uncertain. Instead, as also stated by several researchers mentioned above, user interaction is required to overcome semantic differences. The different domain models are created and maintained by stakeholders, derived from their own specific use cases. These human stakeholders are needed in the process of harmonisation; they are the ones that can assess the change of a concept definition. However, they must first become aware that it is beneficial to do so. This is a social process.

The number of concepts from all domains within the Dutch SDI together is in practice too high to search by hand for all concepts that possibly overlap. However, when the matching is done by a group of people, aided by tools, a high number of concepts can be considered in the study to find possible overlaps. For these reasons we chose not to use automatic matching techniques, but instead offer tools and methods that support data users in discovering matches between domain models. The focus was not on matching per se, but on finding and describing the issues with possible matches. This led to some interesting insights, as is discussed in the rest of this paper.

4.4 The road to semantic harmonisation: methodology

The aim of our research is to harmonise concepts defined in different domain models within the model driven framework of the Dutch SDI. This framework

Table 4.1: Methodology

What	How	Main result
Part I: Identifying semantic overlap and resolving these (section 4.5)		
Identifying semantic differences and obtaining insight into how to resolve these	desk study, interviews and workshops with domain model-owners with the focus on overlap with one specific information model (i.e. IMGeo) In depth discussions with domain model-owners	Identification of: Similar concepts modelled in different models Semantic overlap of these similar concepts Semantic difference of these similar concepts Action plans to resolve differences
Part II: Designing tools for obtaining insight in overlaps, similarities and difference and to solve the differences (section 4.6)		
Generate overview of semantic landscape of SDI	Import all classes from all domain models in one environment Apply domain-independent classification on classes	Visualisation of semantic landscape of SDI
Resolve conflicts in semi-automated manner	By discussing domain-independent semantic posters (based on neutral classification) with model-owners	Insights into needs for cross domain harmonisation via semantic posters
Publish concept-definitions crossing model-boundaries for reuse	Develop semantic concepts registry	Information modellers can easily find related concepts and assess if they are suitable for reuse

was explained in section 4.2. This section presents the methodology that we developed and applied to realise better harmonisation across the pyramid of information models.

Although 'top down' harmonisation from the most generic concepts in NEN 3610 to more specific ones in domain-specific models has been realised (i.e. the NEN 3610 pyramid assures a common way of modelling for all models extending NEN 3610), 'cross-sector' harmonisation has not because similar concepts from other domains have not been reused. Therefore, certain concepts occur in more than one domain model (e.g. building, road, water). Our goal is to achieve a high level of cross-sector harmonisation, eliminating overlapping concepts with slightly differing meanings and reusing concepts from other domains if appropriate. The methodology that we applied respects that harmonisation is a social process that starts with presenting and sharing the semantics between domains. See the steps in Table 4.1.

The first main step of this community driven ontology matching methodology was obtaining an overview of concepts defined in domain specific UML class diagrams, identifying possible conflicts (i.e. same concept; different definition) and solve these conflicts if possible. This was a labour intensive process. In a second step we therefore developed a register to better facilitate this process so that better and more harmonisation can be achieved. The

register captures and publishes all concept definitions valid in the SDI, i.e. all concepts that have been defined by the different information models. As domains all have their proper view and related language on reality, we developed domain independent classifications based on Theme, Function and Use Case. The concepts from different models were then visualised on semantic posters and semantic conflicts analysed and discussed with model-owners. This clearly showed the existing but not yet quantified need and potential for cross domain harmonisation. It was then concluded that a combined semantic register of all domains was needed, firstly to get a centralised semantic library containing all semantic concepts available and secondly to provide a base for reuse and harmonisation. A semantic register with decentralised governance was finally developed to meet these needs.

4.5 Step 1: Identifying differences between information models

The first step was obtaining an overview of concepts defined in domain specific UML class diagrams, identifying possible conflicts (i.e. same concept; different definition) and solve these conflicts if possible. This section details the methodology that was applied for this first step (4.5.1), results (4.5.2) and conclusions (4.5.3).

4.5.1 Methodology

To obtain insight into the differences and overlap between domain models that model similar concepts, we performed a comparison study starting from an information model that presumably has significant overlap with other models, i.e. Information model Geography (IMGeo). IMGeo, published in 2012, contains object definitions for large-scale topographical representations of roads, water, land use, land cover, bridges, tunnels etc. (approximately at scale 1:1k). For the comparison, we first did a desk study (where each individual model was compared with IMGeo and initial conclusions were drawn on semantic differences and overlaps). In a second step we interviewed each domain model-owner to confirm the results of the desk study. Finally, we organised two workshops with all domain model-owners to harmonise findings.

For all domain models that are extensions of NEN 3610, we studied the following questions:

a. Which concepts defined in IMGeo are also modelled in the other domain models and what is the semantic overlap and/or difference between

those similar concepts? Do the differences in modelling represent differences in reality or underlying requirements and are they justified, or can (should) the differences be harmonised?

b. What is required to achieve better harmonisation: either solving definition differences or explicitly modelling intended differences?

The models that were considered are:

- IMBRO – base registry soil
- IMBAG – base registry buildings and addresses
- IMNa - nature
- IMWa - water
- IMWOZ – taxes on real estate
- IMBRT- small scale topography
- IMLB - agriculture
- IMKICH – cultural heritage

4.5.2 Results

The results of comparing concepts in IMGeo and other domain models are presented in this section. Because of limited space, we here address the two questions for three representative domain models only. IMBRO was selected to show how some semantic differences can be solved straightforwardly in a process with all stakeholders. IMWA was selected to show the findings for an information model that was independently modelled from IMGeo and IMLB to show the findings of an information model that was developed in close harmony with IMGeo. After the presentation of findings for these three models, the main findings and conclusions of the complete study (based on all domain information models) are summarised.

4.5.2.1 Comparison IMGeo and IMBRO

The Information Model Basisregistratie Ondergrond (IMBRO) contains definitions of objects relevant for soil and subsoil. IMBRO covers four data categories, each with their own focus: measurements, permits, infrastructures and 2D/3D models.

4.5.2.1.1 a. Similar concepts and semantic overlaps with IMGeo

Only the infrastructure category has possible overlap with IMGeo. This category models public works as well as networks in the underground. Relevant features are Boreholes and Wells used for different purposes, such as monitoring water quality and quantity.

The equivalent IMGeo features are Well-borehole and Sensor-pipe (used to measure the water table), both subclasses of Engineering Element.

These feature types are modelled in IMGeo if they are visible as physical objects (i.e. topography) in the terrain. That means that non-visible pipelines and wells are not modelled in IMGeo. On the contrary, IMBRO contains a register of all pipes and boreholes, independently of their physical appearance in the field. Consequently, IMGeo contains a subset of the IMBRO boreholes and pipes, i.e. only those relevant for topography. Ideally IMGeo should reuse the IMBRO boreholes and pipes, when relevant (i.e. visible in the terrain).

4.5.2.1.2 b. How to resolve the semantic differences?

After this analysis, it was decided with all stakeholders that IMBRO would become the source for the relevant IMGEO features including their definitions. Boreholes and pipes are relatively unimportant features for topography, because they are small. Therefore, it was possible to adjust the definitions in IMGeo and to agree on a process where IMBRO will provide the information about these objects.

4.5.2.2 Comparison IMGeo and IMWA

The Information Model Water (IMWA) is the domain model for the water sector in The Netherlands. It was developed in 2001 as extension of NEN 3610 and adjusted in 2005 to align to the updates of NEN 3610. The current version originates from 2010, before IMGEO 2.0 was developed.

In IMWA all features are specialisations of a main class with a geographical description, called 'Geo-Object'; examples are Protected Areas, Groundwater Withdrawal, Engineering Element, Water, Measurement, Road and Construction.

4.5.2.2.1 a. Similar concepts and semantic overlaps with IMGeo

A comparison with IMGeo showed significant overlap with IMGeo concepts. However, since both models were developed independently, overlap or differences between similar concept definitions are mostly by coincidence, yet problematic for reusing data.

Table 4.2: Water classifications in IMGeo and IMWA

IMGeo	**IMWA**			
Sea	Backwater	Natural		
Canal	*Not available*	Non natural	Channel	Ditch
Water area	*Not available*	*Not available*	*Not available*	Sail-ditch
Dry ditch	*Not available*	*Not available*	*Not available*	Dike ditch
Not available	Water areas	*Not available*	*Not available*	*Not available*
Not available	Coast and transition	*Not available*	*Not available*	*Not available*
Not available	Wetlands	*Not available*	*Not available*	*Not available*
Not available	Wells	*Not available*	*Not available*	*Not available*

A representative example is the difference in the definition of water. IMWA defines Water as "ground surface in principle covered with water". IMGeo defines Water as: "Smallest independent area of water with homogenous characteristics, permanently covered with water". Is the same Water object type meant here or does the difference in definition serve differences in application requirements? Moreover, the classification of Water differs significantly, see Table 4.2.

Also certain concepts which are expected to be the same have been modelled very differently. Three examples are:

Constructions: IMGeo Constructions are limited to constructions that cannot be modelled as bridge, tunnel or building. In IMWA this class also contains tunnels and bridges, separate classes in IMGeo.

Weir is an important feature type in IMWA and therefore modelled with a separate class for dikes, dams, dunes, constructions and high grounds. In IMGeo these features are modelled via a wide variety of classes (Construction, Land use, Separating construction).

"Function of Roads" in IMGeo modelled as one attribute of the class Road and in IMWA as two attributes of Road: road nature and road type.

In conclusion, IMWA and IMGeo have a different modelling approach for a number of concepts. Most of the differences are a result of independent development of both models.

4.5.2.2.2 b. How to resolve the semantic differences? To better align both models, it was decided with IMWA-stakeholders that IMWA will adopt the modelling of non-water sector specific objects like roads. The differences in the types of Engineering objects will be studied in more detail to see if more alignment can be achieved. This was also decided for the differences in code lists (for example for Sensors). Some other differences do not give problems for example "wetland" is a type of Water in IMWA and type of Vegetation in IMGeo. The link between those two concepts needs to

be explicitly modelled to be able to reuse these concepts that refer to the same thing.

4.5.2.3 Comparison IMGeo and IMLB

The Information Model *Landbouw* (IMLB, i.e. rural areas) supports the subsidy provision of farmers.

4.5.2.3.1 a. Similar concepts and semantic overlaps with IMGeo The development of this information model was aligned with the development of IMGeo, and therefore a high level of semantic harmonisation has been achieved. This shows for example from how IMLB models the class Area. This can either be a physical or administrative area. The physical area can either be built objects or not. For IMGeo only the physical areas are relevant. Via a modelled relationship (i.e. association named "occurs in IMGeo as. . . .") the nonbuiltPhysicalAreas are linked to 0, 1 or more areas of different IMGeo types, for example Vegetation, Water, Non-vegetated areas and SolitaryVegetationObjects. Whenever there is a link, the semantics are reused. The strong alignment is also achieved because the IMGeo code lists are adopted within IMLB and IMLB objects are specialisations of IMGeo objects, when appropriate.

4.5.2.3.2 How to resolve the semantic differences? The conclusion of the comparison between IMLB and IMGeo is that the models are highly aligned which enables reuse of data, i.e. IMLB objects can be reused and possibly aggregated in IMGeo objects.

Some research questions remain. For example, the modelling of equivalent concepts is similar, but are the acquisition rules also the same so that data can be reused? Another question is if the links can be modelled more formally so that they can be used in automated matching approaches. Now the links are only visible when exploring the additional information for each class.

4.5.3 Conclusions of initial research on semantic overlaps and differences

This in-depth study on semantic overlaps and differences between existing information models confirms that the pyramid approach does not assure cross-domain harmonisation. Different domains are hardly aware of each other's information models. Therefore, if concepts are reused this is on an ad hoc basis. More often, domains define their own concepts when they need them,

without looking for other information models that might have modelled similar concepts.

In more detail, the following conclusions can be drawn.

4.5.3.1 Differences between models that were or were not developed independently

The issue of differently defined concepts is heterogeneous. Sometimes the domain is already working together with another domain and therefore semantics have been harmonised, while other domains that define similar concepts operate completely independently from each other.

An example of domain information models that were developed in harmony is IMLB and IMGeo. Whenever there is a link between classes in the information models, the underlying instances from IMGeo will be reused in IMLB data by referring to their unique identification. If domain models have been developed independently, similar concepts are easily defined in different ways, as was shown with IMWA and IMGeo.

4.5.3.2 Differences in modelling approaches

Another conclusion is that while all the studied information models comply with NEN 3610 which, as described in section 4.4, establishes a standard modelling method, there are still significant differences in the way of modelling on a more detailed level (for example model explicit links to other models or not; modelling a concept as class or attribute). This requires agreements on how to model a domain model and also how to model links to other models. If domain models use common approaches it is easier to harmonise them.

4.5.3.3 Reuse of concepts not authentic to the domain

We also observed that domain models define concepts, which are not authentic to that domain. Authentic means that classes, properties and relationships are essential to the domain and/or that they originate from that domain. Objects that are instances of these classes can serve as reference objects to other domains. A concept should preferably be defined by the domain to which it belongs while its definition is taken over by other domains. For example, buildings are defined in the information model “addresses and buildings” (IMBAG). They are also collected according to the IMGeo model as objects on the large scale map. In that case, IMGeo reuses both the concept and the instances; although with another geometry (BAG models

building geometries as seen from above; IMGeo models building geometries at surface levels).

4.5.3.4 Other issues that were revealed by the comparison study

1. Domain model experts are hardly aware of possible overlap with another model.
2. It is hard to discover and access the original, currently valid source of the models.
3. It is not always clear who the owner of a definition is, i.e. who has the authority to change it.
4. Definitions are ambiguous; best practises for specifying understandable and unambiguous definitions are lacking.
5. Most models act independently, i.e. relationships between them are not defined.
6. Overlap of similar concepts is not always 100%, because the context may be different. Data may be eligible for reuse in another context, but not with the exact same definitions. Design patterns for linking equivalent but yet not the same concepts may be accomplished via Linked Data.

In conclusion, although semantics within domains is well organised for SDIs, little overview nor steering on semantics across domains exists. This seriously limits the reuse of data.

4.6 Step 2: Tools for obtaining insight in overlaps, similarities and differences

The first part of our research aimed at obtaining in-depth understanding in semantic differences and overlap in existing models and used a labour intensive approach to find and resolve these. The second part of our research aimed therefore at providing tools to ease this process. Such an environment enables us to easily identify overlaps of similar concepts in different domain models. This is a necessary step before semantic differences can be solved, ultimately leading to an SDI with, as much as possible, harmonised semantics on the object level. We call such an environment a 'concept library'. This enables us to move away from an SDI as a set of standalone datasets that are

structured according to agreed information models, to an SDI that provides information on individual concepts.

This section first describes the design process of the concept library (4.6.1) and finishes with results and conclusions (4.6.2).

4.6.1 Methodology

The concept library was realised in two iterations: 1) a prototype, and 2) a production system.

4.6.1.1 Concept library prototype – Creating semantic posters.

In the first iteration the concept library was created as a prototype which was an extension of a UML modelling environment. To establish the "information concept" based approach, we first imported all UML models of the Dutch SDI into one UML modelling environment. All Dutch spatial information models were requested from the organisations that are responsible for their maintenance. These were usually not published on the web, but had to be provided to us by the maintainers of the models. The 34 INSPIRE thematic data specifications were already available for download as one integrated UML project. These were also included in the prototype. In this stage we considered all UML Classes to be "concepts" and we called the whole a "concept library". Concepts also included semantic concepts defined in code lists. This was done because harmonisation and reuse of code lists is an important part of the harmonisation process. Note that at this stage the content of the code lists, the values, were not included.

All concepts from the UML-based concept library were automatically loaded into a spreadsheet, so that all classes of all information models were listed with their definitions. The list contains almost 500 concepts from Dutch models and around 900 from INSPIRE models (1400 in total). Information analysts then applied classifications to the concepts, by assigning keywords of the classifications to all individual concepts. Three classifications were used representing different high level concepts (see Table 4.3) based on 1) a functional viewpoint, i.e. classifying concepts based on their function in the world, 2) a thematic viewpoint which corresponds to how things are often classified in topography, and 3) a use case viewpoint based on high level use cases.

The classification types were chosen in a plenary stakeholder session and developed on our own experience, i.e. no reference to existing classifications was made. The motivation for the three classifications (functional, thematic and use case) is that these different viewpoints were expected to differentiate

Table 4.3: keywords for the classification of concepts in the concept library

Thematic classification	**Functional classification**	**Use case classification**
Physical – nature	Addresses	Noise
Physical - manmade	Living / accommodation	Air
Administrative Regulatory	Cadastre / rights	Nature
Populations	Maintenance / administration	Archaeology
Health and risks	Traffic / transportation	Safety
Supportive	Agricultural	Water
Network topology	Nature and landscape	Soil
Economical production units	Recreation	Space
Measurements - monitoring	Retail trade	Other
Other	Public service	Not applicable
	Technical infrastructure	
	Economic activity / Industry	
	Other	

enough within models, and aggregate enough between models. If this proved to be not the case classifications could be added: the approach allows for an arbitrary number of classifications. All classification lists contained an 'other' keyword to allow unclassified concepts that can be subject of further discussion. The principal goal of the applied classifications was to relate concepts from different models according to similarity in assigned keywords. As was expected, the classifications indeed provided a common semantic view overarching the individual domain classifications.

As an example: the thematic keyword Physical-nature was assigned to concepts like Plantcover, SolitaryVegetationObject, Biomarker, and Wetland. Concepts can also have several keywords assigned. A water retention can for example be natural or manmade. Then the spreadsheet was re-imported into the UML-based concept library and based on the assigned keywords, the concepts were added to corresponding collections. Each unique keyword led to the creation of a collection of concepts, with all concepts that had that keyword assigned members of that collection.

After adding the keywords and re-importing the thus enriched concepts in the concept library, the third step consisted of queries that could now be used to find all classes from different models that were part of a certain collection or combination (i.e. intersection) of collections. The three different classifications allowed for combinations of keywords that provided useful common

semantics for grouping concepts. Of course not all combinations were useful, for instance a functional keyword Addresses, combined with the use case keyword Nature will not provide any concept. But a theme Administrative combined with use case Agriculture combines all agricultural administrative concepts like Parcel, Site, Production unit, Animal unit. Concepts that belong to the same collection could be visualised together on a diagram even though they might be from different information models. In this way, a number of diagrams were created semi-automatically based on generated queries for all combinations of functional and thematic keywords. So the total of over 1400 concepts was now presentable in subsets according to the assigned classes (Figure 4.2). The resulting diagrams were a good basis for further study of classes that, based on the collections they were members of, had been grouped together.

These diagrams were the basis for different "posters" that were printed, showing a combination of concepts from different domain models for a specific use case. The use case posters were compiled by querying different combinations of function and theme, for example function "health and risks" in combination with theme "agriculture". These posters[1] were used to analyse differences and overlap, both as a desk study and in several interactive workshops with stakeholders (data owners and model owners). This created awareness about the possibilities and opportunities for aligning concepts from different information models. During these sessions, several (dis)alignment issues were discovered. The fact that this was done by the stakeholders themselves, improves their willingness to cooperate in aligning similar concepts.

An example poster of all relevant classes for the combination of theme 'administrative' and function 'Addresses' is shown in Figure 4.3.

These overviews enabled to browse the concepts in a thematic, cross-domain fashion and thus to reveal semantic overlap and differences that were hidden until now. The resulting list is public and is being used to solve alignment issues and to reach a higher level of semantic harmonisation[2].

Examples of such harmonisation issues are:

- The concept "building" appears in INSPIRE, IMDBK, IMLB, IMBAG, IMGeo, TOP10NL and IMWOZ: how can the reuse of concepts be modelled?

- IMWA and IMGeo both model "well" with different definitions, can this be harmonised?

[1] see `http://www.geostandaarden.nl/geoconceptregister/posters/index.htm`
[2] See `http://www.geonovum.nl/melding/project/52/geosemantiek`

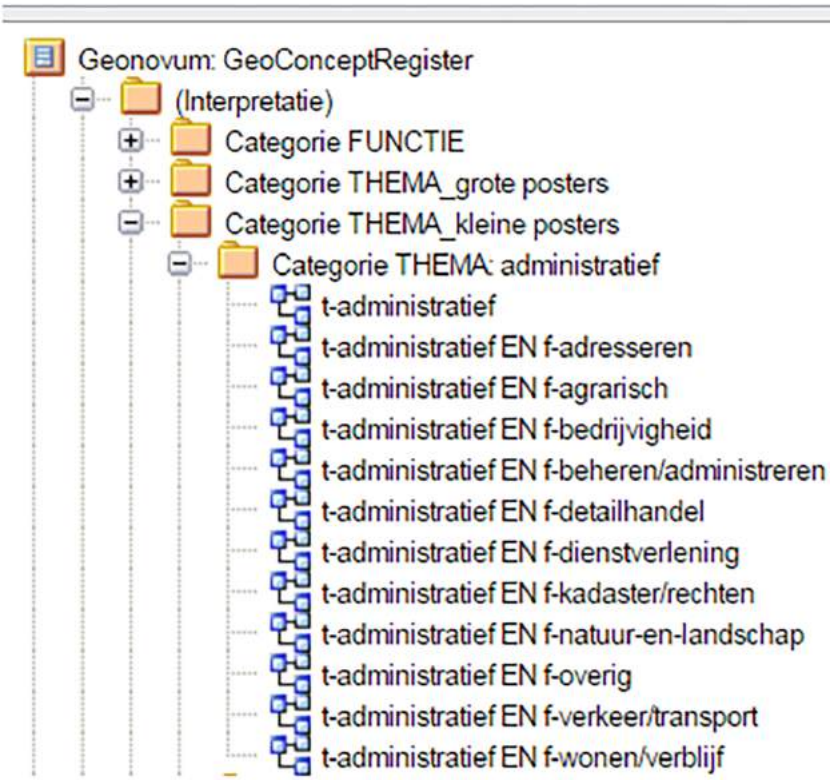


Figure 4.2: Snapshot of the concept library prototype showing a list of diagrams with the results of combining the thematic keyword 'administrative' (prefix 't-' stands for 'thematic') with all functional keywords: 'addresses', 'agricultural', economic activity, administration, retailing, services, cadastre/rights, nature and landscape, other, traffic/transportation, living/residence. (prefix 'f-' stands for 'functional') .

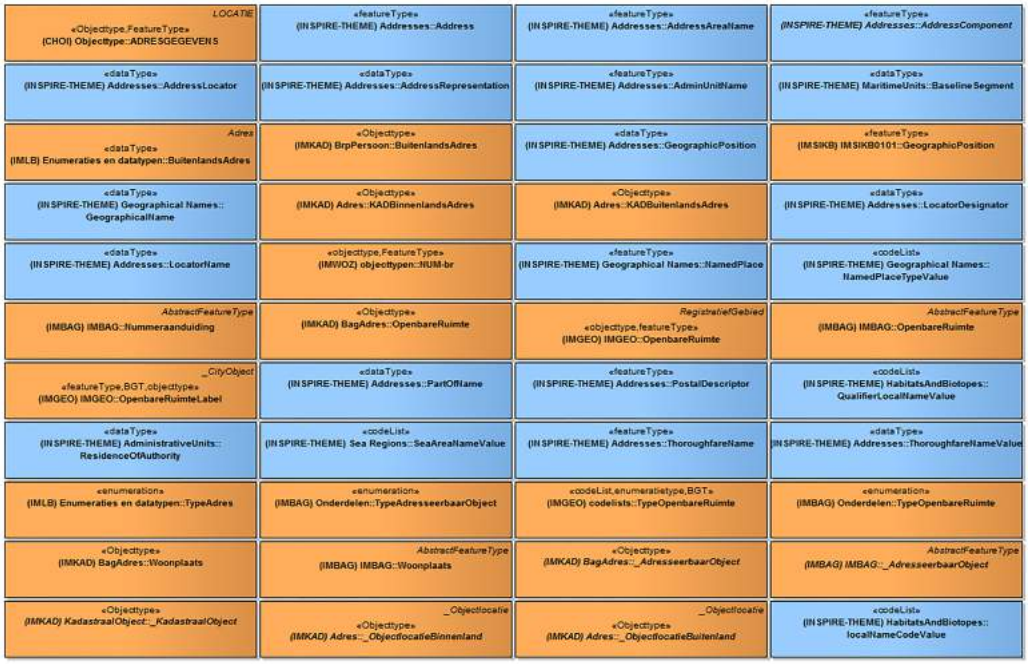


Figure 4.3: Example poster: Administrative & Addresses. Concepts from Dutch models are in orange, INSPIRE concepts in blue. The figure shows that related concepts such as 'Woonplaats' (residence) from different models, IMKAD, IMBAG, INSPIRE, IMGEO, IMSIKB, IMLB are presented in one view to facilitate discussion on semantics .

- Do (INSPIRE)LandWaterBoundary and (IMWA)Oever (Bank in English) model the same concept?
- Definitions of measurements are different in IMKL and IMRO
- INSPIRE Air transport does not have an equivalent in Dutch domain models
- Different attributes are used to model the same concepts for railway and roads in IMWA, IMGeo and TOP10NL
- etc.

4.6.1.2 Concept library – Web based registry (production system)

The UML-based concept library from the first iteration was useful for the creation of thematic posters that could be used in discussions with stakeholders. However the concept library was only available in a desktop UML tool and therefore not publicly accessible. In the second iteration, a concept library was created in which all concepts and their definitions are published on the web so that they can easily be accessed. The concepts were harvested from UML information models; classes as well as terms from code lists were included in the library.

The concept library allows the domain model owners and other stakeholders to view and compare concepts from other domains. Search functionality allows users to find concepts that contain the same word or part of the word across all domains, and view their definitions. The problem that was discovered in part 1 of our study: domain model experts hardly being aware of possible overlap with other models because of the difficulty in discovering and accessing them, is solved with this instrument.

A second advantage of the concept library is that the concepts are published using a persistent http URI, making it possible to create stable links to the concepts. For example, this can be done from an information model, a GML application schema or using JSON-LD (Sporny et al., 2014). The meaning of the data can then be accessed during data exchange. Links between concepts from different domains can now also be created, which further meets our goal of semantic harmonisation.

The concept library is open and can be accessed at `http://definities.geostandaarden.nl`. At the moment more domains are being added to the library. Five domain models are registered at the time of writing, as can be seen in the green menu bar: IMGeo (large scale topography), NEN 3610 itself, IMRO (spatial planning), IMBRT (small scale topography), and IMKL

(utility networks). Work is ongoing and cross-domain harmonisation has not yet been realised.

Figure 4.4 shows the metamodel that underlies this concept library. In this phase a metamodel was designed jointly by stakeholders. The metamodel had to provide for the semantic information in the UML class diagrams, the versioning of concepts and the publication on the web. The developed metamodel was partly related to the source data (UML models) and partly to the target. The SKOS metamodel was therefore not yet considered at the conceptual level. In the implementation stage this metamodel was mapped to SKOS terminology. All entities in the model have an http URI as identifier. A Concept is always part of one Domain (i.e. the information model it originates from) and has one or more versions. To differentiate between concepts in UML stereotyped as feature types and concepts as values in code lists a Value class was added and it's associated ValueList.

All properties of the concept are versioned except for its URI identifier, which is persistent. A concept's properties include its name(s), definition, possibly an additional clarification and/or illustrations, and some metadata. A concept version has a 'super type', indicating its parent concept in the taxonomy of concepts. In addition concepts can be associated with zero or more value lists.

A concept version can also link to one or more related concepts via the ConceptRelation. Four possible relationship types are specified in the RelationType enumeration. At the introduction of the concept library the relations between the concept instances were not specified since the concepts are provided per domain and across domains concepts are not yet harmonised. These are specified during the harmonisation process by domain model owners. The first relationship type they apply is 'similar', indicating that two concepts that have been thus linked are candidates for harmonisation.

We considered the following alternatives for classifying the relationships between concepts more precisely. We selected the first alternative, SKOS, as it combines simplicity with as much functionality as we needed at this stage.

- In SKOS concepts can be linked to other SKOS concepts via semantic relation properties (Miles and Bechhofer, 2009). The SKOS data model provides support for hierarchical and associative links between SKOS concepts. For hierarchical links the properties skos:broader and skos:narrower can be used indicating that the related concepts are respectively broader and narrower. To assert an associative link the skos:related property can be used. These semantic relations are only intended to be used if the related concepts have been designed together and are hence part of the same concept scheme. When aligning concepts

of different concept schemes, mapping properties can be defined between the concepts. The various mapping properties indicate different levels of similarity between the concepts. Properties that can be used include: skos:mappingRelation, skos:closeMatch, skos:exactMatch, skos:broadMatch, skos:narrowMatch and skos:relatedMatch. In case of an exact match the preferred outcome of the harmonisation process may be a full merge of the concept resulting in a library with one less concept.

- OWL 2 is an ontology language for the semantic web. It has many well defined language elements to describe the relationships between concepts and parts thereof. Using OWL, formal ontologies are created. However, this level of semantic precision was beyond our requirements at this stage.

- A well-known process in mapping is cartographic generalisation; in this process the number and complexity of objects on a map is reduced to create a map of a different scale. Typical operations that are used in generalisation are: selection, simplification, combination, smoothing or enhancement (Burghardt et al., 2014). When concepts in the concept register are related via a cartographic operation this could be registered in the concept register, for example the building in TOP10NL is a simplified version of the building in IMBAG. However, a standardised relationship to express this was not found.

- The concept of aggregation is available as a relation type in UML modelling and often used in geospatial modelling, where it is used to model that instances of two classes have a part-whole relationship. More specifically, the geometries of the part-objects together form the geometry of the whole object. For example, administrative units at different levels often have part-whole relationships. This relation type is not available in SKOS or OWL, although part-whole relationships are available on the data level in e.g. GeoSPARQL (Perry and Herring, 2010), Dublin Core and WordNet (Miller, 1995).

This metamodel can easily be mapped to SKOS, which is the language underlying the concept library.

Note that instead of creating one environment in which all concepts and definitions are published, the linked data approach could have been used: all concepts and definitions could have been published on the Web with http URIs by the domain model owners, and links could have been realised between concepts even though they are published in a distributed manner on

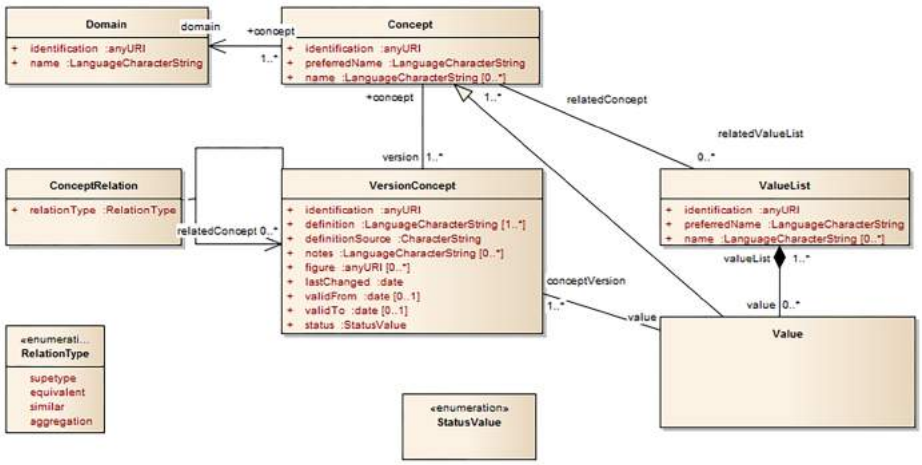


Figure 4.4: Metamodel of the concept library

the web. However, it would have required all domain model owners to create their own solution for publishing their concepts on the web. Therefore, we created a central repository instead from which all the concepts are published. An additional advantage is that functionality for browsing and searching concepts across domain models is easy to provide. For the future however it is foreseen that using linked data to provide a distributed platform in which semantics are published on the web and maintained by the domains is the preferred solution. The concept libraries architecture provides for this future requirement.

4.6.2 Designing the concept library for harmonisation: results and conclusions

The realisation of the concept library led to new conclusions and insights about semantic harmonisation within SDIs. The most important conclusion is that gathering all concepts from different domain models in one environment, and providing ways to browse (possibly) similar concepts across these domains, supports stakeholders in discovering overlaps and also in their willingness to address these overlaps. Other conclusions are:

Although we have not specified relations between concepts from different domains, we have found existing relation types from SKOS and OWL that can be applied to harmonise our domain models. However, some relation types that are needed in the geospatial domain, like the part-whole relation, are missing on the class level.

Differences between similar concepts in different vocabularies are, in some

cases, justified. These are often related to requirements for the spatial properties of concepts and stem from different domain specific needs. In addition, we observed that concepts in the geospatial domain models often describe different representations of things because they represent a different view on the same real world objects. This could partly explain the overlap we observed.

The establishment of a concept library as a derived product from existing UML information models raises the question of where the source of semantics is published. The information models started as a source of semantics but the concept library evolves to a product that eventually becomes the authentic source of semantic concepts.

4.7 Conclusions and Future work

This paper presents the methodology that we developed to fundamentally improve harmonisation of a model-based SDI. The starting point was a set of different domain models, modelled in UML, operating well within the own domain but all with very little linkages to or awareness of other domain models that may model similar concepts. The methodology consists of two steps. The first step analyses all existing domain models on overlap and differences in order to understand the differences and to identify which concepts in the models relate to the same concept in reality (section 4.5). The second step aims at establishing an environment that provides information on all individual concepts within the SDI and explicitly modelling overlap between concepts to be able to reuse semantics in the SDI (section 4.6).

One of our observations is that it requires human interpretation to solve semantic irregularities, (i.e. slightly different concept definitions which may or may not be intended) between existing domain models. Therefore, we chose not to use automatic matching techniques; instead our methodology is based on harmonisation by humans. Since the number of concepts from all domains within the Dutch SDI is too high to search by hand for all concepts that possibly overlap, we developed tools to help domains modellers in discovering matches between domain models themselves, by organising and publishing available concept definitions. This approach makes it possible to include a high number of concepts in the study to find semantic overlaps. In addition, information modellers can easily reuse existing concepts in new models to be developed.

After differences have been resolved and similarities between concepts in different domain models have been identified, the next step is to apply a standardised way of modelling such relationships in order to reuse concepts

(and eventually data). From our inventory of concept relationship types, we can conclude that the spatial domain has semantic relationships between concepts that are not yet well described in the semantic domain. Further steps are necessary to study these semantic relationships, define new relationship types where necessary, and make them available for use.

Another observation is that in the geospatial domain, information models that have semantic overlap often describe different representations of things because they represent a different view on the same real world objects. For example, there are two distinct concepts for building in two Dutch domain models, IMGeo and BAG, because one models the building as a ground level representation, and the other as a representation of the building as seen from above. On the instance level, there are two sets of instances, which both represent the same set of buildings in the real world. In general, we observed that within the information models of the Dutch SDI, there is no concept representing the real world but only views thereof. Thus, it is not possible to model, for example, all Building concepts as representations of the same real world thing. Further research is necessary to confirm this and provide ways of dealing with the harmonisation issues accordingly.

Bibliography

DCMI Usage Board. DCMI metadata terms. Dublin core metadata initiative. Available online: http://dublincore.org/documents/dcmi-terms/ (accessed 2-February-2016), 2012.

Dirk Burghardt, Cécile Duchene, and William Mackaness. Methodologies and Applications of Map Generalisation. In *Publications of the International Cartographic Association.* ICA, 2014.

Open Geospatial Consortium et al. Geography Markup Language (GML) Encoding Standard. Version 3.2.1, doc nr OGC 07-036 [online]. Available online: http://portal.opengeospatial.org/files/?artifact%5Fid=20509, 2007.

Gianluca Correndo, Harith Alani, and Paul Smart. A community based approach for managing ontology alignments. In *Proceedings of the 3rd International Conference on Ontology Matching-Volume 431*, pages 61–72. CEUR-WS. org, 2008.

Isabel F Cruz, William Sunna, and Anjli Chaudhry. Semi-automatic ontology alignment for geospatial data integration. In *Geographic Information Science*, pages 51–66. Springer, 2004.

S Dewaraja. *CLONTY: a scalable approach to collaborative ontology alignment*. PhD thesis, B. Sc. thesis, The University of Westminster, Westminster, UK, 2010.

Jérôme Euzenat and Pavel Shvaiko. *Ontology matching, second edition edn.* Springer, 2013.

Dragan Gaševic, Dragan Djuric, and Vladan Devedžic. *Model driven architecture and ontology development*. Springer Science & Business Media, 2006.

Joao Hespanha, Jan van Bennekom-Minnema, Peter van Oosterom, and Christiaan Lemmen. The model driven architecture approach applied to the land administration domain model version 1.1-with focus on constraints specified in the object constraint language. In *fig working week 2008, integrating generations*, page 19, 07 2008.

INSPIRE. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Available online: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002 (accessed 2-February-2016), 2007.

INSPIRE. D2.5: Generic Conceptual Model, Version 3.4. Available online: http://inspire.ec.europa.eu/documents/Data%5FSpecifications/D2.5 %5Fv3.4.pdf, 2014.

ISO. ISO 19107:2003 Geographic information – Spatial schema., 2003.

ISO. ISO 19110:2005 Geographic information – Methodology for feature cataloguing., 2005.

ISO. ISO 19131:2007 Geographic information – Data product specifications., 2007.

ISO. ISO 19146:2010 Geographic information – Cross-domain vocabularies., 2010.

ISO. ISO 19115-1:2014 Geographic information – Metadata – Part 1: Fundamentals., 2014.

ISO. ISO 19103:2015 Geographic information – Conceptual schema language., 2015a.

ISO. ISO 10109:2015 Geographic information – Rules for application schema., 2015b.

Kilian Kiko and Colin Atkinson. A detailed comparison of UML and OWL. *Reihe Informatik*, TR-2008(4), 2008.

Robert McCann, Warren Shen, and AnHai Doan. Matching schemas in online communities: A web 2.0 approach. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 110–119. IEEE, 2008.

Alistair Miles and Sean Bechhofer. SKOS simple knowledge organization system reference. Available online: https://www.w3.org/TR/2009/REC-skos-reference-20090818/, 2009.

George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

NEN. NEN3610 Basismodel Geo-informatie - Termen, definities, relaties en algemene regels voor de uitwisseling van informatie over aan de aarde gerelateerd ruimtelijke objecten. Available online: https://www.nen.nl/NEN-Shop/Norm/NEN-36102011-nl.htm, 2011.

Javier Nogueras-Iso, F Javier Zarazaga-Soria, Javier Lacasta, Rubén Béjar, and Pedro R Muro-Medrano. Metadata standard interoperability: application in the geographic information domain. *Computers, environment and urban systems*, 28(6):611–634, 2004.

MDA OMG. Guide Version 1.0.1. *Object Management Group*, 62:34, 2003.

Matthew Perry and John Herring. OGC® GeoSPARQL-A geographic query language for RDF data. Available online: http://www.opengeospatial.org/standards/geosparql (accessed 2 February 2016), 2010.

Thorsten Reitz and Arjan Kuijper. Applying instance visualisation and conceptual schema mapping for geodata harmonisation. In *Advances in GI-Science*, pages 173–194. Springer, 2009.

M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, and N. Lindström. JSON-LD 1.0: a JSON-based serialization for linked data, 2014.

Lorenzino Vaccari, Pavel Shvaiko, Juan Pane, Paolo Besana, and Maurizio Marchese. An evaluation of ontology matching in geo-service applications. *GeoInformatica*, 16(1):31–66, 2012.

Linda van den Brink, Jantien Stoter, and Sisi Zlatanova. UML-Based Approach to Developing a CityGML Application Domain Extension. *Transactions in GIS*, 17(6):920–942, 2013.

P Villa, T Reitz, and M Gomarasca. HUMBOLDT project for data harmonization in the framework of GMES and ESDI: introduction and early achievements. *The international archives of the photogrammetry, remote sensing and spatial information sciences*, 37:1741–1746, 2008.

W3C. OWL 2 web ontology language overview. Available online: http://www.w3.org/TR/2009/REC-owl2-overview-20091027/, 2009.

Geospatial Linked Data

Chapter 5

Linking spatial data: automated conversion of geo-information models and GML data to RDF

Authors: L. van den Brink, P. Janssen, W. Quak & J. Stoter (*International Journal of Spatial Data Infrastructures Research*, 9:59-85, 2014).

This paper has been published in a peer-reviewed scientific journal in 2014 and is published unchanged in this chapter, except for footnotes where necessary to supply current information about an outdated statement. It studies how geospatial data can become part of the wider semantic web, by studying how to apply the generic Linked Data paradigm to disseminate geospatial data outside the traditional geospatial data sector. The first part of the paper focuses on deriving linked data from GML data in a generic way. This includes the creation of URI identifiers and selecting an encoding for geometries in linked data. The second part describes how more meaningful RDF, where the data is enriched with semantics, can be created from GML, given the underlying information model, by transforming it from UML to RDFS/OWL.

Contributions: my research was used by the Joint Research Centre (JRC) as input for the development of a method and guideline for deriving OWL vocabularies from the INSPIRE UML models.

text of published paper starts after this line

Abstract Linked data provide an alternative route for the dissemination of spatial information compared to the traditional SOA-based SDI approach. The traditional approach has provided a wealth of standardized and structured location data based on Geography Markup Language (GML), while linked data provides an open mechanism for sharing and combining this data with anything, once the data is available as linked data. The first part of the paper focuses on deriving linked data from GML data. In the second part, we study how more meaningful data, expressed in Resource Description Framework (RDF) can be created from GML, given the underlying information model, by transforming it from Unified Modeling Language (UML) to Web Ontology Language (OWL).

Keywords: SDI, Linked Data, Semantic Web, GML, RDF

5.1 Introduction

Linked data provide an alternative route for the dissemination of spatial information compared to the traditional service-oriented architecture (SOA)-based Spatial Data Infrastructure (SDI) approach. Where the latter is built on predefined structuring of semantics within domains, linked data is open to linking information to any data over the Web. In this respect both are complementary: the traditional approach provides a mechanism for a basis of standardized and structured data within domains, while linked data provides an open mechanism for sharing and combining data. GML (OGC, 2012) as the ISO standard for exchange of service based spatial data and Resource Description Framework (RDF) (Klyne and Carroll, 2004) as the linked data format are therefore related. GML provides the format in which many spatial datasets are available and exchanged. This standardization process and effort has been realized on a large scale. The web of linked data could profit from this effort, as large amounts of standardized spatial information could be made available as linked data. This article will focus on the use of GML structured data as a source for deriving RDF structured data.

The first part of the paper focuses on deriving linked data from GML data. The first version of GML, v1.0 (Lake and Cuthbert, 2000), was based on RDF. From version 2.0 onwards GML was based on Extensible Markup Language (XML) and XML Schema, but the object-property structure was retained. We describe a transformation for translating any correctly structured GML to RDF automatically, using Extensible Stylesheet Language Transformations (XSLT) (Kay, 2007). Because GML's object-property structure translates very well to triples, the transformation is straightforward. Well-known GML content elements such as names and descriptions are mapped to their

RDF equivalent. Geometries are transformed to Well Known Text (WKT), a compact format for expressing geometries described in Simple Features (ISO, 2014). However, any semantics specific to the input GML data (a.k.a. the application schema) are ignored in this translation.

In the second part, we study how more meaningful RDF can be created from GML, given the underlying information model, by transforming it from UML to Web Ontology Language (OWL) (W3C, 2012a). There exists a straightforward mapping to convert a UML model into an OWL vocabulary. However, the re-use of existing concepts in vocabularies takes a central role in OWL while in UML the use of vocabularies is not supported. We describe how annotating the UML model could improve this translation.

5.2 Spatial data as reusable web resources

Linked data is receiving increasing interest as a technology relevant for geographic information. It provides an alternative route for the dissemination of spatial information compared to the already considered traditional SOA-based SDI approach. The key difference is that linked data is much more flexible and open. Where the SDI approach is built on predefined structuring of semantics within domains, linked data is open to linking information to any data over the Web. In this respect it much more appeals to the web 3.0 philosophy: unique information features that are available on the web and can be accessed and extended at any time, by anyone and for any purpose (CEN, 2012). In the Eye on Earth White Paper (Smits et al., 2011) the Linked Data approach and technologies are recommended for accessing resources in environmental information systems. However, the implication of this on what has been done and realized in the spatial information infrastructure until now is unclear. We argue that it is complementary. The traditional approach provides a mechanism for a basis of standardized and structured data within domains, while linked data provides an open mechanism for cross-domain sharing and combining. The traditional approach is characterized by a service based dissemination of GML structured data. In that approach data specifications provide clear definitions of semantics in predefined domains and use cases. These are implemented in XML schema, providing a well-defined and verifiable means of information exchange. The strong point of it is that the proper purpose of standardization and harmonization, interoperability, can be addressed through agreement and sharing of vocabulary. Once agreed the requirements and rules for communication are set and can be implemented in a verifiable way. The quality of implementation can be measured and therefore managed.

However, there is a downside: the semantics are defined within information domains. This results in predefined information silos, each related to different information domains. Within the context of a silo interoperability is assured by shared and foreseen concepts, but between silos little harmonization takes place, and for not yet foreseen concepts and relations the structure is too rigid. This is exactly the weak spot where linked data can be of help. It allows data to become part of a web of data where it is integrated with other data and data models can be interrelated and harmonized.

In a Spatial Data Infrastructure (SDI), GML data are generated and served from different feature based sources. Generally transformation services will do the transformation from these local sources to the GML structured data. In many SDI projects and programmes a lot of effort has been put into that activity. For linked data, including the related RDF format and the Geographic Query Language for RDF Data (GeoSPARQL) (Perry and Herring, 2010), a similar approach can be followed: transformation services that act on the same local sources, generating linked data instead of GML, and making it available through GeoSPARQL endpoints. However, another option is to reuse the already existing GML sources. There are large amounts of existing feature data represented either in a GML file (or similar serialization) or in a data store supporting the general feature model (ISO, 2005). For example, the INSPIRE Directive of the European Union (INSPIRE, 2007) will lead to a lot of geospatial data becoming available in a standardized way. In addition, on the national level, e.g. in the Netherlands, more and more data is being published using Open Geospatial Consortium (OGC) standards. Since these data are already structured in a standardized and defined way, RDF transformation can be standardized as well. Linked data can therefore build on the structure already provided. One would expect a simple rule of benefit: profiting twice by reusing work that has been done once.

5.3 Related work

The problem of publishing spatial data as linked data is addressed in Schade and Cox (2010), who show that GML and RDF are isomorphic and suggest the addition of a façade on top of SDI web services to transform GML response data into RDF. Smits et al. (2011) describe linked open data as a new paradigm for sharing resources and recommend RDF for making data available on the web. They suggest that linked data based approaches avoid the heavyweight platform, application, and interface issues inherent in many more traditional SDI approaches. López-Pellicer et al. (2012) discuss prin-

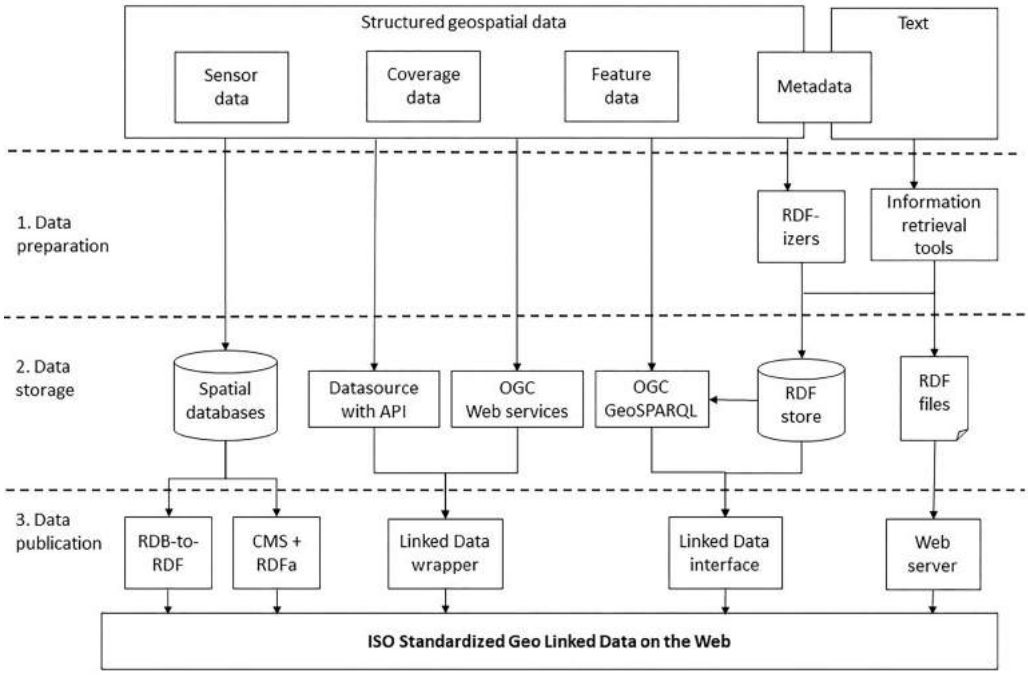


Figure 5.1: Source: López-Pellicer et al, 2012

cipal ways of integrating linked data in a spatial data infrastructure. The following diagram (Figure 5.1) from their publication depicts several ways of building linked data on top of an SDI:

As shown in this diagram, GML to RDF transformation can be positioned in the RDF-izer part. The Linked Data Wrapper is also of interest in this context. As mentioned, the idea was already described by Schade and Cox (2010). An implementation of such a wrapper, which creates linked data on top of OGC web services is described in Ciardi et al. (2013). An ongoing effort to create spatial linked data sets is LinkedGeoData, an RDF representation of OpenStreetMap data, interlinked with other spatial data sets (Stadler et al., 2012).

A recent, thorough overview of ways to encode geometry in RDF can be found in Athanasiou et al. (2013). This report was produced in the context of the GeoKnow project, a 3-year project funded by the European Commission. GeoKnow addresses the problem of linking geospatial data from heterogeneous information systems and uses the web as a platform to exploit this data (Project, 2013). Another relevant overview is given in the Core Vocabularies specification, which was developed as part of the Interoperability Solutions for European Public Administration (ISA) programme of the Eu-

ropean Union (ISA, 2012a). It has a section on location, the Location Core Vocabulary, which was recently installed in World Wide Web Consortium (W3C) space (Perego et al., 2013). The Location Core Vocabulary defines three classes, Location, Address, and Geometry, and a few properties for describing places in terms of their name, address or geometry.

An early effort to create and implement rules for transforming UML models into OWL ontologies was carried out in the context of an OGC Web Services Testbed (OWS-8) and is described in Hobona and Brackin (2011). Standardization efforts for such rules are currently underway within ISO/TC 211 project 19150-2 (ISO, 2013). This work will be published as an ISO Standard, and is currently in Draft International Standard (DIS) status[1]. Experimentation with this mapping is being undertaken, for example by Cox (2013), who describes an explicit OWL representation of the ISO/OGC model for observations and measurements based on the ISO 19150-2 rules.

Tschirner et al. (2011) achieved good results in transforming GML data models into OWL ontologies. We therefore used the experiences of these researchers and further develop the ideas and initial experiences into implementations to convert GML data to RDF instances accordingly. Also, we introduce more linked data geo vocabularies than GeoSPARQL to express geo data as linked data on the web and we provide, as main contribution, a method to link information models to existing vocabularies. New in our work is also the use of the above-mentioned standard ISO19150-2 that provides a set of rules to convert UML to OWL.

Another important topic in this context is how to create persistent, unique identifiers for data on the web. Developing a Uniform Resource Identifier (URI) strategy is one of the first steps needed to publish (geo) data as linked data. The W3C (and other experts groups) has led the development of guidelines and best practices on how to design "good" URIs in general and for the government domain in particular (W3C, 2012b). In addition, several coordinating organizations issued guidelines on persistent URIs, such as the guidelines of ISA for European governments (ISA, 2012b), the guidelines for the UK public sector (Davidson, 2009) and the URI Design Principles of Linking Open Government Data (LODG, 2013). Abbas and Ojo (2014) worked on the improvement of the implementation of such guidelines by consolidating existing URI design rules, distilling core URI design aspects or facets from these rules, and abstracting the rules into a set of consistent URI Design Patterns specifications. In our work we use the URI strategy designed by Overbeek and van den Brink (2013); Overbeek and Brentjens (2013) that describe guidelines for formulating URIs for the Dutch public sector.

[1] [Added2018] Published as an ISO standard in 2015.

5.4 Research questions and method

Geospatial data is already structured in a standardized and defined way, making it possible to standardize RDF transformation of this data as well. The GeoSPARQL standard specification recognizes this by mentioning the development of standard processes for converting (or virtually converting and exposing) these data to RDF as beneficial future work (Perry and Herring, 2010).

The challenge therefore is to investigate a way of generating linked data out of GML. Both GML and linked data have a level of data instances and a level of data model. In GML, data structures are defined in GML application schemata as extension of the abstract GML elements. Usually this data structure is expressed in a UML information model which defines the semantics and from which the GML application schema is automatically derived. In a linked data context, the data instances are expressed in RDF and the underlying model can be described as an ontology in OWL. Rules for transforming UML models into OWL ontologies are being developed within ISO/TC 211 project 19150-2 (ISO, 2013). There are several issues with the rules as described in this ISO standard, related to the closed world assumption of UML versus the open world assumption of linked data; the connection between concepts in a UML model and related concepts in vocabularies from the Semantic Web; and modelling conventions and restrictions in UML which are absent in OWL.

Several questions related to generating linked data out of GML are examined in this paper. First, is it possible to describe a generic transformation, without knowledge of the information model of the data, from GML to RDF? Second, how is geometry to be encoded in the resulting RDF? And third, how can GML application schemata, or the UML information models in which their structure and meaning are defined, be automatically transformed into an OWL representation, and integrated with other ontologies in the Semantic Web?

To address the first question, in section 5.4.1 we examine GML and to what extent it is similar to RDF. In section 5.4.2 we consider how geometry should be encoded in RDF by studying existing vocabularies and use cases. This gives us an overview of what options and criteria there are and allows us to select a method of encoding geometry. In section 5.4.3 we address the construction of URIs for our resources in RDF based on the Dutch URI strategy. The next step, which we describe in section 5.4.4, is to implement an experimental generic GML to RDF transformation in Extensible Stylesheet Language Transformations (XSLT) (Kay, 2007) and to test it with existing GML data. This allows us to determine if a generic GML to RDF

transformation is possible and answers the first question. Note that the geometry2RDF tool (Group, 2014) could have been an alternative to transform GML to RDF. However the source data of this tool is Oracle with geometry described in spatial columns and not in GML, as in our case.

Finally, in section 5.4.5 we describe our study of the third question by experimenting with an automated mapping from UML to OWL and with annotating the UML to better integrate the resulting ontology with existing ontologies or vocabularies on the web. For the experiments, we use a sample GML file containing land use plans, conforming to the Dutch standard IMRO (Information Model Ruimtelijke Ordening - spatial planning), and the IMRO UML information model (Geonovum, 2012).

5.4.1 GML: A Triple Structure

GML (Geography Markup Language) is a standard for the storage and transport of geographic information. The first version of GML, v1.0 0, was published in May of the year 2000. Key concepts in the GML model of the world are the "feature": an abstraction of a real world phenomenon; the "geographic feature": a feature which is associated with a location relative to the Earth; and the "feature collection", a collection of features which can itself be regarded as a feature and gives a digital representation of the real world. Features have properties; geographic features have properties whose value may be a geometry.

GML 1.0 used a geometry model called 'Simple Features', with definitions for point, line string, polygon, and some other basic geometric shapes. In addition it provided a coordinates element for encoding coordinates, and a Box element for defining extents. In its simplest form, GML contains no more semantics than this: geographic features with associated geometric shapes. The standard, however, includes an extension mechanism which makes it possible to define application-specific extensions with added semantics, for example distinct object classes for River and Road, each with their specific properties.

GML 1.0 described three encoding profiles for geographic features, two of which were based on XML, while the third was based on RDF and RDF Schema (RDFS). The model of GML was consistent with RDF. GML features have properties, which can have either literal values or have a geometry object as value ('geometric value'). All features had an optional ID which could be used together with the GML document URI as a fragment identifier. In this manner, GML objects could be referenced as resources. GML 1.0 example, in which 'yourhouse' and 'myhouse' have the same location:

```
<Building ID = ‘yourhouse’ .. >
  <location>
    <Point ID = ‘134’>
      <coordinates>2455.12, 3443.78</coordinates>
    </Point>
  </location>
</Building>
<Building ID = ‘myhouse’ .. >
  <location>
    <Point resource = ‘#134’ />
  </location>
</Building>
```

From version 2.0 onwards GML was based on XML and XML Schema, and the RDF profile was no longer used. But an interesting fact is that the object-property structure, in which objects have properties and properties have either literal values or objects as values—basically a triple structure – always stayed, and is present in the current version, 3.3 (OGC, 2012).

5.4.2 Encoding Location in RDF

Resource Description Framework (RDF) (Klyne and Carroll, 2004) is a cornerstone standard of the Semantic Web. Where the World Wide Web as we know it is a web of interlinked documents, the Semantic Web is a web of interlinked data. RDF provides the foundation for publishing and linking data on the web.

RDF is slightly older than GML. The first version of the suite of standards documents related to RDF was published in February of 1999. A second version was published in 2004. RDF is to linked data what HyperText Markup Language (HTML) is to the World Wide Web: where HTML is a language for (re)presenting documents on the World Wide Web, RDF is a language for representing information about resources on the web. A resource can be a web document, about which metadata can be represented in RDF; the term ‘resource’ however can be interpreted more widely to mean anything that can be *identified* on the Web. This can be any (meta)data about anything, whether or not these things can be directly *retrieved* on the Web (Manola and Miller, 2004). In RDF, things are identified using Uniform Resource Identifiers (URIs) (Berners-Lee et al., 2005). Resources have properties and property values, property values can be URIs of other resources. RDF statements can be represented as a graph of nodes and arcs: the nodes represent the resources and values, the arcs represent the properties. The

basic resource - property - value structure of RDF can also be expressed as a triple consisting of subject, predicate, and object (always in that order). Each triple corresponds to a single arc plus the two nodes it connects in the graph.

As explained, RDF is a language for expressing any data, including geographic information, as resources. As a generic language, it has no specific features for encoding geometry, which is central to geographic information. However, several vocabularies and extensions have been proposed for this purpose. Athanasiou et al. (2013) give an overview of currently available ways to encode a geometry in RDF and of RDF stores with geospatial support. The RDF stores are evaluated by measuring their performance with several geospatial queries against RDF triples.

When storing coordinates directly in RDF instead of referencing a spatial feature from another dataset, there are several options. These differ in what they offer, ranging from only lat/long point geometries, point line and surface geometries, topology, to the possibility to use any coordinate reference system. We give a short overview of the most well-known ones. A more extensive overview is given in Athanasiou et al. (2013).

An early vocabulary for representing location data in RDF is W3C Basic Geo (Brickley, 2003). This vocabulary is explicitly created to be simple, providing just a few basic terms that can be used in RDF for describing lat(itude), long(itude) and other information about spatially-located things, using World Geodetic System 1984 (WGS84) as a reference datum. This RDF vocabulary was created with cross-domain data mixing in mind. By defining an encoding for point geometries in RDF, it makes it possible to "describe not only maps, but the entities that are positioned on the map." (Brickley, 2003). Although a W3C activity, this vocabulary is not a W3C standard nor is it in the process of becoming one. As is evident from the name, the vocabulary is very basic and has only classes 'geo:SpatialThing' (similar to GML's Feature) and 'geo:Point', and properties 'geo:latitude', 'geo:location', 'geo:longitude', and 'geo:altitude'. It has no classes or properties for topology. Acceptance is high: it is used in both GeoNames and DBPedia, both in turn highly used data sets, and in web applications and services including Yahoo! Maps.

GeoRSS is an extension of Really Simple Syndication (RSS), a family of formats for news feeds on the web. GeoRSS adds the ability to encode geometries in RSS. It offers two methods for geometry encoding: Simple and GML. GeoRSS Simple is a very basic format with point, line, box and polygon properties, and allows only WGS84. GML is a GML application profile that supports a greater range of features and adds support for other coordinate reference systems. GeoRSS can be used in RDF to make simple geographical assertions about objects. However, this is only applicable for

GeoRSS Simple, not for the GML variant. GeoRSS Simple is used in e.g. DBPedia and implemented in e.g. OpenLayers, GeoServer, Drupal, and the Google Maps Application Programming Interface (API).

GeoSPARQL (Perry and Herring, 2010) is an OGC standard for representing and querying geospatial data on the Semantic Web. It defines a set of SPARQL Protocol and RDF Query Language (SPARQL) extension functions for spatial queries, a set of Rule Interchange Format (RIF) rules (Boley et al., 2010) for transforming simple topological relation tests into queries involving concrete geometries, and a core RDF/OWL vocabulary for geographic information. In this article the vocabulary is of interest, specifically where it addresses geometry. GeoSPARQL is based on accepted standards from the geospatial domain: the General Feature Model (ISO, 2005), Simple Features (ISO, 2014), Feature Geometry (ISO, 2003), and Structured Query Language Multimedia (ISO/IEC, 2011). It defines three classes, 'geosparql:SpatialObject', representing "anything that can have a spatial representation" (Perry and Herring, 2010), (p. 6), and its subclass 'geosparql:Feature'; and 'geosparql:Geometry', representing the top-level geometry type. Also, several properties are defined for associating features with geometries and for recording metadata on geometries. In addition, it defines several sets of properties representing topological relations. The vocabulary allows two serializations for geometry, WKT and GML.

WKT is a text based format for encoding geometries, defined in the Simple Features specification. It is not only used in RDF, but is supported in several spatial databases and APIs. It is supported in several RDF semantic stores with geospatial capabilities, such as Virtuoso and uSeekM. The WKT option in GeoSPARQL allows only simple feature geometry types, but this is still a wide range of geometry types such as points, curves, surfaces and geometry collections. The GeoSPARQL vocabulary defines a property 'geosparql:asWKT' in which a geometry can be recorded as a text value. It is possible to use any coordinate reference system (CRS); a reference to the used CRS is recorded with the coordinates.

The GML option in GeoSPARQL allows all ISO 19107 spatial schema geometry types, which is a much wider range than the simple features allowed in WKT, including a lot of less commonly used types. The vocabulary defines a property 'geosparql:asGML' in which a geometry can be recorded as a GML literal, i.e. a geometry element from the GML schema can be embedded in the RDF. The 'geosparql:asGML' serialization is implemented in several tools (Athanasiou et al., 2013), and is offered as an option to record geometry in several vocabularies, such as the Location Core Vocabulary (part of the Core Vocabularies Specification, created as part of the European ISA programme). The Location Core Vocabulary (Perego et al., 2013) also allows WKT.

An alternative, more simple way of indicating a location in RDF is by referencing an already existing named place resource. DBPedia and GeoNames are existing RDF datasets with spatial features, i.e. besides place names and a lot of other information, a set of coordinates is also available for the objects in these datasets. The advantage of referring to these named place resources is that it makes clear that different resources which refer to, for example, http://dbpedia.org/page/Utrecht, are all referring to the same city. If these resources would not use a URI reference but a literal value "Utrecht" this could mean the province Utrecht, the city Utrecht (both places in the Netherlands), the South African town called Utrecht, or maybe something else entirely. In DBPedia and GeoNames (in the free service) however, spatial features are limited to the lat/long coordinates of a point, encoded using W3C Basic Geo. This means it is only possible to refer to a point on a map, which could be the location of a small object such as a traffic light, but also the centre point of a city or a country. These datasets do not contain, for example, the exact boundaries of a country, as this would be represented by a polygon feature instead of a single point.

Another ongoing effort to create spatial linked data sets is LinkedGeoData, an RDF representation of OpenStreetMap data (Stadler et al., 2012). This dataset contains nodes, ways and relations, representing all kinds of spatial features, such as roads or boundaries. The dataset is interlinked with DBpedia and GeoNames. Part of the effort is to establish a lightweight ontology based on OpenStreetMap, and an OWL vocabulary for the exchange and reuse of geographic data. LinkedGeoData uses points, lines and polygons to represent spatial features. For every way (such as a road), there exists a triple that contains the positions of all its nodes as a sequence. Each node is a point, represented as an OGC WKT literal. In addition the geometries of nodes are encoded as Basic Geo point geometries.

Since the IMRO test data uses the Dutch Rijksdriehoeksstelsel (RD) coordinate reference system and not WGS84, and uses both point and polygon geometry types, the best option is to use the GeoSPARQL vocabulary. For the experiments, WKT was selected because there was no need to go beyond simple feature geometry types and WKT is more compact than the GML serialization.

5.4.3 URI strategy

An important question when creating RDF data is which criteria the URI identifiers which are formulated should follow and how to meet these criteria. In the Netherlands, we proposed a URI strategy describing guidelines for formulating URIs (Overbeek and van den Brink, 2013; Overbeek and Brentjens,

2013). The main criterion for linked data URIs is persistence: the URIs must still work even when things have changed, e.g. after the organization that minted them ceases to exist. For that reason a neutral domain name, with no organization name in it, is preferred. Other criteria to take into account in a URI strategy are scalability, intelligibility, trust, machine-readability and human-readability.

The Dutch URI strategy describes the following URI pattern based on these criteria:

```
http://{domain}/{type}/{concept}/{reference}
{domain} = {internet domain}/{path}
```

This pattern is adopted from ISA (2012b). The domain serves two purposes. It is first of all an important instrument in obtaining unique identifications: two objects that are administered in two different databases can coincidentally be designated with the same identification. Should both objects be published as linked data, then two unique URIs will still be generated: both will start with a different internet domain name.

Secondly, a well-chosen domain will ensure recognisability and trust. Plots in the Land registry with a URI such as http://data.brk.nl/perceel/-010101 (BRK stands for the Dutch base registry for cadastral information) seem more reliable than http://data.findithere.com/perceel/010101, for example.

The 'persistence' criterion leads to two recommendations: the domain should be reserved exclusively for the publication of the register, not for other publications as well as this may result in a re-organisation of the publications now and then; and its name should not include the name of an organization as these tend to change over time.

The path can be used if various collections of objects exist within a register, in which double identifiers (IDs) may be present. The path can then be used to create extra namespaces. It is recommended to use this with restraint.

The type indicates which kind of URI is involved. This may be either 'id', 'doc', or 'def'. The first two are used to identify resources and information about these resources respectively, and the latter is used for ontology terms.

The concept part serves mainly to give the human reader an indication of the type of concept that is identified by the URI. In addition it offers a solution if there are objects within the registration that have no unique identifiers, but that are unique per type of object. One should also consider persistence when choosing the concept.If it is conceivable in a registration

that object types (classes) can change names, while continuing to represent the same class, then it is not wise to include this component in the URI. A more abstract class should be used in such cases.

The reference is the identifying name or code of the individual object. This may be an identifying number, alphanumerical code, a word or name, etcetera.

In the case of IMRO, these recommendations on URI creation were applied in the following manner.

```
{domain} = http://data.ruimtelijkeplannen.nl
```

ruimtelijkeplannen.nl is the location on the web where Dutch spatial plans are publicly available. This is an appropriate, well-known and trustworthy domain name and not tied to a specific organization. Note that although we used this domain to assign URIs to our resources, no linked data is actually there in practice at this point in time. The optional path is not used.

```
{type} = 'id'
```

Only URIs of type 'id' are created. When information on the resource is supplied at the URI at a later stage, the type 'doc' will be used when supplying this information through content negotiation.

```
{concept} = 'ruimtelijkplan'
```

'ruimtelijkplan' (English: spatial plan) is a generic, conceptual name for the different types of plans that are described in the data. As the IMRO data model is subject to change every few years, and names of classes may change, we did not opt to use the class names of resources directly in the URI in the interest of persistence.

```
{reference} = [numeric id from the source data]
```

For the last part of the URI the numeric id of objects is taken from the source data. If a numeric id is not present, as is the case for nested, anonymous objects, an id is generated.

URIs for an IMRO vocabulary are also minted. These are created from the following pattern:

```
{domain} = http://data.ruimtelijkeplannen.nl
{type} = 'def'
```

This is followed by a hash '#' and the name of the vocabulary term. For the vocabulary, hash URIs are used to make the vocabulary recognisable as not part of the content; also the vocabulary is relatively small and can be requested as a whole in this way.

5.4.4 Experimental Transformation Implementation

Earlier in this article, we described GML in terms of its inherent triple structure. Because GML and RDF both have a triple structure, it is easy to define a transformation for translating any correctly structured (that is, conformant to the object-property triple structure) GML data to RDF automatically. As an experiment, we implemented such a transformation using XSLT 2.0. In this Generic-GML2RDF script, well-known GML content elements such as names and descriptions are mapped to their RDF equivalent. Objects, including nested features, data types and properties are recognized based on their place in the triple structure and are transformed accordingly.

The experimental implementation has 10 templates; counting whitespace and comments it has 98 lines. This shows the simplicity of the transformation. However, the stylesheet is presented here as a proof-of-concept and would need to be extended for general applicability. The stylesheet was tested on a sample GML file containing land use plans, conforming to the Dutch standard IMRO.

The transformation starts at the top of the GML file and selects all features, even the ones that are nested as property value of another feature. The features can be recognized because they always have an even number of ancestors (levels in the XML hierarchy). The GML file starts with a feature (usually a feature collection), which has properties, which in turn have features as values. A simple filter can take advantage of this fact. Those elements that have an even number of ancestors (levels in the XML hierarchy) are transformed to rdf:Description elements. The rdf:about attribute is filled with gml:id if available; if not, an id is generated.

Well-known, standardized GML properties are transformed to an appropriate property. When possible a standard property from RDF or RDFS is used. For example, gml:description is transformed to rdfs:comment, gml:name to rdfs:label. Properties that are not known (i.e. not standard GML, but from the domain-specific IMRO extension) are not changed, but receive the

```
<xsl:template match="/">
     <rdf:RDF>
       <xsl:apply-templates select="//*[count(ancestor::*) mod 2 = 0]
           [not(ancestor::schema-element(gml:_Geometry))]"/>
     </rdf:RDF>
</xsl:template>

<xsl:template match="*[count(ancestor::*) mod 2 = 0]
               [not(self::schema-element(gml:_Geometry))]">
     <rdf:Description rdf:about="{$uridomain}/id/{local-name()}/
               {if (@gml:id) then @gml:id else generate-id(.)}"
         rdf-type="{$uridomain}/def#{local-name()}">
         <xsl:apply-templates/>
     </rdf:Description>
 </xsl:template>

<xsl:template match="gml:Point">
    <xsl:element name="sf:{local-name()}" rdf:about="{$uridomain}/id/{local-name()}/
               {if (@gml:id) then @gml:id else generate-id(.)}">
         <geo:asWKT rdf:datatype="http://www.opengis.net/ont/geosparql#wktLiteral">
             &lt;<xsl:value-of select="@srsName"/>>
             Point(<xsl:value-of select="gml:pos"/>)
         </geo:asWKT>
     </xsl:element>
 </xsl:template>

 <xsl:template match="gml:description">
     <rdfs:comment><xsl:value-of select="text()"/></rdfs:comment>
 </xsl:template>

 <xsl:template match="gml:name"><rdfs:label><xsl:value-of select="text()"/></rdfs:label>
 </xsl:template>
```

Figure 5.2: Sample XSLT Fragment

same name in the RDF output.

In the IMRO data properties sometimes have nested content: they have an object as value, which is not referenced, but included in the hierarchical XML structure as a child element representing the object which in turn has child elements representing the properties. This construct receives special treatment in the stylesheet. The nested object is recognized by its even number of ancestors, and transformed to an rdf:Description. The property with nested content is transformed to a property that references the object that was nested, using an rdf:resource attribute containing the identifier of the feature prefixed with a hash '#'. Usually an identifier is not present in these cases, and one is generated automatically.

Properties that reference another feature in the GML data are transformed to an RDF ObjectProperty with an rdf:resource attribute containing the URI of the referenced feature.

The IMRO sample file contains simple Point and Surface geometries. These are transformed to a WKT serialization conform GeoSPARQL. The stylesheet currently only implements transformation of these two geometry types, but can easily be extended.

The XSLT stylesheet described above transforms GML data to RDF in a generic way, based on GML's object-property structure. But it ignores any domain-specific semantics the GML data may have. The IMRO sample file has a lot of domain-specific semantics, defined in the IMRO GML application schema. As shown in Figure 5.3, imro:Bouwaanduiding (an element which gives rules about the external appearance of buildings in a planning area) is translated to an rdf:Description of rdf:type http://data.ruimtelijkeplannen.nl/def#Bouwaanduiding. All properties of imro:Bouwaanduiding are transformed to RDF properties of the same name (see Figure 5.3). These should all be defined in the IMRO ontology, which does not exist, at least not in RDF/OWL at this stage. Some of the properties could be mapped to Linked Data vocabularies. For example, it would be appropriate to translate imro:naam to rdfs:label, but this is not known within the transformation, as it is a generic tool and is not aware of the meaning of the IMRO vocabulary.

This aspect must be addressed, because usually GML is extended for a certain domain. It contains rich semantics, which would be lost in the translation to RDF. However, these semantics are of crucial importance in the context of the Semantic Web. Semantic extensions of GML are usually described in a standardized way in a well-documented UML model, from which the so-called GML application schema is automatically derived. For the Dutch IMRO standard such a UML model and application schema are available. Therefore not only the GML data, but also either the UML model or the GML application schema should be translated to Linked Data stan-

```
<rdf:Description
    rdf:about="http://data.ruimtelijkeplannen.nl/id/Bouwaanduiding/NL.IMRO.0268.ID101733-00"
    rdf-type="http://data.ruimtelijkeplannen.nl/def#Bouwaanduiding">
     <imro:identificatie>NL.IMRO.0268.ID101733-00</imro:identificatie>
     <imro:typePlanobject>bouwaanduiding</imro:typePlanobject>
     <imro:plangebied rdf:resource="#NL.IMRO.0268.BP5000-VG01"/>
     <imro:naam>onderdoorgang</imro:naam>
     <imro:labelInfo rdf:resource="http://data.ruimtelijkeplannen.nl/id/Label/d34e17"/>
     <imro:geometrie rdf:resource="http://data.ruimtelijkeplannen.nl/id/Surface/d34e33"/>
  </rdf:Description>
  ...
  <sf:Surface rdf:about="http://data.ruimtelijkeplannen.nl/id/Surface/d34e33">
     <geo:asWKT rdf:datatype="http://www.opengis.net/ont/geosparql#wktLiteral">
               &lt;urn:ogc:def:crs:EPSG::28992&gt;
               Surface((187894.954 428671.17,187894.637 428671.314,187893.061 428672.029,
                        187889.716 428673.547,187886.166 428665.573,187891.442
                        428663.115,187894.954 428671.17))
     </geo:asWKT>
  </sf:Surface>
```

Figure 5.3: Sample IMRO RDF Fragment with WKT Geometry

dards. In addition, domain-specific knowledge about the application schema could improve the mapping, taking into account established Linked Data languages and vocabularies like RDF and RDFS, as well as Dublin Core, Simple Knowledge Organization System (SKOS), or other vocabularies. In our experiments we looked at this, and the next section describes some interesting aspects regarding the translation to RDF of specific semantics from an application-specific GML structure like IMRO.

5.4.5 Creating Meaningful RDF from Geo-Information Models

Meaning in the Semantic Web comes from vocabularies. However, the method described in the previous section does not provide or use a vocabulary. The mapping from GML to RDF would become more useful when it is accompanied by a OWL ontology for IMRO. Such an ontology can be automatically derived from the IMRO information model. The IMRO information model is available as UML diagram which, in current practice, is automatically converted into a GML application schema which supplies the rules for information exchange. This gives us two options to generate an OWL ontology: either from the UML model directly, or from the corresponding GML application schema. The first option has the advantage that UML is more mainstream information technology (IT) than GML application schemata and that a well-defined mapping from UML to OWL is defined by the OMG (OMG, 2009). The second option (mapping from the GML application schema) has the advantage that it is spatially aware (since a GML application schema has well defined spatial semantics) which would result in a better mapping for spatial objects. A combination of both would be best; this can be achieved by defining specific mappings from UML for spatial modelling constructs. Currently these mappings are partially stable: for spatial datatypes a mapping is described in GeoSPARQL (Perry and Herring, 2010). How to map UML stereotypes used in spatial models (such as <<FeatureType>>) is still under development. Rules for transforming UML models into OWL ontologies, and rules for developing Application Schemas directly in OWL are currently being developed within ISO/TC 211 project 19150-2 (ISO, 2013). This work will be published as an ISO Standard, and is currently in Draft International Standard (DIS) status. Experimentation with this mapping is being undertaken, for example in Cox (2013).

5.4.5.1 UML to OWL mapping issues

There are several issues with the mapping as described in the ISO mapping that should be discussed.

The first issue, also described in Cox (2013), is related to the open-world assumption of OWL as opposed to the closed-world assumption of UML. ISO 19150-2 states that "OWL ontologies are complementary to UML static views and serve different purposes." (p. vii) The OWL representation of UML models as described in ISO 19150-2 leans towards a more closed model: properties belong to one specific class, and any property that is not defined as belonging to a class, cannot be used with that class. Basically under the closed world assumption, anything that is not explicitly stated is not true within the closed system. By contrast, a popular saying to describe the open world assumption is that anyone can say anything about anything. Properties are independent things that can be used to make a statement about a thing belonging to any class, and anyone can do this. Under an open world assumption, anything that is not explicitly stated is unknown.

Two options are available: either the OWL representation should as closely as possible reproduce the frame-based closed UML model from the standard, or it should embrace the open-world assumption of OWL. In the latter case the OWL model would be a representation of the underlying model, but without exactly reproducing the restrictions present in the UML model. This affects property scoping and object property restrictions. The latter choice requires more interpretation during conversion.

The second issue is that the mapping rules convert the UML diagram to an OWL ontology without linking the terms in this vocabulary to terms in existing OWL vocabularies. This may be acceptable in the closed world context of ISO standards, but in OWL interoperability is achieved via the re-use of vocabularies. Janowicz et al. (2014) give vocabularies that reuse existing vocabularies 3 out of 5 stars in their proposed five star model for linked data vocabulary use (Janowicz et al., 2014).

The third issue is that some modelling conventions in UML (like the avoidance of multiple inheritance) can result in awkward UML constructions. When such a UML model gets mapped to OWL the awkwardness of the UML modelling is preserved, even in cases where these modelling conventions do not make any sense in the OWL domain and could be expressed in a better way in OWL, representing the real world more closely than is possible in UML. This, however, is probably not solvable in an automated way.

5.4.5.2 Experimentation

In this experiment we address the second issue, while also gaining more insight in the first and third issues. We propose to improve the mapping from UML to OWL by providing the missing information that is needed to create a better mapping. This is done by using tagged values in the UML model as mapping annotations which link terms from the IMRO model to terms from existing, well-known vocabularies. The transformation thus becomes a semi-automated process.

We experimented with these extended mapping rules by implementing them in an existing automated mapping tool: ShapeChange. ShapeChange implements an experimental, earlier set of UML to OWL mapping rules, described in Hobona and Brackin (2011). Feature types are mapped to owl:Class, their properties to owl:DatatypeProperty (for properties with literal values) or owl:ObjectProperty (for relations). Property names in the ontology are composed of the name of the UML class they belong to and their own name. Furthermore, rdfs:domain and rdfs:range are used to associate properties to their class. Codelists and enumerations are mapped to SKOS representations of the codelist.

In the first step of our experimentation we left the original mapping to OWL intact. The automated UML to OWL transformation was successfully applied to the IMRO model resulting in a IMRO OWL vocabulary. By slightly adapting the Generic-GML2RDF script it is possible to generate IMRO RDF from the GML data that refers to the IMRO vocabulary.

However, the IMRO vocabulary which was automatically generated by ShapeChange in this first step, is not harmonized with existing RDF vocabularies. For example, it would be meaningful to map imro:naam (name) to rdfs:label. However, the knowledge that imro:naam in fact has the same semantics as rdfs:label is not available in the UML model and cannot be automatically mapped. In order to improve the UML model for better mapping to RDF we extend the UML model by annotating the UML attributes that have a matching, standardized meaning in some well-known vocabulary in RDF, with a link to their RDF counterpart. Classes can be annotated in the same way. The annotation is recorded in UML via a tagged value. For example, the imro:naam attribute in the UML model is given the following annotation: 'owl:equivalentProperty=rdfs:label'. The enrichment of the UML model with the mapping of IMRO classes and properties to existing OWL classes and properties with equivalentProperty and equivalentClass statements solves the second issue of the provided ISO mapping.

We extended ShapeChange with a few lines of code to transform this annotation to a triple stating that imro:naam is an equivalent property to

```
<rdf:Description
  rdf:about="http://data.ruimtelijkeplannen.nl/id/Bouwaanduiding/NL.IMRO.0268.ID101733-00"
  rdf-type="http://data.ruimtelijkeplannen.nl/def#Bouwaanduiding">
     <imro:identificatie>NL.IMRO.0268.ID101733-00</imro:identificatie>
     <imro:typePlanobject>bouwaanduiding</imro:typePlanobject>
     <imro:plangebied rdf:resource="#NL.IMRO.0268.BP5000-VG01"/>
     <imro:naam>onderdoorgang</imro:naam>
     <imro:labelInfo rdf:resource="http://data.ruimtelijkeplannen.nl/id/Label/d34e17"/>
     <imro:geometrie rdf:resource="http://data.ruimtelijkeplannen.nl/id/Surface/d34e33"/>
</rdf:Description>
```

Figure 5.4: IMRO Bouwaanduiding with annotation in UML

```
<DatatypeProperty xmlns="http://www.w3.org/2002/07/owl#" rdf:about="http://www.geonovum.nl/imro2008#Bouwaanduiding.naam">
  <owl:equivalentProperty rdf:resource="rdfs:label"/>
  <domain xmlns="http://www.w3.org/2000/01/rdf-schema#" rdf:resource="http://www.geonovum.nl/imro2008#Bouwaanduiding"/>
  <range xmlns="http://www.w3.org/2000/01/rdf-schema#" rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <definition xmlns="http://www.w3.org/2004/02/skos/core#" rdf:resource="http://www.geonovum.nl/imro2008#Bouwaanduidingen"/>
</DatatypeProperty>
```

Figure 5.5: IMRO Bouwaanduiding vocabulary entry as generated by ShapeChange

rdfs:label.

This is a very simple example, but the implication is that a national standard such as IMRO, which in the Dutch context has very well-defined and agreed upon semantics, can now be mapped to international standards and thus be understood in an international context. If, for example, the INSPIRE data specifications were derived as OWL ontology from their UML source, national standards could easily be related to INSPIRE, thus becoming part of a big, international web of meaning.

5.4.5.3 Analysis of Results

The resulting IMRO OWL model is improved with links to terms from existing, well-known vocabularies. In this respect, the experiment is a success. It requires manual annotation of the UML model, based on which the UML to OWL transformation can be automated. Thus, the experiment provides a semi-automated method for UML to OWL transformation. However, the other issues with UML to OWL mapping may prevent the semi-automatic creation of a useful OWL model from UML.

ShapeChange produces a closed-world oriented representation of the UML in OWL: by this we mean it reproduces as closely as possible the frame-based, closed UML model. We prefer the open-world oriented representation described by Cox (2013), but nevertheless used ShapeChange in this experiment because it provides an automated approach which we could extend to support mapping annotations in the UML model. To be able to create an open-world oriented ontology as representation of a UML model, an open-world mapping should be defined. We suggest that ShapeChange can be modified to leave out the rdfs:domain and rdfs:range statements; these are logical axioms that 'close' the knowledge model. In addition, to make properties independent of classes, the class name should not be part of the property name (contrary to the rule in ISO 19150-2 par. 6.2.6).

In IMRO, properties that have the same name but belong to different UML classes can be treated as the same, as in each case the semantics of the property are the same. This was ascertained on a case-by-case basis. With an adapted mapping rule, which does not add domain and range to

properties and does not use the class name in the property name, multiple properties with the same name would be generated in the UML to OWL transformation. The duplicates must either be removed by hand or with some automated method.

In other models, properties with the same name that belong to different UML classes may have class-dependent semantics, in which case they cannot be considered the same. For these cases, the proper mapping to a class-specific property in OWL, i.e. with the domain specified and a class-specific property name, can be configured in a UML annotation, which indicates that the property is class-specific.

Finally, whether a closed-world or open-world mapping is most applicable for a specific UML model can be decided when performing the mapping. The choice of open-world or closed-world mapping could be recorded on the package level or the class level by, for example, adding a tagged value 'rdfMappingRule= open'.

It is unlikely that the third issue, UML anomalies getting mapped to RDF instead of being fixed, can be fixed by extending the UML model. IMRO is not a complex model and UML anomalies were not encountered during the experiment. To examine this in other cases, the generated OWL model would have to be analysed, providing conclusions and recommendations for each case.

An example of a UML anomaly is the avoidance of multiple inheritance. In UML models multiple inheritance is allowed, but often viewed as unwanted and therefore avoided. A result of this is often that classes have a number of properties that are exactly the same (often hard-copied from one class to the next). When this UML model gets mapped to OWL the awkwardness of the UML modelling is preserved (i.e. duplicate properties are created while there could have been just one). The adapted mapping rule, which maps properties with the same name in UML to just one property in OWL, would solve this.

Another example are the UML modelling constructs aggregation and composition citepkiko2008detailed, for modelling mereological relationships. A similar construct is not available in OWL. In addition, a construction like the UML association class is not directly available in OWL. Kiko and Atkinson (2008) suggest that a common solution for representing n-ary and association class relations in OWL is reification, i.e., the creation of an individual, which stands for an instance of the relation and relates the things that are involved in that instance of the relation. Aggregation, composition, and association classes were not encountered in the IMRO model.

More experimentation with automatically creating an open world oriented OWL model from UML, as well as with more complex UML models

containing modelling artefacts such as composition and association classes, is needed. Our experiment results in a working semi-automated method for mapping terms from UML models to existing OWL vocabularies, but does not address the other issues with UML to OWL transformation. This should be addressed in future work.

5.4.6 Source code availability

The IMRO2012 UML model, sample GML file containing IMRO spatial plans, result XML/RDF, genericGML2RDF XSLT, and adapted ShapeChange version and configuration file are downloadable from: `http://www.pilod.nl/-wiki/BrinkEtAl-GML2RDF-files`. The downloadable archive contains the source UML and GML files used in this experiment, the generic-GML2-RDF XSLT stylesheet, and the adapted ShapeChange source and configuration file.

5.5 Conclusions and future work

Linked data is complementary to GML published data. GML is advantageous for a controlled publication of data according to predefined semantic data specifications, while linked data facilitates an open publication environment in which additional information and data from other sources can easily be added. In this paper we demonstrated that it is possible to transform any GML to RDF using generic transformation rules, and without knowing the data model. OWL ontologies can be automatically derived from the original data specifications in UML. Once the data specification is available as an OWL ontology, it can be augmented with semantic relations to classes and properties from other ontologies, thus contributing to cross domain semantic harmonization. We introduced and implemented a method to annotate the data model in UML for this purpose. Deriving an OWL ontology from a UML model thus becomes a semi-automated process: manual annotation of the UML model with semantic annotations, followed by automatic transformation to OWL.

The implication of these results is that existing geo data can now be published on a large scale as semantically rich linked data. Linked data publication can be integrated in a SOA/GML based architecture, thus taking advantage of a large amount of geospatial data already being made available in national geoportals or, for example, in the context of INSPIRE. The IMRO information model and other Dutch information models could be derived as OWL ontologies, enriched with semantic relations to other ontologies from

the Semantic Web, using the annotation method we described. It would also be beneficial to annotate the INSPIRE UML data models with mappings to existing, relevant ontologies and then to derive an INSPIRE ontology as well. The ontologies can then be interlinked and better harmonized, and INSPIRE data can be made available as linked data. In the Netherlands, the next step is to extend our existing SDI with the option to create linked data. The web of linked data can then take advantage of the existing wealth of standardized, open geospatial data.

Future work includes the implementation of ISO 19150-2 in an automated mapping tool such as ShapeChange and the addition of an option for creating open world-oriented ontologies. The problem of modelling conventions and restrictions in UML which lead to awkward modelling constructs, and could be better modelled differently in OWL, must still be addressed. Whether this is possible in an automated way as well remains an open question and seems, in some cases, unlikely.

Bibliography

Sonya Abbas and Adegboyega Ojo. Applying Design Patterns in URI Strategies–Naming in Linked Geospatial Data Infrastructure. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 2094–2103. IEEE, 2014.

S. Athanasiou, L. Bezati, G. Giannopoulos, K. Patoumpas, and D. Skoutas. GeoKnow: Making the web an exploratory place for geospatial knowledge. *Market and Research Overview.*, 2013.

T. Berners-Lee, Roy Fielding, and L. Masinter. Uniform resource identifier (URI): Generic syntax. Available online: http://www.ietf.org/rfc/rfc3986.txt (accessed 3 June 2014, 2005.

Harold Boley, Gary Hallmark, Michael Kifer, Adrian Paschke, Axel Polleres, and Dave Reynolds. RIF core dialect. Available online: http://www.w3.org/TR/rif-core/ (accessed 11 February 2014), 2010.

Dan Brickley. Basic Geo (WGS84 lat/long) Vocabulary. Available online: http://www.w3.org/2003/01/geo/ (accessed 11 February 2014), 2003.

CEN. TR 15449-4:2012 Geographic information – Spatial Data Infrastructures – Part 4: Service Centric View., 2012.

G. Ciardi, A. Abrescia, and S. Pezzi. Linked open data from OGC compliant web services: the case of Regione Emilia-Romagna GeoPortal. In *INSPIRE conference*, 2013.

Simon Cox. An explicit OWL representation of ISO/OGC Observations and Measurements. In *Proceedings of the 6th International Workshop on Semantic Sensor Networks*, pages 1–18, 2013.

Paul Davidson. Designing URI sets for the UK public sector. *UK Chief Technology Officer Council*, 2009.

Geonovum. Informatiemodel Ruimtelijke Ordening 2012. Available online: http://www.geonovum.nl/wegwijzer/standaarden/informatiemodel-ruimtelijke-ordening-imro2012 (accessed 3 June 2014, 2012.

Ontology Engineer Group. geometr2rdf. Available online: http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/technologies/151-geometry2rdf (accessed 3 June 2014), 2014.

Gobe Hobona and Roger Brackin. OGC® OWS-8 Cross Community Interoperability (CCI) Semantic Mediation Engineering Report. Available online: https://portal.opengeospatial.org/files/?artifact%5Fid=46342 (accessed 11 February 2014), 2011.

INSPIRE. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Available online: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002 (accessed 2-February-2016), 2007.

ISA. ISA Programme Core Vocabularies Specification. Available online: https://joinup.ec.europa.eu/system/files/project/Core%5FVocabularies-Business%5FLocation%5FPerson-Specification-v1.00%5F0.pdf (accessed 11 February 2014), 2012a.

ISA. D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC. Available online: https://joinup.ec.europa.eu/sites/default/files/D7.1.3 - Study on persistent URIs%5F0.pdf (accessed 11 February 2014), 2012b.

ISO. ISO 19107:2003 Geographic information – Spatial schema., 2003.

ISO. ISO 10109:2015 Geographic information – Rules for application schema., 2005.

ISO. ISO/CD 19150-2 (2013). Geographic information - Ontology - Part 2: Rules for developing ontologies in the Web Ontology Language (OWL). (Draft International Standard), 2013.

ISO. 19125-1:2004. Geographic information - Simple feature access - Part 1: Common architecture., 2014.

ISO/IEC. ISO/IEC 13249-3 Information technology - Database languages - SQL multimedia and application packages - Part 3: Spatial., 2011.

Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and II Vardeman. Five stars of linked data vocabulary use. *Semantic Web Journal*, 5(3):173–176, 2014.

M. Kay. XSL Transformations (XSLT) version 2.0. Available online: http://www.w3.org/TR/xslt20/ (accessed 11 February 2014), 2007.

Kilian Kiko and Colin Atkinson. A detailed comparison of UML and OWL. *Reihe Informatik*, TR-2008(4), 2008.

Graham Klyne and Jeremy J Carroll. Resource description framework (RDF): Concepts and abstract syntax, version 2004-02-10. Available online: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, 2004.

Ron Lake and Adrian Cuthbert. Geography markup language (GML) v1. 0. Available online: htp://portal.opengeospatial.org/files/?artifact%5Fid=7197 (accessed 11 February 2014), 2000.

LODG. Linking Open Government Data, URI Design Principles: Creating Persistent URIs for Government Linked Data. Available online: http://logd.tw.rpi.edu/instance-hub-uri-design (accessed 3 June 2014), 2013.

F.J. López-Pellicer, L.M. Vilches-Blázquez, F.J. Zarazaga-Soria, P.R. Muro-Medrano, and O. Corcho. The Delft Report: Linked Data and the challenges for geographic information standardization. *Revista Catalana de Geografía*, XVII(44), 2012.

Frank Manola and Eric Miller. RDF primer. Available online: http://www.w3.org/TR/2004/REC-rdf-primer-20040210/ (accessed 11 February 2014), 2004.

OGC. OGC® Geography Markup Language (GML) - Extended schemas and encoding rules, version 3.3.0. Available online: http://www.opengeospatial.org/standards/gml, 2012.

OMG. Ontology Definition Metamodel (version 1.0). Available online: http://www.omg.org/spec/ODM/1.0/ (accessed 11 February 2014), 2009.

H. Overbeek and T. Brentjens. Draft URI Strategy for the NL Public Sector. Available online: http://www.w3.org/2013/04/odw/odw13%5Fsubmission%5F14.pdf (accessed 3 June 2014), 2013.

H Overbeek and L van den Brink. Towards a national URI–Strategy for Linked Data of the Dutch public sector. Available online: http://www.pilod.nl/wiki/Bestand:D1-2013-09-19%5FTowards%5Fa%5FNL%5FURI%5FStrategy.pdf (accessed 3 June 2014), 2013.

A Perego, M Lutz, and P Archer. Programme Location Core Vocabulary. *ISA Programme Core Vocabularies Working Group*, 2013.

Matthew Perry and John Herring. OGC® GeoSPARQL-A geographic query language for RDF data. Available online: http://www.opengeospatial.org/standards/geosparql (accessed 2 February 2016), 2010.

GeoKnow Project. GeoKnow Project description. Available online: http://geoknow.eu/Project.html (accessed November 2013, 2013.

Sven Schade and Simon Cox. Linked Data in SDI or How GML is not about Trees. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science-Geospatial Thinking*, pages 1–10, 2010.

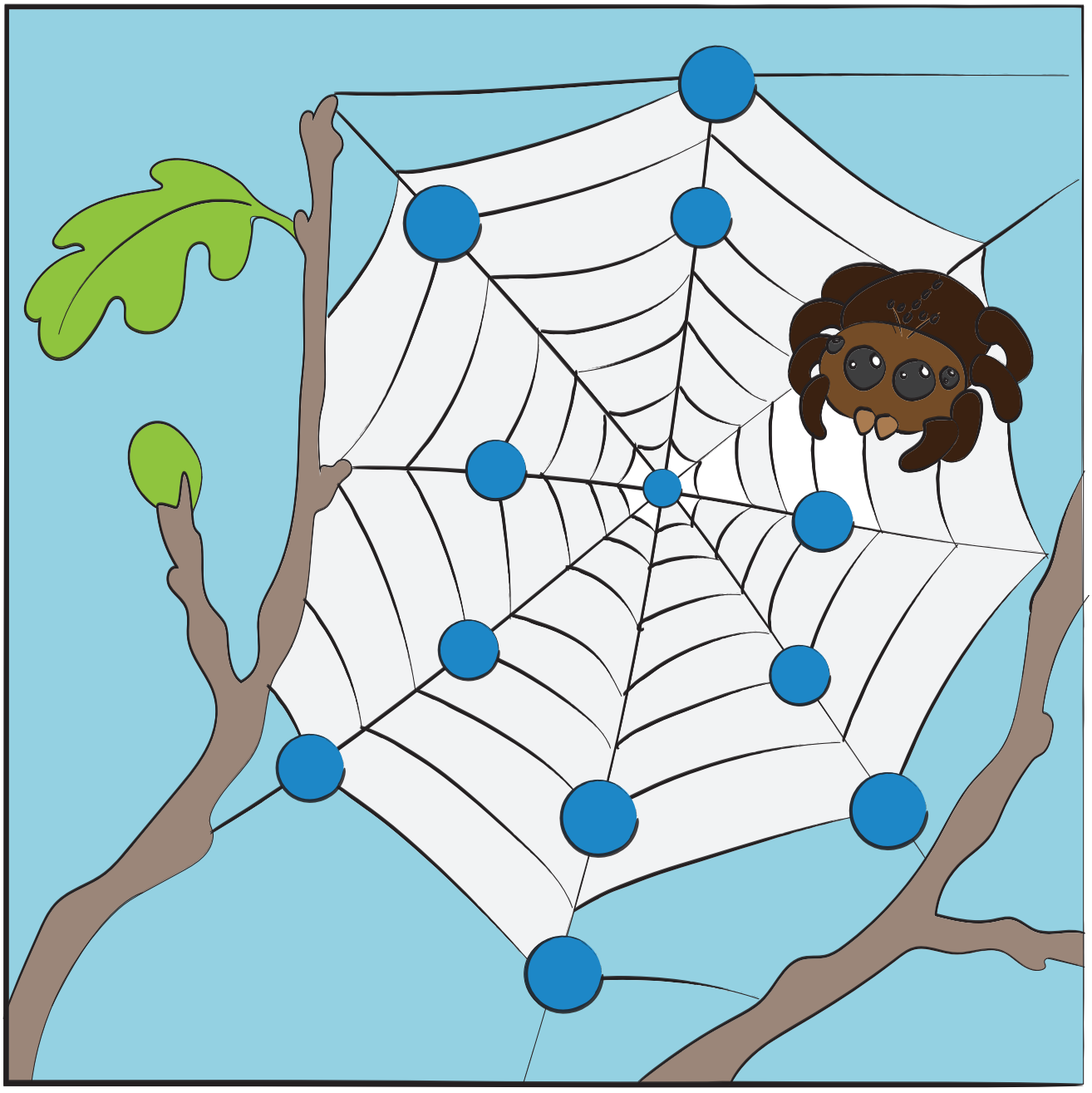
P Smits, D Arctur, R Atkinson, B Bargmeyer, L Boerboom, SF Browdy, EV Praag, G Hodge, S Jensen, TS Ulgen, and M. Wilson. Recommendations for the technical design of a global interoperable information network. *Eye on Earth Working Group 3 – Technical Infrastructure, White Paper 1 EOE/WG3/1*, 2011.

Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web Journal*, 3(4): 333–354, 2012.

Sven Tschirner, Ansgar Scherp, and Steffen Staab. Semantic access to INSPIRE. In *Proceedings of the Terra Cognita 2011 Workshop on Foundations, Technologies and Applications of the Geospatial Web In conjunction with the International Semantic Web Conference (ISWC2011)*, page 75, 2011.

W3C. OWL 2 Web Ontology Language Document Overview (Second Edition). Available online: http://www.w3.org/TR/2012/REC-owl2-overview-20121211/ (accessed 3 June 2014), 2012a.

W3C. 223 Best Practices URI Construction. Available online: http://www.w3.org/2011/gld/wiki/223%5FBest%5FPractices%5FURI%5FConstruction (accessed 3 June 2014), 2012b.



Web of Data

Chapter 6

Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web

Authors: Linda van den Brink, Payam Barnaghi, Jeremy Tandy, Ghislain Atemezing, Rob Atkinson, Byron Cochrane, Yasmin Fathy, Raúl García-Castro, Armin Haller, Andreas Harth, Krzysztof Janowicz, Sefki Kolozali, Bart van Leeuwen, Maxime Lefrançois, Josh Lieberman, Andrea Perego, Danh Le-Phuoc, Bill Roberts, Kerry Taylor, Raphaël Troncy. (*Semantic Web Journal*, pre-press, 2018)

This paper has been accepted for publication in a peer-reviewed scientific journal in 2018 and is published unchanged in this chapter. It explores the question: "How can we apply general Web based principles to improve the discoverability and accessibility of spatial data?". It summarises the efforts that have been undertaken in the joint W3C/OGC Working Group on Spatial Data on the Web, in particular the effort to describe a set of best practices, distilled from practical experiences, for publishing, discovery and retrieving spatial data on the Web. This paper presents those best practices and their rationale, the principles that guided their selection, as well as pointing out gaps in current practice, such as a single standardised vocabulary for geospatial data, and a standardised way to publish dynamic, large geospatial datasets on the web.

Contributions: the Spatial Data on the Web Best Practice was published as a W3C Note and an OGC Best Practice. Implementations of the guidelines are in progress in dissemination portals for geospatial data such as at the Dutch Geoportal and the data portal of the Dutch Cadastre, as well as in existing standards such as OGC Web Feature Service 3.0 (Portele and Vretanos, 2018).

text of published paper starts after this line

Abstract: Data owners are creating an ever richer set of information resources online, and these are being used for more and more applications. Spatial data on the Web is becoming ubiquitous and voluminous with the rapid growth of location-based services, spatial technologies, dynamic location-based data and services published by different organizations.

However, the heterogeneity and the peculiarities of spatial data, such as the use of different coordinate reference systems, make it difficult for data users, Web applications, and services to discover, interpret and use the information in the large and distributed system that is the Web. To make spatial data more effectively available, this paper summarizes the work of the joint W3C/OGC Working Group on Spatial Data on the Web that identifies 14 best practices for publishing spatial data on the Web.

The paper extends that work by presenting the identified challenges and rationale for selection of the recommended best practices, framed by the set of principles that guided the selection. It describes best practices that are employed to enable publishing, discovery and retrieving (querying) spatial data on the Web, and identifies some areas where a best practice has not yet emerged.

Keywords: Geographic information systems, Spatial data, Web technologies, World Wide Web, W3C, Open Geospatial Consortium, OGC

6.1 Introduction

Spatial data is important. Firstly, because it has become ubiquitous with the explosive growth in positioning technologies attached to mobile vehicles, portable devices, and autonomous systems. Secondly, because it is fundamentally useful for countless convenient consumer services like transport planning, or for solving the biggest global challenges like climate change adaptation (Taylor and Parsons, 2015). Historically, sourcing, managing and using high-quality spatial data has largely been the preserve of military, government and scientific enterprises. These groups have long recognized

the importance and value that can be obtained by sharing their own specialized data with others to achieve cross-theme interoperability, increased usability and better spatial awareness, but they have struggled to achieve the cross-community uptake they would like. Spatial Data Infrastructures (SDIs) (Masser, 1999), which commonly employ the mature representation and access standards of the Open Geospatial Consortium (OGC), are now well developed, but have become a part of the "deep Web" that is hidden for most Web search engines and human information-seekers. Even geospatial experts still do not know where to start looking for what they need or how to use it when they find it. The integration of spatial data from different sources offers possibilities to infer and gain new information; however, spatial data on the Web is published in various structures, formats and with different granularities. This makes publishing, discovering, retrieving, and interpreting the spatial data on the Web a challenging task.

By contrast, the linked data Web, as a platform of principles, tools, and standards championed by the World Wide Web Consortium (W3C) enables data discoverability and usability that is readily visible in, for example, search engine results for consumer shopping. The principles are based on proven aspects of the Web such as resolvable identifiers, common representation formats, and rich interlinking of independently-published information, but adds explicit vocabulary management and tooling that targets the huge Web developer community. Can these principles be successfully applied to the world of complex spatial data to achieve the desired usability and utility?

There are, already, many good examples of projects and Web services that deliver to these goals, such as spatial data publication platforms in the Netherlands,[1] Nanaimo City in Canada,[2] or the UK Environmental Agency,[3] but also popular spatial data collections such as Geonames.[4] However, spatial data custodians struggle to find the *best* way to publish their data in order to optimize the future impact as more data appears, more tools are developed, and the consumer community grows. Similarly, Web developers as data consumers and tool developers as foundation-stonemasons are demanding an expert consensus to guide their product development.

The W3C and OGC standardization bodies jointly convened a large workshop in London in 2014 where these issues were extensively discussed over two days[5]. As a result of the interest, enthusiasm and challenges identified

[1] `https://data.pdok.nl/datasets`

[2] `http://maps.nanaimo.ca/data/`

[3] `http://environment.data.gov.uk/bwq/profiles/`

[4] `http://www.geonames.org`

[5] Linking Geospatial Data (LGD'14). 5-6 March 2014, London. See: `https://www.w3.org/2014/03/lgd/`

Table 6.1: Standardized aspects of SDIs

Aspect	Description	Reference standards
Discoverability	Annotate resources with metadata	ISO 19115 (ISO/TC 211, 2003; ISO, 2014), ISO 19119 (ISO/TC 211, 2005a; 211, 2005b)
Accessibility	Web services for discovering, viewing, downloading, sharing geospatial raster data (coverages) etc.	OGC CSW (Nebert et al., 2016), OGC WMS (de la Beaujardiere, 2006), OGC WFS (Vretanos, 2010), OGC WCS (Baumann, 2012)
Portrayal	Defining rules for displaying spatial data	ISO 19117 (211, 2012), OGC SLD (Lupp, 2007), OGC SE (Müller, 2006), OGC KML (Burggraf, 2015)
Information modeling	Describing the contents of information resources, including geometry	ISO 19103 (ISO, 2015a), ISO 19107 (ISO, 2003), ISO 19109 (ISO, 2015b), ISO 19110 (ISO, 2005)
Data exchange	Defining formats for exchanging the data	OGC GML (Consortium et al., 2007)
Spatial reference systems	Specifying the location on Earth of geographical information	ISO 19111 (211, 2005a, 2007)

there, they proceeded to establish a joint working group to develop, amongst other things, a compendium of best practices for spatial data on the Web, published in September 2017 (Tandy et al., 2017). The working group completed its work in 2017, but was succeeded by an interest group[6] in which the work is continued.

This paper is a companion publication to the Spatial Data on the Web Best Practices Note that was endorsed by the working group, including the authors of this paper. It summarizes the work and describes the best practices themselves, but additionally presents the principles that guided the selection of these best practices, and the rationale underlying particular selections. It also identifies some areas where a best practice seems to be needed but has not yet emerged.

6.1.1 Background: spatial data, the Web, and semantics

Any data (or metadata) that has a location component can be viewed as spatial data: its spatial nature means certain operations such as proximity and containment functions have a meaning within the spatial domain. A location component is a reference to a place on Earth or within some other space (e.g., another planet, or a shopping mall) and can be many things: a physical object with a fixed location, such as a building or canal; an administrative unit, like a municipality or postal code area, or the trajectory of a moving object like a car. The power of spatial data is in the opportunity to combine and integrate information based on location.

While spatial data refers to any data that has a location component either on Earth or within some other space such as another planet (as defined earlier), geospatial data refers to any data that has a location in terms of geographic coordinates (e.g., GIS coordinates) within Earth. The focus of the Spatial Data on the Web Best Practices, and consequently of this paper, is on geospatial data. Although non-geospatial use cases were brought to the working group and were included in the Spatial Data on the Web Use Cases and Requirements (UCR) (Knibbe and Llaves, 2016), there were no active participants in the working group who had expertise with spatial data other than geospatial. The focus was therefore narrowed to geospatial data; requirements of non-geospatial data might be included in future work. This paper likewise deals almost exclusively with geospatial data. That said, many of the best practices are applicable to wider spatial data concerns. In the remainder of this paper, we simply refer to spatial data for brevity.

Spatial data can also have a temporal dimension. Temporal data varies over time. On the other hand, spatio-temporal data captures spatial data (i.e. current location) associated with the time the data is taken at that particular location (Fathy et al., 2017).

However, in order to use spatial-temporal data, it has to be available and accessible. Common practice is to publish this data using an SDI. SDIs are based on a service-oriented architecture (SOA), in which existing resources are documented using dataset-level metadata, published in catalogs which are the most important discovery and access mechanism (Nogueras-Iso et al., 2005). More detailed metadata describing structure and content of datasets, as well as service requests and payloads, are far less commonly shared, and whilst standards suitable for some aspects of this requirement are emerging, defining a common or best practice remains a challenge.

[6] `https://www.w3.org/2017/sdwig/charter.html`

The Open Geospatial Consortium (OGC), founded in 1994, publishes technical standards necessary for SDIs to work in an interoperable way. These standards are based on the more abstract standards for geospatial information from the ISO Technical Committee 211 on Geographic information / Geomatics (ISO/TC 211). Different aspects of the SDI are standardized (see Table 6.1).

The OGC developed the first standards for spatial data Web services as of 1998 and has responded to early architectural trends in the Web (e.g., SOA). While SDIs and related standards were developing, so did the World Wide Web. Web standards like HTML, XML and RDF (Cyganiak et al., 2014) were created in the nineties as well. While the Web started off as mostly documents with hyperlinks, over the years it evolved to much more sophisticated Web applications, including mass applications in which geospatial data was used, like Bing Maps, Google Maps, Google Earth and OpenStreetMap. More recently, XML has been replaced in many Web applications by more lightweight formats (e.g. JSON, and RDF).

As a generic model or framework, RDF can be used to publish geographic information. Its strengths include its structural flexibility, particularly suited for rich and varied forms for metadata required for different purposes. However, it has no specific features for encoding geometry, which is central to geographic information. Several vocabularies and extensions have been proposed for this purpose, including a core RDF/OWL vocabulary for geographic information which is part of OGC's GeoSPARQL (Perry and Herring, 2010b). Figure 6.1 provides an illustration of the main areas of overlap between the different communities of practice of the Spatial, Semantic and the "mainstream" Web—the last being characterized by large amounts of unstructured as well as structured data in various forms, and ad-hoc approaches to data publishing, driven by Web-centric skills and technologies, without explicit support for semantic or spatial aspects. Since the data is often published over HTTP in various formats and with different structure, XML formats are often used to integrate different data structures from different resources into the Web.

Spatial data can be published on the Web in the JSON format through a RESTful API or through a Spatial Data Infrastructure (SDI) based on a Service Oriented Architecture (SOA). RDF data is also widely used to describe and link resources on the Web. In the spatial data domain, for example, OGC's GeoSPARQL is an RDF-based representation that is used to query spatial data.

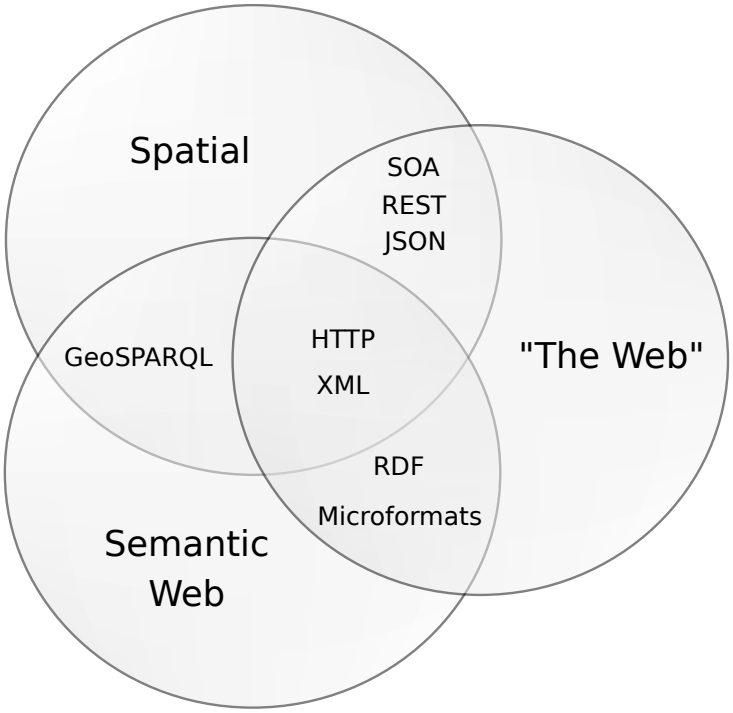


Figure 6.1: Commonalities between different communities of practice publishing data on the Web.

6.1.2 Contributions

With the prevalence of sensor and actuator devices and the increase in location-based services, the use of spatial technologies is rapidly growing. The existing geospatial data, online maps combined with new forms of dynamic location-based data and services, create an opportunity for various new applications and services. However, to make spatial data more effectively available across different domains, a set of common practices are required.

The Spatial Data on the Web (SDW) Working Group[7] has been committed to determining how spatial data can best be published, discovered, queried and integrated with other data on the Web. This paper is the result of considerable effort on identifying these best practices for publishing and integrating spatial data on the Web. These Spatial Data on the Web Best Practices were published as a W3C Note (Tandy et al., 2017) and as an OGC Best Practice.

This companion paper was written by members of the Spatial Data on the Web Working Group. It summarizes the key requirements for publishing,

[7] `http://www.w3.org/2015/spatial/`

retrieving and accessing spatial data on the Web to make it more interoperable, accessible, interpretable and understandable by humans and machines, accompanied by know-how and best practices addressing those requirements. The main contribution of the paper is its additional background information about the rationale underlying the selection of the best practices. In addition, the paper describes areas where best practices are still missing.

6.1.3 Paper organization

The rest of the paper is organized as follows: Section 6.2 explains a set of principles for setting out the scope of the problems that the best practices will address. Section 6.3 states the key requirements for publishing and sharing spatial data on the Web, presents the related best practices as identified by the working group, and discusses how the best practices address the described requirements. Section 6.4 discusses several gaps that still exist in current practice. Finally, Section 6.5 draws conclusions and discusses the future direction.

In the remainder of this paper, we refer to the Spatial Data on the Web Best Practices as "SDWBP".

6.2 Principles for describing best practices

As explained in the introduction, the aim of the work is to improve the discoverability, interoperability and accessibility of spatial data. The key principle follows from this: that through the adoption of the best practices identified in the SDWBP, the discoverability and linkability of spatial information published on the Web is improved.

A second principle concerns the intended audience of the spatial data in question. The aim is to deliver benefits to the broadest community of Web users possible—not to geospatial data experts only. The term 'user' signifies a data user: someone who uses data to build Web applications that provide information to end users—website visitors and app users—in some way. These data users are therefore among the intended audience of not only the spatial data, but also the Best Practice. The SDWBP provides value and guidance for application, website and tool builders to address the needs of the mass consumer market. Furthermore, the best practices should provide guidance for spatial data custodians. The best practices can offer a comprehensive set of guidelines for publishing spatial data on the Web.

A third principle is to have a broad focus. The first working draft of the best practices solicited several comments about a perceived "RDF bias".

While to develop spatial data following the 5-star Linked Data principles[8] was one of the goals at the start, the solutions described in the best practices for discoverability and linkability should also be applicable to other spatial data on the Web. The best practices promote a linked data approach, but without asserting a strong association between linked data and RDF. Linked data requires only that the formats used to publish data support Web linking (Jacob and Walsh, 2004). Furthermore, any ontologies developed within the working group should not be tightly coupled with upper ontologies ("compatible with" rather than "dependent upon"); this avoids data publishers having to commit to a given world view, as specified within a particular upper ontology, in order for them to use the best practices and any ontologies related to them.

A fourth principle follows from the term 'best practice' very directly: that its contents are taken from practice. The aim is not to reinvent or provide 'best theories;' in other words, the intention of the best practice is not to create new solutions where good solutions already exist or to invent solutions where they do not yet exist. The contents of the best practice should be made up of the best existing practices around publishing spatial data on the Web that can be found. Consequently, the aim is for each of the best practices in the document to be linked to at least one publicly available example(s) of a non-toy dataset that demonstrates the best practice.

Lastly, the best practices should comply with the principles of the W3C Best Practices for Publishing Linked Data (Hyland et al., 2014) and the W3C Data on the Web Best Practices (Farias Lóscio et al., 2017). Where they do not, this will be identified and explained. The Data on the Web Best Practices (referred to as DWBP henceforth) form a natural counterpart to the work on spatial data on the Web. The best practices are aligned with DWBP in the following ways: a) by using the same best practice template, and b) by referring to the DWBP instead of repeating it. While the focus of the best practices is on spatial data, they may include recommendations on matters that are not exclusively related to spatial data on the Web, but are considered by the Working Group to be essential considerations in some use cases for publishing and consuming spatial data on the Web, and not covered in enough detail in DWBP or other documents.

Table 6.2: The Spatial Data on the Web Best Practices (Tandy et al., 2017)—and where they are discussed in the paper

Best Practice 1	Use globally unique persistent HTTP URIs for spatial things	§6.3.3	Best Practice 2	Make your spatial data indexable by search engines	§6.3.9
Best Practice 3	Link resources together to create the Web of data	§6.3.3	Best Practice 4	Use spatial data encodings that match your target audience	§6.3.1, §6.3.6
Best Practice 5	Provide geometries on the Web in a usable way	§6.3.1, §6.3.10	Best Practice 6	Provide geometries at the right level of accuracy, precision, and size	§6.3.1, §6.3.5, §6.3.8
Best Practice 7	Choose coordinate reference systems to suit your user's applications	§6.3.2, §6.3.5, §6.3.7	Best Practice 8	State how coordinate values are encoded	§6.3.2
Best Practice 9	Describe relative positioning	§6.3.10	Best Practice 10	Use appropriate relation types to link spatial things	§6.3.1
Best Practice 11	Provide information on the changing nature of spatial things	§6.3.3, §6.3.6, §6.3.7	Best Practice 12	Expose spatial data through 'convenience APIs'	§6.3.3
Best Practice 13	Include spatial metadata in dataset metadata	§6.3.4	Best Practice 14	Describe the positional accuracy of spatial data	§6.3.5

6.3 The key requirements and best practices for publishing spatial data on the Web

The following sections discuss the main topics covered by SDWBP, and explains how they are addressed in the defined best practices. The topics are presented in a summarized fashion and are grouped thematically. Examples of datasets in which these best practices are implemented are not provided, since these are already present in the SDWBP. For the convenience of the reader, Table 6.2 outlines the best practices as they are stated in SDWBP, and indicates in which of the following sections they are discussed.

6.3.1 Geometries and spatial relationships

Tobler's first law of geography states, "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Like statistics, in addition to portraying information, spatial data provides a basis for reasoning, in this case based on location. Integration of data based on a location (i.e. relating things based on being located at or describing the same location) is often very useful. For this to be possible, either explicit geometries or topological relationships are necessary. Ideally, to enable their widest reuse, geometries should be described having in mind the geospatial, Linked Data and Web communities. This may not always be feasible,

[8] `http://5stardata.info`

but the objective should at least be to describe geometries (also) for Web consumption.

One of the key best practices (namely, Best Practice 5) is therefore about providing geometries on the Web in a usable way. A single best way of publishing geometries was not identified; what is 'best' is in this case primarily related to the specific use case and tool support, which determine the geometry format to be used, the coordinate reference system (CRS) (more details are included in Section 6.3.2), as well as the level of accuracy, precision, size and dimensionality of geometry data. Note that these aspects are interrelated: for instance, the dimensionality of a geometry constrains the CRSs that can be used, as well as the geometry encodings.

Best Practice 5 identifies three scenarios in which geometries can be used: specific geospatial applications, linked data applications, and Web consumption. The practice also offers guidelines for choosing the right vocabularies from several available ones for describing geometry in each scenario. Currently, there are two geometry formats widely used in the geospatial and Web communities, respectively, GML (Consortium et al., 2007) and GeoJSON (Butler et al., 2016). GML provides the ability to express any type of geometry, in any CRS, and up to 3 dimensions (from points to solids) but is typically serialized in XML. GeoJSON supports only one coordinate reference system (CRS84—i.e., WGS 84 longitude / latitude), and geometries up to 2 dimensions (points, lines, surfaces) but it is serialized in JSON, which is often easier for browser-based Web applications to process. In the Linked Data community, several specific vocabularies for RDF-based representations of geometries are available, such as GeoSPARQL (Perry and Herring, 2010b), W3C Basic Geo (Brickley, 2003), GeoRSS (Lieberman et al., 2007), or the ISA Core Location vocabulary (Perego and Lutz, 2015). Appendix A of SDWBP offers a comparison of the most common spatial ontologies.

Instead of catering for a single scenario, spatial data publishers should offer multiple geometric representations when possible, balancing the benefit of ease of use against the cost of the additional storage or additional processing if converting on-the-fly. This can be implemented using HTTP content-negotiation; however, this only works for media-type, character set, encoding and language. Consequently, it is not possible to select one representation that conforms to a given "profile" (e.g., data model, complexity level, CRS) from several that all share the same media-type.

Note that publishing geometries on the Web need not always be called for. Although spatial relationships can often be derived mathematically based on geometry, this can be computationally expensive. Topological relationships such as these can be asserted, thereby removing the need to do geometry-based calculations. Exposing such entity-level links to Web applications,

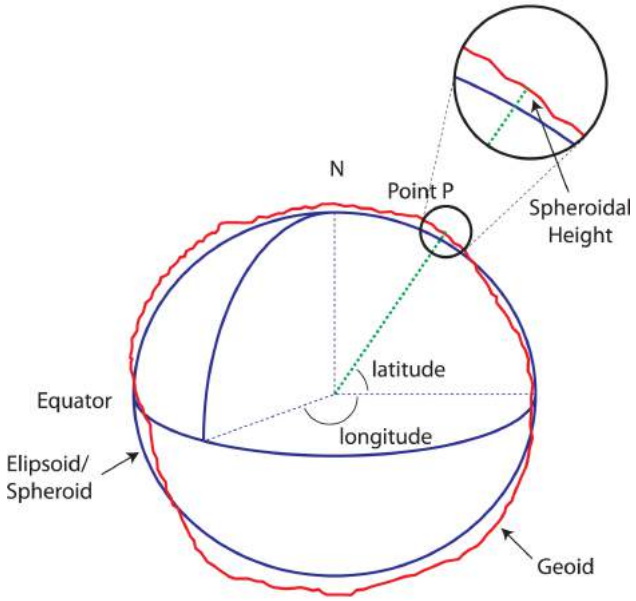


Figure 6.2: Elipsoid and spherical coordinates. ("Creative Commons Attribution 3.0 Australia" license from ICSM.gov.au—Intergovernmental Committee on Surveying and Mapping).

user-agents and Web crawlers allows the relationships between resources to be found without the data user needing to download the entire dataset for local analysis.

6.3.2 Coordinate reference systems and projections

The key to reasoning and sharing spatial information is the establishment of a common coordinate reference frame in which to position the data. For Earth-based data, this may be done in spherical coordinate values such as latitude, longitude and (optionally) elevation, or in a projected Cartesian coordinate space. The latter involves the flattening of a sphere in exchange for making it vastly easier to accurately measure area and distance. Regardless of the Coordinate Reference System (CRS) chosen, a distortion of the data will occur either in relative angles (positions), sizes (areas), or distances. Best Practice 7 identifies the World Geodetic System 1984 (WGS 84—EPSG:4326), which provides a good approximation at all locations on the Earth, as the most commonly used CRS for spatial data on the Web, but also explains when EPSG:4326 is not recommended, especially in use cases that require a higher level of accuracy than WGS 84 can offer.

In the Spatial Data on the Web Working Group, there was a lot of dis-

cussion on this topic. There are a lot of concerns with WGS 84 in the geospatial community. However, WGS 84 is dominant on the Web; it is used by most online mapping providers (e.g. Google, Bing, OpenStreetMap) and popular formats such as GeoJSON. Some explanation is required in order to understand the concerns of the geospatial community. All spatial coordinate frameworks begin with a mathematical model of the object being mapped. For the Earth, (which is not a true spheroid) an ellipsoid model, most fitting to the area being mapped, is commonly used (Figure 6.2). A Geodetic Datum is then placed on top of the reference ellipsoid to allow numerical expression of position. This may include a Vertical Datum, usually an approximation of sea level, to reference height and depth. WGS 84 (EPSG:6326) is an example of a Geodetic Datum based on the WGS 84 (EPSG:7030) Ellipsoid. It uses latitude and longitude coordinates (anchored to the equator and poles) to indicate position.

Geodetic CRSs are useful for collecting information within a common frame. The measurements they use are good for plotting directions but difficult to use when calculating area or distance. Latitude and longitude are angular measurements that do not convert easily to distance because their size in true units of measure (e.g., meters) varies according to location on the sphere. They become smaller as you near the poles.

In order to measure size and distance, and to display spatial information on a screen or paper, Projected Coordinate Reference Systems are used. A Projected CRS flattens a sphere to enable it to be portrayed and measured in 2 (or 3) dimensions. Figure 6.3 shows an example of the effects of this based on WGS 84 / Pseudo-Mercator (EPSG:3857), which is used commonly on the Web. As demonstrated by this figure, flattening a sphere introduces distortions. Imagine flattening the skin of an orange. You can preserve fairly accurately measurements for a portion of the skin but the rest will necessarily be distorted, either through stretching, compression or tears (see Figure 6.4). Projected CRSs are optimized to preserve distance, area or angular relations between spatial things for a chosen region. Distortion grows the further you move from that region. This is one reason why there are so many CRSs.

To summarize, WGS 84 is by far the most common CRS for spatial data on the Web. However, other CRSs exist for good reasons. Best Practice 7 therefore recommends that spatial data be published in CRSs to suit the potential user's applications. Spatial data on the Web should be published at least in WGS 84, and additionally in other CRSs if there are use cases demanding this.

The ability to unambiguously identify the CRS used is fundamental for the correct interpretation of spatial data. For instance, part of the information defined in a CRS concerns the order in which the geographic coordinates

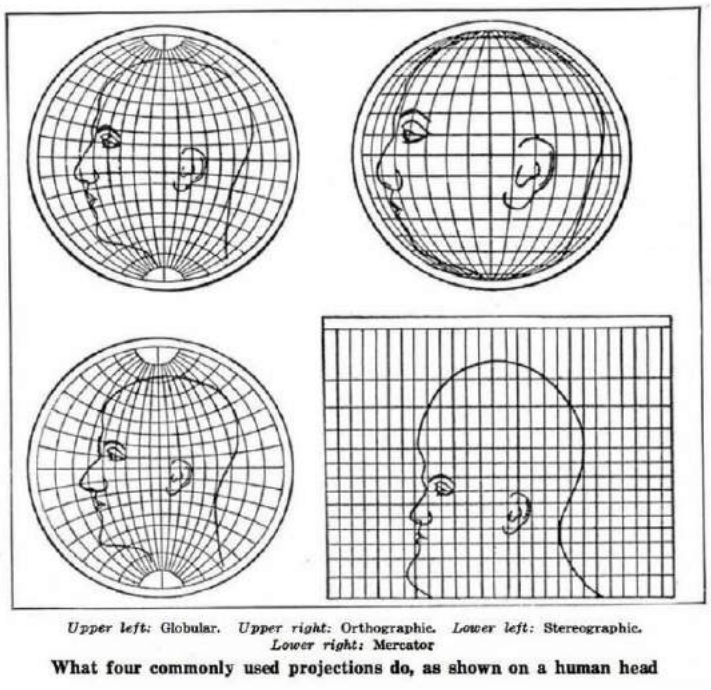


Figure 6.3: Projected CRS. (Public domain image from *Scientific American* circa 1923).

(i.e., latitude, longitude, etc.) are encoded, the units of measurement used for these coordinates, as well as the *datum* used. Mistaking the coordinate order (i.e., the *axis order*) is a very common error that results in plotting the data in a completely different location. Best Practice 8 specifically addresses this issue, by requiring that this information is made explicit, either by specifying the CRS used—use of EPSG codes, where they exist, is recommended[9].—, or by using data formats / vocabularies where this information is implicit (e.g., GeoJSON (Butler et al., 2016) and the W3C Basic Geo vocabulary (Brickley, 2003) support only one CRS—respectively, CRS84 and WGS 84).

6.3.3 Spatial identifiers

Spatial things should be uniquely identified with persistent Uniform Resource Identifiers (URIs) in order for those using spatial data on the Web (Best Practice 1) to be able to definitively combine and reference these resources (Best Practice 3): they become part of the Web's information space; contribut-

[9]EPSG is a register of CRSs maintained by the IOPG, an oil industry organization. The ESPG register is available online at: `http://www.epsg-registry.org/`

Figure 6.4: The Web site `http://thetruesize.com/` is a good tool for comparing sizes of countries at different latitude and shows the distortion which results from flattening a sphere.

ing to the Web of Data. URIs are a single global identification system used on the World Wide Web, similar to telephone numbers in a public switched telephone network. HTTP(S) URIs are a key technology to support Linked Data by offering a generic mechanism to identify entities ('Things') in the world and to allow referring to such entities by others. Anyone can assign identifiers to entities ('Things') in a namespace they own, i.e. a domain name within the Internet—e.g., hospitals, schools, roads, equipment, etc. 'Spatial things', such as the catchment area of a river, the boundaries of a building, a city or a continent, are examples of such 'Things' on the Web that need to be identified so that it is possible to refer to or make statements about this particular spatial thing.

Spatial things described or mentioned in a dataset on the Web should be identified using a globally unique URI so that a given spatial thing can be unambiguously identified in statements that refer to it. Good identifiers for data on the Web should be dereferenceable / resolvable, which makes it a good idea to use HTTP—or HTTPS—URIs as identifiers. Data publishers need to assign their subject spatial things HTTP URIs from an Internet domain name where they have authority over how the Web server responds.

Typically, this means minting new HTTP URIs. Important aspects of this include authority, persistence, and the difference between information resources and the thing they give information about. The use of a particular Internet domain may reinforce the authority of the information served. HTTP can only serve information resources such as Web pages or JSON documents. Best Practice 1 contains no requirement to distinguish between the spatial thing and the page/document unless an application requires this.

HTTP URIs allow objects to link (through hyperlinking) with the existing objects on the Web (using their URIs), i.e., link to URIs available in community-maintained Web resources such as DBpedia[10] and GeoNames[11] or public government data, such as a national registers of addresses. Mapping and cadastral authorities maintain datasets that provide geospatial reference data. Re-using well-known identifiers is a good practice because it makes it easy to recognize that data from different sources are related. An example of a spatial thing URI is http://sws.geonames.org/2172517/ which identifies the spatial thing *Canberra*, while when resolved in a browser returns information such as the name of the spatial thing, the centroid location, geographical boundaries, etc.

When exposing spatial data through standard SDIs (e.g., WFS services (Vretanos, 2010)), a certain user group is catered for, i.e., users with the expertise and tooling to use these services based on standards from the geospatial domain. To allow more users to benefit from the data it is important to expose the link to the Web representation of the features on the Web. Best Practice 12 identifies two approaches for doing this while keeping the SDI in place. One is to add an attribute to all spatial things, named `rdf_seealso`, which contains the URI of the Web representation of the spatial thing visible on the map. The Web representation is created by mapping the data in the SDI dynamically to crawlable resources on the Web using the R2RML standard (Das et al., 2012) and Linked Data Publication tools. This approach leverages existing SDIs while enriching them with dereferenceable linked data representations of the spatial things. Exposing the data about a spatial thing as linked data makes sure that the attributes themselves will be URIs and thus unambiguous. The overhead of the extra attribute on existing SDIs is negligible and traditional clients will not be hindered by the extra attribute, but more advanced usage allows unlocking of a wealth of connections behind the traditional spatial data.

The other approach is to create a RESTful API as a wrapper, proxy or a shim layer around WFS services. It is worth mentioning that different

[10] `http://dbpedia.org/`
[11] `http://geonames.org/`

spatial data services are often making their data accessible through REST API services. Content from the WFS service can be provided in this way as linked data, JSON or another Web-friendly format. There are examples of this approach of creating a convenience API that works dynamically on top of WFS such as the experimental *ldproxy*.[12] This is an attractive option for quickly exposing spatial data from existing WFS services on the Web. The approach is to create an intermediate layer by introducing a proxy on top of the WFS (data service) and CSW (metadata service) (Nebert et al., 2016) so the contained resources are made available. The proxy maps the data and metadata to schema.org (schema.org, 2018) according to a provided mapping scheme; assigns URIs to all resources based on a pattern; makes each resource available in HTML, XML, JSON-LD (Sporny et al., 2014), GML (Consortium et al., 2007), GeoJSON (Butler et al., 2016), and RDF/XML (metadata only); and generates links to data in other datasets using SPARQL queries (W3C SPARQL Working Group, 2013).

6.3.4 Discovery of spatial information

Cataloging of spatial information has always been difficult, whether the data is digital or not. A roadmap for Wellington NZ may be filed under NZ, Wellington, Transportation, tourism or a large number of other categories. Spatial data therefore has a greater requirement for metadata. Best Practice 13 recommends the inclusion of spatial metadata in dataset metadata. As a minimum, the spatial extent should be specified: the area of the world that the dataset describes. This information is necessary to enable spatial queries within catalog services such as those provided by SDIs and often suffices for initial discovery. However, further levels of description are needed for a user to be able to evaluate the suitability of a dataset for their intended application. This includes at least spatial coverage (continuity, resolution, properties), and representation (for example vector or grid coverage) as well as the coordinate reference system used.

In SDIs, the accepted standard for describing metadata is ISO 19115 (ISO/TC 211, 2003; ISO, 2014) or profiles thereof. To provide information about the spatial attributes of the dataset on the Web, DCAT (Maali and Erickson, 2014) is recommended. An application profile of DCAT for geospatial data, GeoDCAT-AP (ISA GeoDCAT-AP Working Group, 2016), can be applied to more fully express spatial metadata. In addition, several spatial ontologies, already mentioned in Section 6.3.1, allow the description of spatial datasets.

[12]—urlhttps://hub.docker.com/r/iide/ldproxy/

6.3.5 Scale and quality

The quality of spatial data, as that of any other data, has a big impact on the quality of applications that use such data. This is intensified in the Web scenario, where data could be consumed with unforeseen purposes and data that are useful for a particular use case could come from plenty of sources.

Therefore, having quality information about spatial data on the Web significantly facilitates two main tasks to the consumer of such data. One of them is the selection of data, allowing to focus on data that satisfy the needs of a concrete use case. For example, spatial data accuracy is important when using them in the application of self-driving cars; guiding an autonomous vehicle to a precise parking spot near a facility that has a time-bounded service and then booking the charging spot requires accurate location data. Another is the reuse of data, i.e., understanding the behavior of data in order to adapt its processing (e.g., by considering data currentness, refresh rate or availability). A fundamental concept of spatial data is the scale of the representation of the spatial thing. This is important because combining data designed to be used at differing scales often produces misleading results. A scale is often represented as a ratio or shortened to the denominator value of a ratio. A 1:1,000 (or 1,000) scale map is referred to as larger than a 1:1,000,000 (or 1,000,000) scale map. Conversely, a million scale map is said to be a smaller scale than a thousand scale map. Data collected at a small scale is most often more generalized than data collected at a larger scale. Concepts related to scale are resolution (the smallest difference between adjacent positions that can be recorded), accuracy (the amount of uncertainty—how well a coordinate designates a position on the actual Earth's surface) and precision (the reproducibility of a measurement to a certain number of decimal places in coordinate values).

When publishing spatial data on the Web, they should be supplemented with information about the precision and accuracy of such data. Such quality information should at least be available for humans. Best Practice 14 is concerned with positional accuracy and how spatial data should be accompanied with information about its accuracy. Best Practice 7 asserts a link between CRS and data quality, because the accuracy of spatial data depends for a large part on the CRS used, as was explained in a previous section. In order to support automatic machine-interpretation of quality information, such information should be published following the same principles as any other data published on the Web. The CRS of geometries on the Web should always be made known. For describing other quality aspects, the W3C Data Quality Vocabulary (DQV) (Albertoni and Isaac, 2016), which allows specifying data quality information (such as precision and accuracy), could be

used.

Even if the recommendations in Best Practice 14 focus on precision and accuracy, evidently the same advice can be followed for other relevant spatial quality information.

6.3.6 Thematic layering and spatial semantics

Spatial data is typically collected in layers. Although this sounds very map-oriented, these layers can be thought of as collections of instances of a class within a spatial and temporal frame. In other words, layers are usually organized semantically. Although SDWBP does not address layers directly, it does address spatial semantics. DWBP recommends the use of vocabularies to communicate the semantics, i.e., the meaning of data, and preferably standardized vocabularies. There are several vocabularies about spatial things available, such as the W3C Basic Geo vocabulary (Brickley, 2003), GeoRSS (Lieberman et al., 2007), GeoSPARQL (Perry and Herring, 2010b), the ISA Core Location Vocabulary (Perego and Lutz, 2015) or schema.org (schema.org, 2018); overviews on their uses are provided in the literature (Battle and Kolas, 2012; Athanasiou et al., 2013; van den Brink et al., 2014). Best Practice 4 identifies the main vocabularies in which spatial things can be described when the aim is data integration; however, it does not recommend one of them as the best. Currently there is no common practice in the sense of the same spatial vocabulary being used by most spatial data publishers. This depends on many factors; furthermore, describing spatial data multiple times using different vocabularies maximizes the potential for interoperability and lets the consumers choose which is the most useful. Appendix A of SDWBP offers guidance for selecting vocabularies by providing a table comparing the most commonly used ones.

The most important semantic statement to be made when publishing spatial data—or any data—is to specify the type of a resource. The W3C Basic Geo vocabulary has a class `SpatialThing` which has a very broad definition. This can be applicable (as a generic concept) to most of the common use-cases. Thematic semantics and general descriptions of spatial things and their properties should be provided as linked data. They should have URIs that return human- and machine-readable definitions when resolved.

6.3.7 Temporal dimension

It is important to consider the temporal dimension when publishing spatial data. The “where” component of the data is seldom independent of the “when” which may be implicit or explicit to the data. Hence, capturing the

temporal component makes spatial data more useful to potential users, since it allows them to verify whether the data suits their needs. For instance, tectonic movements over time can distort the coordinate values of spatial things.

This is included in Best Practice 7 on coordinate reference systems as valuable knowledge when dealing with spatial data for precision applications. Furthermore, it is recommended in Best Practice 13 to include metadata statements about the (most recent) publication date, the frequency of update and the time period for which the dataset is relevant (i.e., temporal extent).

Apart from the need for enhancing spatial data with their temporal context, the temporal dimension also affects the very nature of spatial things, since both spatial things and their attributes can change over time; this is covered in Best Practice 11. When dealing with changes to a spatial thing, its lifecycle should be taken into account; in particular, how much change is acceptable before a spatial thing can no longer be considered as the same resource (and requires defining a new resource with a new identifier). Creating a new resource will depend on whether domain experts think the fundamental nature of the spatial thing has changed, taking into account its lifecycle (e.g., a historic building replaced by another that has been built on top of it). In this case, the temporal dimension of spatial data can be expressed by providing a series of immutable snapshots that describe the spatial thing at various points in its lifecycle, each snapshot having a persistent URI.

In those cases when the spatial thing itself does not change over time but its attributes do, the description of the spatial thing should be updated to reflect these changes, and each change published as a snapshot. In contrast, when a spatial thing has a small number of attributes that are frequently updated (e.g., the GPS-position of a runner or the water level from a stream gauge), the time-series of data values within such attributes of the spatial thing should be represented. This is relevant in relation to recent advances in embedding smart sensors and actuators in physical objects and machines such as vehicles, buildings, and home appliances, which has led to the publishing of large volumes of data that are spatio-temporal (Aggarwal et al., 2013).

6.3.8 Size of spatial datasets

Spatial data tends to be very large. This can pose difficulties when sharing or consuming spatial data over the Web—particularly in low bandwidth or high latency situations. Accurate (polygon) geometries tend to contain a high number of coordinates. Especially when querying collections of spatial things with geometries over the Web, this results in very large response payloads wasting bandwidth and causing slow response times, while for some very

common use cases, like simply displaying some things on a Web map, high accuracy is not required. The primary basis for simplification is scale. For example, when searching for a Starbucks on a city scale, an accuracy of 3 meters is acceptable, but when providing street-level directions to a shop for a self-driving car it is not.

For those use cases that do not require high accuracy, common ways of dealing with reducing the size of spatial data include degrading precision by reducing the number of decimals, and simplifying geometries using a simplification algorithm such as Ramer-Douglas-Peucker (Ramer, 1972; Douglas and Peucker, 1973) or Visvalingam and Whyatt (1993). These methods result in lower accuracy and precision.

Big spatial data is often not vector data, i.e., a representation of spatial things using points, lines, and polygons (ISO, 2015a), but coverage data, i.e., gridded data: a data structure that maps points in space and time to property values (ISO/TC 211, 2005b). For example, an aerial photograph can be thought of as a coverage that maps positions on the ground to colors. A river gauge maps points in time to flow values. A weather forecast maps points in space and time to values of temperature, wind speed, humidity and so forth. For coverage data, other methods are required to manage size.

DWBP recommends to provide 1) bulk download and 2) subsets of data. Providing bulk-download or streaming access to data is useful in any case and is relatively inexpensive to support as it relies on standard capabilities of Web servers for datasets that may be published as downloadable files stored on a server. Subsets, i.e., extracts or "tiles", can be provided by having identifiers for conveniently sized subsets of large datasets that Web applications can work with (Brizhinev et al., 2017). Actually, breaking up a large coverage into pre-defined lumps that you can access via HTTP GET requests is a very simple API. A second way of supporting extracts, more appropriate for frequently changing datasets, is by supplying filtering options that return appropriately sized subsets of the specific dataset.

6.3.9 Crawlability

Search engines use crawlers to update their search index with new information that has been found on the Web. Traditional crawlers consist of two components: URL extractors, i.e., HTML crawlers that extract links from HTML pages in order to find additional sources to crawl, and indexers, i.e., different types of indexes, typically using the occurrence of text on a Web page, that are maintained by search engines.

A major issue with crawlers identified in the early 2000's by Bergman (2012) was the inaccessibility of the so-called "deep Web": information that

was hidden to traditional crawlers as it is only accessible through services (e.g., forms) that require user input. Several solutions have since been introduced to access and index information on the "deep Web" (He et al., 2007; Madhavan et al., 2008). Spatial Data services (e.g. OGC Web services and/or other APIs) typically make information available only after user input has been provided, leading to a similar problem, i.e., a Deep Spatial Web. Further, these services are built to be accessed and searched by domain-specific applications rather than general Web services and/or search engine crawlers. For example, in the OGC architecture, catalog services are intended to be used for searching spatial assets by using specific client tools, and are not optimized for indexing and discovery by general purpose search engines. However, a typical Web user does not know that these catalogs exist and is accustomed to using general purpose search engines for finding information on the Web. Therefore, making sure that spatial data is indexable by Web search engines is an important approach for making spatial data discoverable by users directly. The addition of structured data to Web services that are otherwise not accessible to search engines increases the visibility of a service or dataset in major search engines (Guha et al., 2016). Schema.org (schema.org, 2018), a single schema across a wide range of topics that includes people, places, events, products, offers, etc., and widely supported by Bing, Google, Yahoo! and Yandex, is the predominant way of marking up content and services on the Web with structured data to improve the presentation of the result in a search engine (Guha et al., 2016). To verify if schema.org markup on a Web page is recognized by Web agents, Google's Structured Data testing tool[13] can be used. Experimental work such as Geonovum's Spatial Data on the Web Testbed[14] describes ways to make spatial data indexable by publishing an HTML Web page for the spatial dataset and each spatial thing that it contains; by using structured data, schema.org and links, as well as publishing XML sitemaps containing links to all data resources for spatial data services. Another example is the Dutch geoportal PDOK.nl which extensively publishes dataset metadata, for example, the national roads dataset,[15] resulting in better accessibility through common search engines. Several examples of spatial things published in this way are provided in Best Practice 2.

Currently, spatial information, even when published in accordance with these guidelines, is not widely exploited by search engines. However, by increasing the volume of spatial information presented to search engines, and the consistency with which it is provided, it is expected that search en-

[13] `https://search.google.com/structured-data/testing-tool`
[14] `https://github.com/geo4web-testbed/general`
[15] `https://data.pdok.nl/datasets/nationaal-wegenbestand-wegen`

gines will begin offering spatial search functions. Evidence is already seen of this in the form of contextual search, such as prioritization of search results from nearby entities. In addition, search engines are beginning to offer more structured, custom searches that return only results that include certain schema.org types such as `Dataset`, `Place` or `City`.

6.3.10 Other aspects of spatial data

Spatial things may have 0 to 3 dimensions (points, lines, areas, 3D), and it may be difficult to combine similar things if the dimensions in which they are represented differ. Although SDWBP does not address this at length, Best Practice 5 does recommend describing the number of dimensions in metadata and notes that one of the selection criteria for choosing a geometry format on the Web is the dimensionality of the data.

In common language, and for spatial things inherently related to mobile things, it may be convenient to describe positions of spatial things *relative* to other spatial things. Just as for other spatial things, we would like such descriptions to be both human- and machine-interpretable. We advise, in Best Practice 9, that positions or geometries of the target reference things should be retrievable via link relations, in accord with general principles for linkability in Section 6.2. The geocentric use case (i.e., position relative to the Earth) is generally addressed throughout SDWBP and does not require an explicit link relation to the Earth. As the active contributors to SDWBP were primarily interested in the geocentric case, SDWBP does not explore relative positioning in depth. We find that practices in this area vary widely in the details of implementation and so we are able to offer only broad advice and examples. Spatial relationships as described for Best Practice 10 (Section 6.3.1) may be useful, but we find no evidence for suitable common vocabularies for many needs. For some spatial data, the *symbology* associated with the data is of high importance because it communicates meaning. However, as rendering maps is explicitly out of scope, symbology is not addressed in SDWBP.

6.4 Gaps in current practice

The best practices described in brief in Section 6.3, and in full in the SDWBP document, are compiled based on evidence of real-world application. This is in line with the fourth principle described in Section 6.2 of this paper. However, there are several issues that inhibit the use or interoperability of spatial data on the Web, for which no evidence of real-world applied solutions

was found. These issues are denoted "gaps in current practice" and described as such in this section. An issue is considered a gap when no evidence of real-world applied solutions in production environments was found. The term 'production environment' signifies a case where spatial data has been delivered on the Web with the intention of being used by end users and with a quality level expected from such data. In contrast, a "testing environment" is published with the intent of being tested so that bugs can be discovered and fixed and an experimental publication of spatial data on the Web is published with the intent of, for example, exploring possibilities, learning about the technology, or other goals besides publishing with the intent of serious use. In the case of gaps, there might be emerging practice, i.e., a solution that has been theorized for a certain issue and has possibly been experimented on in testing/beta settings, but not in production environments. Gaps and emerging practices in the area of publishing spatial data on the Web are discussed in this section.

6.4.1 Representing geometry on the Web

Location information can be an important 'hook' for finding information and for integrating different datasets. There are different ways of describing the location of spatial things: referencing the name of a well-known named place, describing a location in relation to another location, or providing the location's coordinates as a geometry. The latter allows the integration of data based on location using spatial reasoning, even when explicit links between things are not available, as well as, of course, showing spatial things on a map. Although there are ways to represent geometry in Web data, there are still some gaps in current practice related to selecting a serialization format, selecting an embedded, file-based or Linked Data approach, and making geometries available in different CRSs.

There are several aspects to representing geometries on the Web. First, there is the question of different serialization formats to choose from. In general, the formats are the same as for publishing any other data on the Web: XML, JSON, CSV, RDF, etc. How to select the most appropriate serialization is described in general terms in DWBP. As described in Section 6.3.1, a single best way of publishing geometries was not identified in SDWBP. The currently most widely used geometry formats, namely, GML (Consortium et al., 2007) and GeoJSON (Butler et al., 2016), do not address all requirements. Therefore, in order to facilitate the use of geometry data on the Web, SDWBP recommends that GML-encoded geometries are made available also in GeoJSON, by applying not only the required coordinate reference system transformation, but, if needed, by simplifying the original geometry (e.g., by

transforming a 3D geometry in a 2D one).

A second, related aspect concerns the existing options for publishing geometries on the Web—e.g., in self-contained files such as GML or GeoJSON, or rather to embed geometries as structured data markup in HTML, or in an RDF-based way, i.e., as Linked Data. Choosing between these approaches—or not choosing but rather offering a combination of these—depends largely on the intended audience. As explained in Section 6.3.9, dealing with crawlability, the advantage is that HTML with embedded data is indexed by search engine crawlers. However, the options for embedding geometry in HTML are limited. Typically this is done using schema.org (schema.org, 2018) as Microdata (McCathie Nevile and Brickley, 2017), RDFa (Sporny, 2015), or JSON-LD (Sporny et al., 2014), but for specifying only 0D-2D geometries (points, lines, surfaces)—e.g., the centroid and/or 2D bounding box. An additional issue is that search engines index only a subset of the terms defined in schema.org, and those concerning geometries are not fully supported. RDF-based publication of geometry data is the most advanced option, but the audience for this is smaller than the others. GeoSPARQL (Perry and Herring, 2010a) offers a vocabulary that allows serialization of geometries as GML or WKT ("Well-Known Text") (Herring, 2011), whereas the ISA Core Location vocabulary (Perego and Lutz, 2015) supports also GeoJSON, but the lack of best practices on the consistent use of the existing spatial data vocabularies prevents interoperability (see Section 6.4.2).

Third, there is a question of how to make geometries available in different CRSs. Section 6.3.2 explains the existence of many CRSs and why spatial data should be published in CRSs that are most common to potential users. It follows that, on the one hand, the CRS should be specified for geometries published on the Web and, on the other hand, users should be able to find out which CRSs are available and to get geometries using the CRS of their choice.

Sometimes the CRS used is clear from the representation. In other cases the CRS needs to be specified either on the dataset level or the instance level. How this is done differs for each serialization. For example, in GeoSPARQL this is added as a prefix of the WKT literal while in GML an attribute `srsName` can be specified on geometry elements. In an OGC WFS (Vretanos, 2010) request, users can specify the CRS they wish to use by specifying the `srsName` parameter. In GeoSPARQL the `getSRID` function returns the spatial reference system of a geometry, thus making it possible to request a specific CRS at a (Geo)SPARQL endpoint. However, these options require the user to be proficient in either geospatial Web services or Linked Data.

A best practice for returning geometries in a specific requested CRS has not yet emerged. Many options can be found in current practice, includ-

ing creating CRS-specific geometry properties (for example, the Dutch Land Registry does this), and supporting an option for requesting a specific CRS in a convenience API; but one best practice cannot yet be identified. Another option worth exploring might be the use of content negotiation, i.e., negotiate CRS as part of the content format for the geometry, as has also been proposed for encoding format. For example, this could be done with an extra media type parameter (e.g., `application/ttl; geomLiteral="WKT"; crs="CRS84"`) or by adding specific request and response headers for negotiating CRS to the HTTP protocol. A contribution to address this issue might be provided by the W3C Dataset Exchange Working Group (DXWG), which is due to deliver a specification on profile-based content negotiation.[16] However, providing different CRS might be too complicated to handle in the HTTP protocol. For example, multidimensional datasets will in general use multiple CRSs (e.g., horizontal, vertical and temporal, maybe more), and conversion between CRSs will, in general, introduce errors, so data in one CRS are not exactly the same as data in another CRS. Furthermore, CRS is just one of a number of parameters that may characterize a particular geometric representation of a spatial thing, including the type of geometry, its relationship to the Thing, method of interpolation, scale or resolution. However, offering a choice between all these parameters of data objects such as geometries might be an overloading of HTTP content negotiation protocols. It might, therefore, be more appropriate to handle this in the application layer.

6.4.2 A spatial data vocabulary

Although a large amount of geospatial data has been published on the Web, so far there are few authoritative datasets containing geometrical descriptions of spatial things available in Web-friendly formats. Their number is growing (e.g., at the time of writing there are three authoritative spatial datasets publicly available as linked data in the Netherlands containing topographic[17], cadastral[18], and address[19] data), but currently there is no common practice in the sense of the same spatial vocabulary—or the same combination of spatial vocabularies—being used by most spatial data publishers. The consequence is the lack of a baseline during the mapping process for application developers trying to consume specific incoming data. Identifying administra-

[16] For a description of this deliverable, see the DXWG Charter: `https://www.w3.org/2017/dxwg/charter`

[17] `https://brt.basisregistraties.overheid.nl`

[18] `https://brk.basisregistraties.overheid.nl`

[19] `https://bag.basisregistraties.overheid.nl`

tive units, points of interest or postal addresses with URIs could be beneficial not only for georeferencing other datasets, but also for interlinking datasets georeferenced by the direct and indirect location information.

Direct georeferencing of data implies representing coordinates or geometries and associating them to a CRS. This requires vocabularies for geometries and able to specify which CRS is used. Further, indirect georeferencing of data implies associating them to other data on named places. Preferably, these data on named places should be also georeferenced by coordinates in order to serve as the basis for data linking between indirectly and directly georeferenced datasets. Moreover, vocabularies developed for representing specific sets of geospatial data on the Web should reuse as much as possible existing ones. This is the case for the vocabularies developed by IGN France for geometries,[20] topographic entities,[21] and CRS.[22] These vocabularies contain alignments with existing vocabularies, e.g., the class `geom:Geometry` is a subclass of both `sf:Geometry` (OGC Simple Features vocabulary) which is a subclass of the Geometry class of the GeoSPARQL vocabulary; and `ngeo:Geometry` (Neogeo vocabulary). Furthermore, the topographic entity class from the IGN France vocabulary is declared equivalent to the `Feature` class from the Geonames vocabulary.

In W3C Basic Geo (Brickley, 2003), it is assumed that the CRS used is WGS 84. However, publishers might have data in a different, local CRS. Thus, there is a need for a more generic class for, for example, a point geometry with the benefit of choosing the CRS of the underlying data (Atemezing and Troncy, 2012). Existing vocabularies, as GeoSPARQL (Perry and Herring, 2010b) and ISA Core Location (Perego and Lutz, 2015), support this feature, but there are currently no best practices for their consistent use, thus hindering interoperability.

Vocabularies like the one by IGN France are created because, currently, the existing vocabularies do not cover all requirements and no guidance is available on their consistent and complementary use. SDWBP partially addresses the latter issue, by providing examples on how to model spatial information with widely used vocabularies. However, solving the existing gaps would require the definition of new terms in existing or new vocabularies, which was not in scope with the work of the Spatial Data on the Web Working Group. A possible way forward is an update for GeoSPARQL. This would provide an agreed spatial ontology, i.e., a bridge or common ground between geographical and non-geographical spatial data and between W3C and OGC

[20] `http://data.ign.fr/def/geometrie/20160628.en.htm`
[21] `http://data.ign.fr/def/topo/20140416.en.htm`
[22] `http://data.ign.fr/def/ignf/20160628.en.htm`

standards, conformant to the ISO 19107 (ISO, 2003) abstract model and based on existing vocabularies such as GeoSPARQL (Perry and Herring, 2010b), the W3C Basic Geo Vocabulary (Brickley, 2003), GeoRSS (Lieberman et al., 2007), NeoGeo or the ISA Core Location vocabulary (Perego and Lutz, 2015). The updated GeoSPARQL vocabulary would define basic semantics for the concept of a reference system for spatial coordinates, a basic datatype, or basic datatypes for geometry, how geometry and real world objects are related and how different versions of geometries for a single real world object can be distinguished. For example, it makes sense to publish different geometric representations of a spatial object that can be used for different purposes. The same object could be modeled as a point, a 2D or a 3D polygon. The polygons could have different versions with different resolutions (generalization levels). And all those different geometries could be published with different coordinate reference systems. Thus, the vocabulary would provide a foundation for harmonization of the many different geometry encodings that exist today. An alternative approach is to establish best practices for the consistent use of the most popular spatial vocabularies. An example are the guidelines for the RDF representation of INSPIRE data developed in the framework of the EU ISA Programme (ISA Action ARe3NA, 2017) by following the SDWBP recommendations. In such a scenario, the definition of new classes and properties would be limited to cover the gaps in the existing vocabularies.

6.4.3 Spatial aspects for metadata

Even if all spatial data should become discoverable directly through search engines, data portals would still remain important hubs for data discovery—for example, because the metadata records registered there can be made crawlable. But, in addition, different data portals can harvest each others' information provided there is consistency in the types and meaning of included information, even if structures and technologies vary. Discovery of spatial data is improved in the Netherlands, for example, because the national general data portal[23] harvests the spatial data portal and thus all spatial datasets are registered in the general data portal as well.

In the eGovernment sector, DCAT (Maali and Erickson, 2014) is a standard for dataset metadata publication, and harvesting this metadata is implemented by eGovernment data portals. Because DCAT is lacking in possibilities for describing some specific characteristics of spatial datasets, an application profile for spatial data, GeoDCAT-AP (ISA GeoDCAT-AP Working

[23] `https://data.overheid.nl`

Group, 2016; Perego et al., 2017), has been developed in the framework of the ISA Programme of the European Union,[24] with the primary purpose of enabling the sharing of spatial metadata across domains and catalog platforms. To achieve this, GeoDCAT-AP defines RDF bindings covering the core profile of ISO 19115:2003 (ISO/TC 211, 2003) and the INSPIRE (Union, 2007) metadata schema, enabling the harmonized RDF representation of existing spatial metadata. The focus was on the most used metadata elements, whereas additional mappings—as well as the alignment with the latest version of ISO 19115 (ISO, 2014)—could be defined in future versions of the specification, based on users' and implementation feedback.

One of the outcomes of the development of GeoDCAT-AP was the identification of gaps in existing RDF vocabularies for representing some spatial information (Perego et al., 2016)—such as coordinate reference systems and spatial resolution (see also Section 6.4.2 on this topic). But it also highlighted a key issue for spatial data, in that the use of global and persistent identifiers is far from being a common practice. Apart from making it difficult to implement a Linked Data-based approach, this situation has negative effects on the geospatial infrastructure itself. E.g., it makes it impossible to unambiguously identify a spatial thing or a dataset over time, and it prevents an effective implementation of incremental metadata harvesting in a federated infrastructure (such as the INSPIRE one).

Notably, recent activities are contributing to fill at least part of these gaps. For instance, DQV (Albertoni and Isaac, 2016) provides a solution for modeling precision and accuracy, as mentioned in Section 6.3.5. Moreover, the use cases collected by the W3C Dataset Exchange Working Group (Faniel et al., 2017) cover also geospatial requirements, which might be addressed by the work of this group in the revision to DCAT (Maali and Erickson, 2014). However, the consistent use of global and persistent identifiers in the geospatial domain is an issue that, far from being merely technical, affects the data management workflow, and therefore needs to be addressed also at the organizational level.

6.4.4 Describing dataset structure and service behaviors

Datasets may be arbitrarily large and complex, and may be exposed via services to expose useful resources, rather than a "download" scenario. Data gathered using automated sensors, in particular, may be impossible to download in its entirety due to its dynamic nature and potential volumes. It is,

[24] `https://ec.europa.eu/isa2/`

therefore, necessary in these cases to be able to adequately describe the structure of such data and how services interact to expose subsets of it—even individual records in a Linked Data context. Such datasets are common in the information processing world, and commonly organized in "hypercubes"—where "data dimensions" are used to locate values holding results. A standard based on this dimensional model of data is the RDF Data Cube vocabulary (QB) (Cyganiak and Reynolds, 2014). It has been used to publish sensor data; for example to publish a homogenized daily temperature dataset for Australia over the last 100 years (Lefort et al., 2012). However, QB is lacking in possibilities for describing spatio-temporal aspects of data, which are very important for observations. One of the work items in the Spatial Data on the Web Working Group was, therefore, an extension to the existing QB vocabulary to support specification of key metadata required to interpret spatio-temporal data, called QB4ST (Atkinson, 2017).

QB4ST is an extension to QB to provide mechanisms for defining spatio-temporal aspects of dimension and measure descriptions. It is intended to enable the development of semantic descriptions of specific spatio-temporal data elements by appropriate communities of interest, rather than to enumerate a static list of such definitions. It provides a minimal ontology of spatio-temporal properties and defines abstract classes for data cube components (i.e., dimensions and measures) that use these, to allow classification and discovery of specialized component definitions using general terms. QB4ST is designed to support the publication of consistently described re-usable and comparable definitions of spatial and temporal data elements by appropriate communities of practice. One obvious such case is the use of GPS coordinates described as decimal latitude and longitude measures. Another example is the intended publication of a register of Discrete Global Grid Systems (DGGS) by the OGC DGGS Working Group. QB4ST is intended to support publication of descriptions of such data using a common set of attributes that can be attached to a property description (extending the available QB mechanisms for attributes of observations). The Spatial Data on the Web Working Group has demonstrated the use of QB and QB4ST to serve satellite imagery through DGGS and a virtualized triple store (Brizhinev et al., 2017).

6.4.5 Versioning of spatial data

Future Internet technologies will aim more and more to capture, make sense of and represent not only static but dynamic content and up-to-date information. Through Internet technology, connections between devices and people will be realized through the exchange of large volumes of multimedia and

data content. When the communication latency becomes lower, and the capacity of the communication gets higher with fifth generation (5G) mobile communications systems, the Internet will be an even more prominent platform to control real and virtual objects in different parts of our lives, such as healthcare, education, manufacturing, smart grids, and many more (Jaber et al., 2016). The Internet of Things (Barnaghi and Sheth, 2014) and Tactile Internet (Simsek et al., 2016) are some of the technologies that aim to facilitate the interaction between people and devices, observe near real-time phenomena and actuate devices or robots. The Tactile Internet is focused on speeding up this interaction process and reducing the latency in communication systems. Such high-speed communication will bring new challenges for intelligent systems. There is a big gap in the lightweight and semantically rich representation of versioning and temporal aspects of spatial data content. There have been a few attempts to represent changing and moving spatial objects, such as OGC's TimeSeriesML (Tomkins and Lowe, 2016) and Moving Features (Hayashi et al., 2017). However, although these ontologies provide a reasonably good semantic coverage, there is still a need for the development of lightweight and semantically rich representations to conduct enhanced (near) real-time operations. Having heavy semantic expressivity in ontologies can cause a burden on reasoning engines and can slow down the processing time for machines. For instance, it is important to identify and represent the direction and coverage area of a surveillance camera or the orientations and positions of objects or people in the observed environment. In the future, a broken car will be fixed by a robot, or surgeries will be carried out by multiple robots using Tactile Internet (Aijaz et al., 2017). It can also be envisioned that people, who have difficulty in walking, will not need to use a walking stick but merely a strap of an exoskeleton. These must be controlled by wireless systems to monitor the coverage, direction, and identify the objects and people around them including their shapes.

To conduct such activities, a better representation is needed for not only spatial information but also temporal and geometrical aspects of objects. Observed objects can change their size during the actuation process. For instance, a group of surgical robots will need to know about the shape of an organ and the changes regarding its size, geometrical shape, and orientation during surgery and exchange this information among themselves and with doctors to conduct an operation with high precision and low latency. A self-driving wheelchair or a self-driving car will be able to communicate with other sensor objects regarding the surrounding environment and direction to avoid obstacles. This will prevent possible accidents and harm caused by machines, such as not falling down stairs or running into objects with high speed or force. In all these scenarios, lightweight representation and

exchange of temporal-spatial knowledge are essential to understand and react fast enough to prevent disasters or to control the movement of devices. Having the means to represent the semantics of activities and phenomena at such high granularity and lightweight format will play a pivotal role in the development of future Internet technologies. Moreover, it will allow machines to instantly exchange information including spatial and geometrical knowledge and carry out their tasks with high precision. The Spatial Data on the Web Interest Group[25], the successor of the Working Group, will address the topic of representing moving objects on the Web.

6.5 Conclusions

Spatial data has become ubiquitous with the explosive growth in positioning technologies attached to mobile vehicles, portable devices, and autonomous systems. It has proven to be fundamentally useful for countless things, ranging from everyday tasks like finding the best route to a location to solving the biggest global challenges like climate change adaptation. However, spatial data dissemination is heterogeneous and although the Web is commonly used as a publication medium, the discovery, retrieval, and interpretation of spatial data on the Web is still problematic.

SDWBP describes how Web principles can be applied to the world of complex spatial data to solve this problem. Good practices can be observed in current practice and have been collected into the Best Practices based on a set of principles and an examination of practice. In some cases, a best practice has not yet emerged. There are still questions related to representing geometry on the Web, with regard to recommendable serialization forms and formats, and the use of coordinate reference systems. A Web-friendly way of publishing spatial metadata has not yet been described in full, especially with regards to the relevant subset of spatial metadata standards. A standardized ontology for spatial things that covers all the main requirements for publishing spatial linked data is not yet available, and best practices on the consistent use of the existing spatial vocabularies are yet to be established. Finally, there are new approaches emerging such as QB4ST (Atkinson, 2017), an extension to the RDF Data Cube to provide mechanisms for defining spatio-temporal aspects of dimension and measure descriptions.

Notwithstanding these gaps and emerging solutions, a useful set of actionable best practices for publishing spatial data on the Web has been described. Following these guidelines will enable data users, Web applications and ser-

[25] `https://www.w3.org/2017/sdwig/charter.html`

vices to discover, interpret and use spatial data in large and distributed Web systems.

Bibliography

ISO/TC 211. ISO 19111:2005 Geographic information – Spatial referencing by coordinates. https://www.iso.org/standard/26016.html, 2005a.

ISO/TC 211. ISO 19119:2016 Geographic information – Services. https://www.iso.org/standard/39890.html, 2005b.

ISO/TC 211. ISO 19111:2007 Geographic information – Spatial referencing by coordinates. https://www.iso.org/standard/41126.html, 2007.

ISO/TC 211. ISO 19117:2012 Geographic information – Portrayal. https://www.iso.org/standard/46226.html, 2012.

C.C. Aggarwal, N. Ashish, and A.P. Sheth. The Internet of Things: A survey from the Data-Centric Perspective. In C.C. Aggarwal, editor, *Managing and mining sensor data*, pages 383–428. Springer, 2013.

A. Aijaz, M. Simsek, M. Dohler, and G. Fettweis. Shaping 5G for the Tactile Internet. In W. Xiang, K. Zheng, and X.S. Shen, editors, *5G Mobile Communications*, pages 677–691. Springer, 2017.

R. Albertoni and A. Isaac. Data on the Web Best Practices: Data Quality Vocabulary. https://www.w3.org/TR/2016/NOTE-vocab-dqv-20161215/, 2016.

G. Atemezing and R. Troncy. Comparing vocabularies for representing geographical features and their geometry. In *ISWC 2012, 11th International Semantic Web Conference, Terra Cognita Workshop*, volume 901, 2012.

S. Athanasiou, L. Bezati, G. Giannopoulos, K. Patoumpas, and D. Skoutas. GeoKnow: Making the web an exploratory place for geospatial knowledge. *Market and Research Overview.*, 2013.

R. Atkinson. QB4ST: RDF Data Cube extensions for spatio-temporal components. https://www.w3.org/TR/2017/NOTE-qb4st-20170928/, 2017.

P. Barnaghi and A. Sheth. The Internet of Things: The Story So Far. *IEEE Internet of Things eNewsletter*, 2014.

R. Battle and D. Kolas. Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL. *Semantic Web Journal*, 3(4):355–370, 2012.

P. Baumann. Ogs® Web WCS 2.0 Interface Standard - Core, version 2.0.1. http://www.opengis.net/doc/IS/wcs-core-2.0.1, 2012.

M.K. Bergman. The Deep Web: Surfacing Hidden Value. https://brightplanet.com/2012/06/the-deep-web-surfacing-hidden-value/, 2012.

Dan Brickley. Basic Geo (WGS84 lat/long) Vocabulary. Available online: http://www.w3.org/2003/01/geo/ (accessed 11 February 2014), 2003.

D. Brizhinev, S. Toyer, and K. Taylor. Publishing and Using Earth Observation Data with the RDF Data Cube and the Discrete Global Grid System. https://www.w3.org/TR/2017/NOTE-eo-qb-20170928/, 2017.

D. Burggraf. OGC KML 2.3. http://www.opengis.net/doc/IS/kml/2.3, 2015.

H. Butler, M. Daly, A. Doyle, S. Gillies, S. Hagen, and T. Schaub. The GeoJSON format, RFC 7946. https://www.rfc-editor.org/info/rfc7946, 2016.

Open Geospatial Consortium et al. Geography Markup Language (GML) Encoding Standard. Version 3.2.1, doc nr OGC 07-036 [online]. Available online: http://portal.opengeospatial.org/files/?artifact%5Fid=20509, 2007.

R. Cyganiak and D. Reynolds. The RDF Data Cube vocabulary. http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/, 2014.

R. Cyganiak, D. Wood, and M. Lanthaler. RDF 1.1 Concepts and Abstract Syntax. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/, 2014.

S. Das, S. Sundara, and R. Cyganiak. R2RML: RDB to RDF Mapping Language. http://www.w3.org/TR/2012/REC-r2rml-20120927/, 2012.

J. de la Beaujardiere. OpenGIS Web Map Server Implementation Specification. http://www.opengis.net/doc/IS/wms/1.3, 2006.

D.H. Douglas and T.K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, 10(2):112–122, 1973.

I. Faniel, J. Pullman, and R. Atkinson. Dataset Exchange Use Cases and Requirements. https://www.w3.org/TR/dcat-ucr/, 2017.

B. Farias Lóscio, C. Burle, and N. Calegari. Data on the Web Best Practices. https://www.w3.org/TR/2017/REC-dwbp-20170131/, 2017.

Y. Fathy, P. Barnaghi, and R. Tafazolli. Large-Scale Indexing, Discovery and Ranking for the Internet of Things (IoT). *ACM Computing Surveys*, 2017.

R.V. Guha, D. Brickley, and S. Macbeth. Schema. org: Evolution of structured data on the Web. *Commun. ACM*, 59(2):44–51, 2016.

H. Hayashi, A. Asahara, K.-S. Kim, R. Shibasaki, and N. Ishimaru. OGC® Moving Features Access. Version 1.0. http://www.opengis.net/doc/IS/movingfeatures-access/1.0, 2017.

B. He, M. Patel, Z. Zhang, and K.C.-C. Chang. Accessing the deep web. *Commun. ACM*, 50(5):94–101, 2007.

J.R. Herring. OpenGIS® Implementation Standard for Geographic information - Simple feature access - Part 1: Common architecture, Version 1.2.1. http://www.opengeospatial.org/standards/sfa, 2011.

B. Hyland, G. Atemezing, and B. Villazón-Terrazas. Best Practices for Publishing Linked Data. http://www.w3.org/TR/2014/NOTE-ld-bp-20140109/, 2014.

ISA Action ARe3NA. Guidelines for the RDF encoding of spatial data. http://inspire-eu-rdf.github.io/inspire-rdf-guidelines, 2017.

ISA GeoDCAT-AP Working Group. GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe. Version 1.0.1. https://joinup.ec.europa.eu/release/geodcat-ap/v101, 2016.

ISO. ISO 19107:2003 Geographic information – Spatial schema., 2003.

ISO. ISO 19110:2005 Geographic information – Methodology for feature cataloguing., 2005.

ISO. ISO 19115-1:2014 Geographic information – Metadata – Part 1: Fundamentals., 2014.

ISO. ISO 19103:2015 Geographic information – Conceptual schema language., 2015a.

ISO. ISO 10109:2015 Geographic information – Rules for application schema., 2015b.

ISO/TC 211. ISO 19115:2003 Geographic information – Metadata. https://www.iso.org/standard/26020.html, 2003.

ISO/TC 211. ISO 19119:2005 Geographic information – Services. https://www.iso.org/standard/39890.html, 2005a.

ISO/TC 211. ISO 19123:2005 Geographic information – Schema for coverage geometry and functions. https://www.iso.org/standard/40121.html, 2005b.

M. Jaber, M.A. Imran, R. Tafazolli, and A. Tukmanov. 5G backhaul challenges and emerging research directions: A survey. *IEEE access*, 4:1743–1766, 2016.

I. Jacob and N. Walsh. Architecture of the World Wide Web, Volume One. http://www.w3.org/TR/2004/REC-webarch-20041215/, 2004.

F. Knibbe and A. Llaves. Spatial data on the Web use cases & requirements. https://www.w3.org/TR/2016/NOTE-sdw-ucr-20161025/, 2016.

L. Lefort, J. Bobruk, A. Haller, K. Taylor, and A. Woolf. A linked sensor data cube for a 100 year homogenised daily temperature dataset. In *Proceedings of the 5th International Conference on Semantic Sensor Networks-Volume 904*, pages 1–16. CEUR-WS. org, 2012.

J. Lieberman, R. Singh, and C. Goad. W3C Geospatial Vocabulary. https://www.w3.org/2005/Incubator/geo/XGR-geo/, 2007.

M. Lupp. Styled Layer Descriptor profile of the Web Map Service Implementation Specification. http://www.opengis.net/doc/IS/sld/1.1, 2007.

F. Maali and J. Erickson. Data Catalog Vocabulary (DCAT). http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/, 2014.

J. Madhavan, D. Ko, Ł. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy. Google's Deep Web crawl. *Proc. VLDB Endow.*, 1(2):1241–1252, 2008.

I. Masser. All shapes and sizes: the first generation of national spatial data infrastructures. *International Journal of Geographical Information Science*, 13:67–84, 1999.

C. McCathie Nevile and D. Brickley. HTML Microdata. https://www.w3.org/TR/2017/WD-microdata-20171010/, 2017.

M. Müller. Symbology Encoding Implementation Specification. http://www.opengis.net/doc/IS/se/1.1, 2006.

D. Nebert, U. Voges, and L. Bigagli. OGC® Catalogue Services 3.0 - General model. http://www.opengis.net/doc/IS/cat/3.0, 2016.

J. Nogueras-Iso, F. J. Zarazaga-Soria, R. Béjar, P.J. Álvarez, and P.R. Muro-Medrano. OGC® Catalog Services: a key element for the development of Spatial Data Infrastructures. *Computers & Geosciences*, 31(2):199–209, 2005.

A. Perego and M. Lutz. ISA Programme Location Core Vocabulary. *ISA Programme Location Core Vocabulary, Namespace Document*, 2015.

A. Perego, A. Friis-Christensen, and M. Lutz. GeoDCAT-AP: Use cases and open issues. In *Smart Descriptions & Smarter Vocabularies (SDSVoc)*, 2016.

A. Perego, V. Cetl, A. Friis-Christensen, and M. Lutz. GeoDCAT-AP: Representing geographic metadata by using the "DCAT application profile for data portals in Europe". In *Joint UNECE/UNGGIM Europe Workshop on Integrating Geospatial and Statistical Standards*, 2017.

Matthew Perry and John Herring. OGC® GeoSPARQL-A geographic query language for RDF data. Available online: http://www.opengeospatial.org/standards/geosparql (accessed 2 February 2016), 2010a.

Matthew Perry and John Herring. OGC® GeoSPARQL-A geographic query language for RDF data. Version 1.0. http://www.opengis.net/doc/IS/geosparql/1.0, 2010b.

Clemens Portele and Panagiotis (Peter) A. Vretanos. OGC® Web Feature Service 3.0 - Part 1: Core. Draft. Available online: https://cdn.rawgit.com/opengeospatial/WFS%5FFES/master/docs/17-069.html (accessed 2018-04-18), 2018.

U. Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer graphics and image processing*, 1(3):244–256, 1972.

schema.org. Schema.org. http://schema.org/, 2018.

M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis. 5G-enabled tactile internet. *IEEE Journal on Selected Areas in Communications*, 34 (3):460–473, 2016.

M. Sporny. HTML+RDFa 1.1 – Second Edition. http://www.w3.org/TR/2015/REC-html-rdfa-20150317/, 2015.

M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, and N. Lindström. JSON-LD 1.0: a JSON-based serialization for linked data, 2014.

J. Tandy, L. van den Brink, and P. Barnaghi. Spatial Data on the Web Best Practices. https://www.w3.org/TR/2017/NOTE-sdw-bp-20170928/, 2017.

Kerry Taylor and Ed Parsons. Where is everywhere: bringing location to the web. *IEEE Internet Computing*, 19(2):83–87, 2015.

W.R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1):234–240, 1970.

J. Tomkins and D. Lowe. Timeseries Profile of Observations and Measurements. Version 1.0. http://www.opengis.net/doc/IS/timeseries-profile-om/1.0, 2016.

European Union. Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), Directive 2007/2/EC of the European Parliament and of the Council, 14 March 2007. http://data.europa.eu/eli/dir/2007/2/oj, 2007.

L. van den Brink, P. Janssen, W. Quak, and J. Stoter. Linking spatial data: automated conversion of geo-information models and GML data to RDF. *International Journal of Spatial Data Infrastructures Research*, 9:59–85, 2014.

M. Visvalingam and J.D. Whyatt. Line generalisation by repeated elimination of points. *The Cartographic Journal*, 30(1):46–51, 1993.

P.A. Vretanos. OpenGIS Web Feature Service 2.0 Interface Standard. http://www.opengis.net/doc/IS/wfs/2.0, 2010.

W3C SPARQL Working Group. SPARQL 1.1 Overview. http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/, 2013.

Part VI

Discussion, Conclusion and Future work

Chapter 7

Developments since the publication of the articles

Since the previous chapters were published, some of them several years ago, there have been relevant developments. These will be discussed in this chapter; figure 7.1 gives an overview of the different developments and how they relate to each other.

7.1 3D standards

CityGML as an international 3D standard has gained acceptance over the past years. With its advanced possibilities for representing 3D geometries at different levels of detail as well as detailed semantics of all the parts of 3D city objects, the standard is increasingly used to model cities in 3D over the world. After the ADE extension method was published as an OGC best practice, several application domain extensions have been developed according to our method to support domain specific applications, such as the CityGML Utility Network ADE (Kutzner and Kolbe, 2016), the CityGML Energy ADE (Nouvel et al., 2015), The Land Administration Domain Model (LAD) CityGML ADE (Rönsdorff et al., 2014), and an ADE for immovable property taxation (Çağdaş, 2013).

The increased adoption of CityGML is aided by the development of tooling for validation of CityGML data. A CityGML Quality Interoperability Experiment (QI) was organized by OGC in 2015-2016 (Wagner and Ledoux, 2016), focussing on geometry validation, semantic validation, and validation of conformance requirements. Earlier work on the validation and repair of 3D geometries and other quality aspects of CityGML data (Kutzner and Kolbe, 2016; Ledoux, 2013; Wagner et al., 2013; Alam et al., 2014; Zhao et al., 2014)

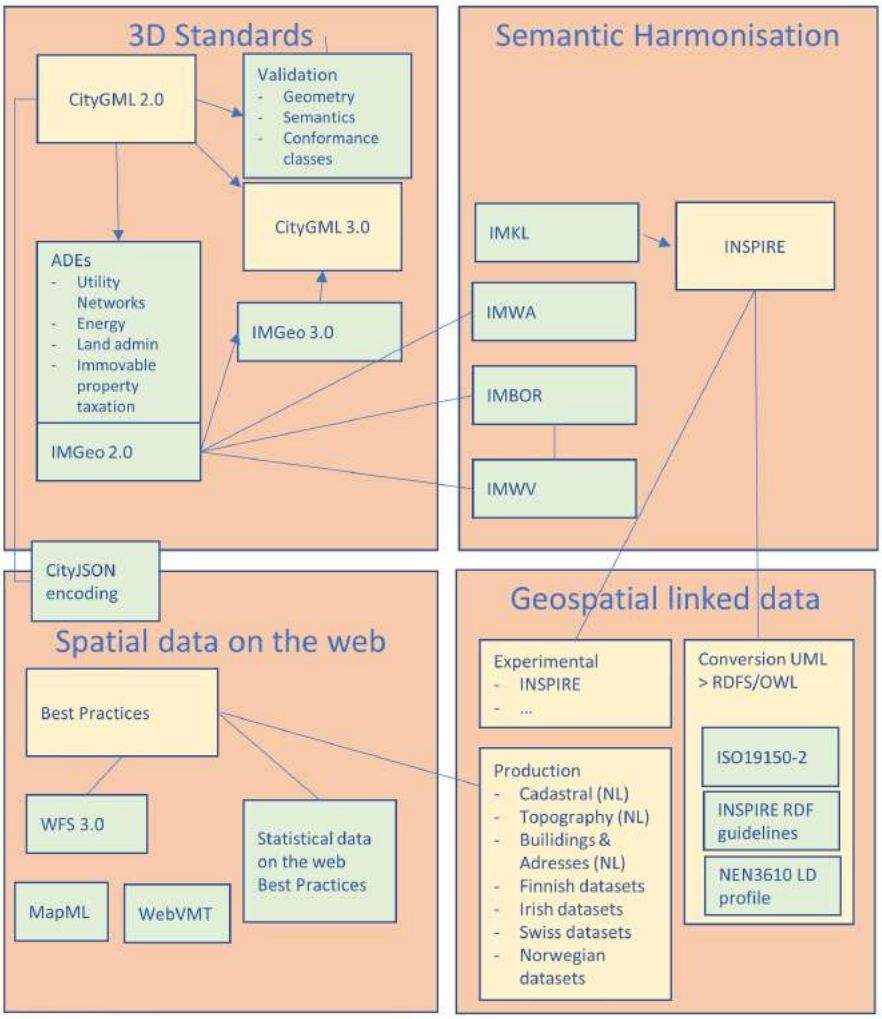


Figure 7.1: Overview of the developments since publication of the studies on 3D standards, semantic harmonisation, geospatial linked data and spatial data on the web.

was used in this QI. Among the results of the QI are common definitions of testing methods and test datasets for CityGML and 3D geospatial data in general.

The next version of CityGML (i.e. version 3.0) is under development. Proposed changes include a new Level Of Detail (LOD) concept (Biljecki et al., 2014; Löwner et al., 2016), explicit modelling of spaces and space boundaries, and an expansion of the semantic model with building-like non-building structures inspired by the Other Construction class in IMGeo and the INSPIRE Buildings theme. The new CityGML standard will include a normative conceptual UML model of all classes and properties, as well as a normative GML-based encoding.

Because conceptual model and encoding are separately defined, it is possible to define alternative encodings. An additional encoding of CityGML, based on JavaScript Object Notation (JSON) rather than GML, has already been developed. This encoding, CityJSON (Ledoux et al., 2017), could help adoption even further, especially as it provides a more lightweight format for exchanging and using both the geometries and semantics of 3D city models on the Web.

The new CityGML release will impact IMGeo in several ways. Firstly, a study will be necessary to see how IMGeo can adopt the new LOD concept. Secondly, since non-building building like structures become part of the CityGML standard, these classes can in future be redefined as extensions of the new corresponding CityGML classes.

IMGeo 3.0 is also under development. One of the changes will be realignment with a new version of its cousin IMBAG (information model buildings and addresses). There are also change proposals related to 3D, such as a proposal to make it possible to include different surface geometries for the footprint and the top view (or view from below, in the case of subsurface structures) of buildings including a relative height indication. Another proposal is to add closure surfaces to 2D IMGeo, in order to align the 2D representation of topographic objects with the representation in 3D.

7.2 Semantic harmonisation

On the topic of semantic harmonisation of geospatial data, there are also some relevant developments that are worth mentioning. Several harmonisation issues that were identified in the case study, and described in Chapter 4, have since been addressed in updates of the involved standards. For example, the Dutch national Water domain model IMWA has been harmonised with IMGeo.

In addition, new areas of harmonisation have been addressed beyond the case study. For example, the national Utility networks domain model IMKL (*Informatie model Kabels en Leidingen*) has been harmonised with the corresponding INSPIRE theme.

Based on the findings described in Chapter 4, harmonisation with existing models is explicitly considered when a new domain model is designed; examples of this include the domain model for public space management, IMBOR (*Informatie model beheer en openbare ruimte*), which was developed as an extension of IMGeo; and the domain model for roads and traffic, IMWV (*Informatie model wegen en verkeer*) which is currently under development and will include the definition of its relations with other domain models, such as IMGeo and IMBOR.

Semantic harmonisation is often still an ignored aspect of data reuse. Hansen et al. (2017), who describe a case where an SDI will be set up for sharing maritime spatial planning data, conclude that semantic harmonisation is one of the important requirements in an SDI. However, it is also a complex and time-consuming task, which requires human data modellers to solve. Although in the Netherlands semantic harmonization between data models is now steadily progressing, it is indeed a complex, careful and long-winded process. In other countries I did not find much evidence of it being addressed systematically.

An interesting recent approach by Jiang et al. (2017) uses human input for semantic harmonization in an automated way. Their method combines large scale user search histories and click streams with similarity calculation methods for discovering alignments between ontologies to obtain a better determination of semantic relationships.

Ultimately, the goal of this approach is to improve data discovery by making it easy to find relevant data without having to know the correct search terms to use. However, for different use cases, such as using and combining datasets in day-to-day work processes (e.g. IMWA and IMGeo for specific water management tasks as well as general public space management), humans may still be needed to solve cases where data have similar semantics, but not quite the same.

In another publication Yang et al. (2017) describe a different case, where the goal is to make different classifications semantically interoperable in order to enable more complex analysis across datasets using different land cover classifications. In this case semantic interpretation is necessary to establish the degree of similarity between terms, as numerous land cover classifications exist, created by different scientific communities with different aims. Yang et al. (2017) apply an automated technique to determine semantic similarity, but conclude that expert knowledge is still needed to resolve semantic

inconsistencies between land cover classes.

In conclusion, both developments in automatic matching and involving expert knowledge in improving semantic interoperability in the geospatial domain are needed and ongoing.

7.3 Geospatial linked data

Since the work described in Chapter 5, there has been a lot of development on linked geospatial (governmental) data. For example, Patroumpas et al. (2015) describe a method to build an abstraction layer on top of the INSPIRE infrastructure based on linked data concepts. They focus on the data level, basing their data model on OGC GeoSPARQL, a standard for modelling and querying geospatial linked data; and developed a working GML to RDF conversion tool, but did not address UML to OWL conversion.

Their conclusion confirms the first part of mine: conversion to RDF is straightforward. This means that making geospatial data like INSPIRE data accessible and discoverable as linked data takes relatively little effort. However, they offer no new conclusions on the difficulties of UML to OWL conversion.

Several linked spatial datasets have been published since our paper on linked data, including governmental, non-experimental linked spatial datasets from the Netherlands such as Cadastral data (Kadaster, 2018); Buildings and addresses (Kadaster, a); Topographic data (Kadaster, b); Ireland (Debruyne et al., 2017), Switzerland (data porta, 2018), Finland (data Finland, 2018), and Norway (Shi et al., 2017). Although European countries seem to be most active on this topic, Australia also has an active Australian Government linked data working group and several spatial linked datasets published (government, 2018).

The method I describe in chapter 5 became input for the development of a method to derive OWL vocabularies from the INSPIRE UML models (Tirry and Vandenbroucke, 2014), in which I was involved. Later a set of INSPIRE RDF Guidelines (ISA Action ARe3NA, 2017) were developed, based on this and other input. The INSPIRE RDF Guidelines constitute an optional encoding rule, aiming at giving guidance to EU countries looking at RDF for spatial data publication, but not taking away the requirement to publish INSPIRE data as specified in the Data Specifications of the different INSPIRE themes. Clause 9 of these specifications gives the requirements for encodings, which is usually GML, although alternative encodings are allowed. Thus INSPIRE explicitly allows linked data as a second, alternative publication method.

Automated conversion from GML application schemas specified in UML to OWL remains problematic. A current project at Geonovum (Geonovum, 2018), aimed at developing a linked data profile for NEN 3610, has lead to some preliminary findings regarding OWL ontologies created from UML models. These findings are based on several national spatial datasets being published as linked data (see earlier reference to Dutch linked spatial data sets), including OWL ontologies, which were developed by hand.

According to the preliminary findings, a 'correct' ontology can be automatically derived from a UML model—'correct' meaning adhering to the relevant standards. However, a 'good' ontology, meaning one that adheres to common OWL design patterns and expresses the knowledge domain in a useful way, cannot. This requires an OWL modelling expert, who can take into account the fundamental differences between UML and OWL, such as the ones described in my paper: the closed versus open world assumption, the usage of external, existing vocabularies which is common in OWL, modelling constructs which exist in UML but not in OWL and vice versa; and other peculiarities.

Another fundamental difference which surfaced explicitly in the current Geonovum project is that when compared to database normalisation, an OWL ontology would be most similar to the 6th normal form, i.e. as close as possible to the real world it is trying to model. In contrast, most information models describe information systems or information exchange schemas and are thus more denormalized: more adapted to the system, but less close to the real world. The current assumption within the Geonovum project is that normalizing a denormalized model is only possible if the semantics of the data are known—for which a human is needed.

7.4 Web of data

Work on the Best Practices for publishing spatial data on the web is the most recent part of this thesis and thus, there are fewer developments since publication. The paper was accepted for publication in January 2018, after the Best Practice itself was published in September 2017 as a W3C Note and OGC Best Practice (Tandy et al., 2017). Follow up work on the Best Practices document has been started in the W3C/OGC Spatial Data on the Web Interest Group (W3C, 2018), with me as one of its chairs.

The Best Practice will be updated based on implementation reports from practice; also, new work items of the interest group include maintenance of the Semantic Sensor Network Ontology (SSN) (Haller et al., 2017) and a Statistical Data on the Web Best Practices document. In addition, the

Interest Group finds, tracks and supports new developments related to both geospatial data and the Web as candidates for standardisation within OGC and/or W3C; e.g. CityJSON, Map Markup Language (MapML), and Video geotagging for maps.

Bibliography

Nazmul Alam, Detlev Wagner, Mark Wewetzer, Julius von Falkenhausen, Volker Coors, and Margitta Pries. Towards automatic validation and healing of CityGML models for geometric and semantic consistency. In *Innovations in 3D Geo-Information Sciences*, pages 77–91. Springer, 2014.

Filip Biljecki, Hugo Ledoux, Jantien Stoter, and Junqiao Zhao. Formalisation of the level of detail in 3D city modelling. *Computers, Environment and Urban Systems*, 48:1–15, 2014.

Volkan Çağdaş. An Application Domain Extension to CityGML for immovable property taxation: A Turkish case study. *International Journal of Applied Earth Observation and Geoinformation*, 21:545–555, 2013.

Linked data Finland. Finnish linked data. http://www.ldf.fi/datasets.html (accessed 2018-04-18), 2018.

European data porta. Swiss linked data. https://www.europeandataportal.eu/nl/news/swiss-geospatial-data-now-available-linked-data (accessed 2018-04-18), 2018.

Christophe Debruyne, Alan Meehan, Éamonn Clinton, Lorraine McNerney, Atul Nautiyal, Peter Lavin, and Declan O'Sullivan. Ireland's Authoritative Geospatial Linked Data. In *International Semantic Web Conference*, pages 66–74. Springer, 2017.

Geonovum. Geonovum project NEN3610-Linkeddata. https://github.com/Geonovum/NEN3610-Linkeddata (accessed 2018-04-18), 2018.

Australian government. Australian government linked data working group. http://linked.data.gov.au/ (accessed 2018-04-18), 2018.

Armin Haller, Krzysztof Janowicz, Simon Cox, Danh Le Phuoc, Kerry Taylor, and Maxime Lefrançois. Semantic Sensor Network Ontology. Available online: https://www.w3.org/TR/vocab-ssn/ (last accessed 2018-04-18), 2017.

Henning Sten Hansen, Ida Maria Reiter, and Lise Schrøder. A System Architecture for a Transnational Data Infrastructure Supporting Maritime Spatial Planning. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 158–172. Springer, 2017.

ISA Action ARe3NA. Guidelines for the RDF encoding of spatial data. http://inspire-eu-rdf.github.io/inspire-rdf-guidelines, 2017.

Yongyao Jiang, Yun Li, Chaowei Yang, Kai Liu, Edward M Armstrong, Thomas Huang, David F Moroni, and Christopher J Finch. A comprehensive methodology for discovering semantic relationships among geospatial vocabularies using oceanographic data discovery as an example. *International Journal of Geographical Information Science*, 31(11):2310–2328, 2017.

Kadaster. Buildings and addresses. https://bag.basisregistraties.overheid.nl (accessed 2018-04-18), 2018a.

Kadaster. Cadastral data. https://brk.basisregistraties.overheid.nl (accessed 2018-04-18), 2018.

Kadaster. Topographic data. https://brt.basisregistraties.overheid.nl (accessed 2018-04-18), 2018b.

Tatjana Kutzner and Thomas H Kolbe. Extending semantic 3D city models by supply and disposal networks for analysing the urban supply situation. In T. P. Kersten, editor, *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e.V. Wissenschaftlich-Technische Jahrestagung der DGPF*, pages 382–394, 2016.

H. Ledoux, C. Nagel, and K. Arroyo. CityJSON Specifications. Available online: http://www.cityjson.org/en/0.5/specs/ (last accessed 2018-03-09), 2017.

Hugo Ledoux. On the validation of solids represented with the international standards for geographic information. *Computer-Aided Civil and Infrastructure Engineering*, 28(9):693–706, 2013.

Marc-O Löwner, Gerhard Gröger, Joachim Benner, F Biljecki, and Claus Nagel. Proposal for a new lod and multi-representation concept for Citygml. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:3–12, 2016.

Romain Nouvel, Robert Kaden, Jean-Marie Bahu, Jerome Kaempf, Piergiorgio Cipriano, Moritz Lauster, Joachim Benner, Esteban Munoz, Olivier Tournaire, and Egbert Casper. Genesis of the citygml energy ADE. In *Proceedings of International Conference CISBAT 2015 Future Buildings and Districts Sustainability from Nano to Urban Scale*, volume EPFL-CONF-213436, pages 931–936, 2015.

Kostas Patroumpas, Nikos Georgomanolis, Thodoris Stratiotis, Michalis Alexakis, and Spiros Athanasiou. Exposing NSPIRE on the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:53–62, 2015.

C Rönsdorff, D Wilson, and JE Stoter. Integration of land administration domain model with CityGML for 3D cadastre. In *Proceedings 4th International Workshop on 3D Cadastres, 9-11 November 2014, Dubai, United Arab Emirates*. International Federation of Surveyors (FIG), 2014.

Ling Shi, Dina Sukhobok, Nikolay Nikolov, and Dumitru Roman. Norwegian State of Estate Report as Linked Open Data. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 445–462. Springer, 2017.

J. Tandy, L. van den Brink, and P. Barnaghi. Spatial Data on the Web Best Practices. https://www.w3.org/TR/2017/NOTE-sdw-bp-20170928/, 2017.

Diederik Tirry and Danny Vandenbroucke. Guidelines on methodologies for the creation of RDF vocabularies representing the INSPIRE data models and the transformation of INSPIRE data into RDF, 2014.

W3C. Spatial Data on the Web Interest Group. https://www.w3.org/2017/sdwig/ (accessed 2018-04-18), 2018.

D. Wagner and H. Ledoux. Ogc® CityGML Quality Interoperability Experiment, 2016.

Detlev Wagner, Mark Wewetzer, Jürgen Bogdahn, Nazmul Alam, Margitta Pries, and Volker Coors. Geometric-semantical consistency validation of CityGML models. In *Progress and new trends in 3D geoinformation sciences*, pages 171–192. Springer, 2013.

Hui Yang, Songnian Li, Jun Chen, Xiaolu Zhang, and Shishuo Xu. The standardization and harmonization of land cover classification systems towards harmonized datasets: a review. *ISPRS International Journal of Geo-Information*, 6(5):154, 2017.

Junqiao Zhao, Jantien Stoter, and Hugo Ledoux. A framework for the automatic geometric repair of CityGML models. In *Cartography from pole to pole*, pages 187–202. Springer, 2014.

Chapter 8

Conclusions and future work

The research described in this thesis spans about six years of work, starting with the definition of a national standard for 3D topography, and ending with a broad effort to define best practices for publication of geospatial data as part of the web of data. Not surprisingly, since the earlier work was published new developments have occurred, as described in the previous chapter. My work fits in with and is a part of the developments I have described, contributing to the amount of reuse of spatial data via the web.

Although a great deal of progress has been made, there is still work to be done to improve reuse of geospatial data across communities via the web even further. An evolution — or even a revolution — of the way geospatial data are disseminated is occurring and will be ongoing for the coming years, opening up countless possibilities for smart, innovative data-driven applications. I will describe some of these possibilities later in this chapter; but first, I will summarise the main conclusions.

8.1 Main Conclusions

The main question of my research was:

> *How to reuse geospatial data, from different, heterogeneous sources, via the web across communities?*

I formulated four questions to cover different aspects of the problem to reuse geospatial data occurring in practice.

1 *How to define a national standard for large-scale topographic objects in 3D for wide re-use of once collected 3D data to solve current ad hoc acquisition and use of such data?*

One of the main causes for ad hoc and project-based collection of 3D data is the lack of a 3D standard. Instead of creating a new standard, the established Dutch 2D standardisation framework was studied for extension into 3D while aligning to the international standardisation developments (van den Brink et al., 2013a). A national standard for 2D, large scale topographic already existed: IMGeo. A comparison of the main international 3D standards showed that OGC CityGML was the best candidate for 3D data reuse, mainly because of its extensive 3D geometry support and its rich semantic model. The alignment of IMGeo with CityGML proved to be possible with little adjustment, resulting in a nationwide 3D standard for geospatial data, covering all topographic classes, and extending the OGC CityGML standard. The extension method was the basis for a generic framework for extending CityGML (van den Brink et al., 2013b) which became an OGC best practice (van den Brink et al., 2014b). The practice of extending CityGML has become more widespread, as is documented at `https://www.citygml.org/ade/` and described in Biljecki et al. (2018). Many of these ADEs use the guidelines described in this best practice.

Establishing a national 3D standard has improved 3D data re-use, although not as much as expected. CityGML-IMGeo is supported in an open source tool for the creation of 3D data from 2D input (which can be 2D IMGeo data) in combination with point cloud data, i.e. 3dfier (`https://github.com/tudelft3d/3dfier`). Kadaster is using 3dfier to create a 3D dataset for the whole country of the Netherlands, compliant to the 3D CityGML-IMGeo data model. This work has shown that CityGML (and GML in general) is verbose and complete, but also sometimes difficult to implement. Future research is needed to develop CityGML encodings that are lighter and easier to work with. See also future work.

2 *How can semantic interoperability between different kinds of geospatial datasets that have been created for different purposes best be achieved?*

In general, a form of semantic interoperability is often achieved by matching the ontologies which describe the semantics of the domains in question; largely in an automated way. However, my goal was to improve the actual sharing and reuse of spatial data, and therefore I needed to find areas where similarities existed between the data models of related domains, but semantic inconsistencies prevented reuse. These problems could only be solved by deciding on the best way to adjust the ontologies in question; i.e. human interpretation is required. The task of finding potential harmonisation issues and presenting these to human experts was made computer-assisted. Human experts decided in a next step the cases where harmonisation would be most

beneficial, and assessed how each semantic problem could best be solved. The combination of human domain experts and tooling which helps them in discovering matches between domain models made it possible to include a high number of concepts in the study to find semantic overlaps (van den Brink et al., 2017). This community-driven approach has led to changes in several national semantic standards and related work processes resulting in improved semantic interoperability between existing geospatial data sets.

3 *How to apply the Linked Data paradigm to disseminate geospatial data outside the traditional geospatial data sector?*

I studied Linked Data as a paradigm that might be employed to disseminate both spatial semantics and spatial data to new groups of data users (van den Brink et al., 2014a). To automatically derive linked data from GML data in a generic way, several steps were necessary. The first step, the technical conversion of standard geospatial data encodings such as GML to RDF proved to be straightforward. The second step, choosing the right vocabulary for geometries, was more difficult as there are several standards to choose from. Since most national datasets, including our test data, use a local coordinate system that is specific to a particular country or region, the best option is OGC GeoSPARQL, which offers support for a complete range of geometry types. The third step, deciding on a URI strategy for the creation of stable, scalable URIs for spatial objects, required careful consideration. The main criterion for linked data URIs is persistence: the URIs must continue to work over a long period of time, notwithstanding changes like the termination or renaming of the organisation that minted them. For that reason a neutral domain name, with no organisation name in it, is preferred. A URI pattern was designed based on the Dutch URI strategy (Overbeek and van den Brink, 2013) which in turn is based on ISA (2012), taking into account persistence, scalability, intelligibility, trust, machine-readability and human-readability.

From my study, it can be concluded that it is to a certain extent possible to automatically derive OWL ontologies from UML class models, which are used in the Dutch SDI for describing semantics. ISO 19150-2 describes the mapping from UML to OWL for geographic information models. However, I found several issues with this mapping. I chose to focus on one of these issues: the mapping rules converted the UML diagram to an OWL ontology without linking the terms in this vocabulary to terms in existing OWL vocabularies, which is considered best practice in the linked data community. This can be solved by annotating the UML model with links to existing OWL vocabulary terms, and using these annotations when automatically deriving the

OWL ontology from the UML model. The resulting OWL ontology may need manual improvement to adjust for the different underlying paradigm of OWL compared to UML (open world assumption versus closed world assumption) and for awkward modelling constructs resulting from limitations of UML. A consequence of the open world assumption is that restrictions, both explicitly expressed and inherently present in the UML model, cannot be exactly reproduced in OWL because of its basic assumption that anything not explicitly stated is unknown. An example of awkward modelling because of UML limitations is that in UML multiple inheritance is usually avoided in geospatial information models because of technical implications on the GML/XML data exchange level, while in OWL and in RDF data multiple inheritance is not problematic. The workaround applied in UML should therefore not be translated to the OWL version of the model. Automated conversion of UML to OWL thus remains problematic: a 'correct' ontology can be automatically derived from a UML model—'correct' meaning adhering to the relevant standards. But a 'good' ontology, meaning one that adheres to common OWL design patterns and expresses the real-world knowledge domain as closely as possible, cannot. This requires manual adjustments by an OWL modelling expert, who can take into account the fundamental differences between UML and OWL.

However, even taking these limitations into account, these results show that existing geospatial data can now be published on a large scale as semantically rich linked data. Linked data publication can be integrated in an SDI, thus taking advantage of a large amount of geospatial data already being made available in national geoportals or, for example, in the context of INSPIRE.

4 *How to apply general Web based principles to improve the discoverability and accessibility of spatial data?*

Linked data is not the only way to publish data on the web; more mainstream and well-known Web standards and technologies are in place and are preferred by many web data users, because of their simplicity and better integration in web browsers. These general Web standards and technologies can be applied to spatial data publication to disseminate the data to a wider audience. During my research, we described how to do this in a set of best practices for publishing spatial data on the web, distilled from good current practices (Tandy et al., 2017; van den Brink et al., 2018). To summarise the key best practices, web principles should be applied to both spatial data and dataset descriptions (metadata) by their publication at a stable URI, in a format that can be indexed by search engines, and with links to other

data on the Web – thus becoming part of the web of data. Furthermore, in order to be useful to a wider audience, geospatial data should be published on the web using a global coordinate system for geometries. The geometries themselves can be encoded and serialized in several ways, depending on the target audiences. Finally, spatial data should be published through easily accessible Application Programming Interfaces (APIs). When implemented, these guidelines for publishing spatial data on the web make it easier to discover, interpret and use geospatial data for data users in general – not just geospatial experts.

8.2 Limitations of the research

Most of the research described in this thesis was, although the subject of study were international standards, limited in its orientation towards the Netherlands; the exception being the last study, about the spatial data on the web best practices. The Dutch culture of standardisation and open data is not exactly the same as in other countries; therefore, results of my research cannot be extrapolated to other countries without reserve. In the Netherlands, the social process of reaching consensus (the "polder model") is very dominant, also in standards-making: standards are not imposed top-down; every involved party has to agree to them. This situation is different in other countries. On the other hand, a lot of other countries are comparable to the Netherlands in that they have an SDI in place; all European countries provide their geospatial data in the same way in order to be INSPIRE compliant; and a growing number of countries have 3D data and linked data available.

My research was in danger of being limited to theoretical findings. In the context in which my research was carried out, there was no data available, only data models. This was solved by arranging the research as a part of pilots in which companies and governmental organisations worked together with researchers.

Another limitation of my research is related to semantic harmonisation specifically. While creating a semantic mapping between standards is often possible, it is much harder to change existing work processes. Ajusting a standard in order to align it semantically to a related standard may cost an organisation that works with the standard a lot of money, because existing work processes need to change. This is something a researcher cannot influence.

Finally, a usual limitation of standards work was also encountered by me. Standards work is often arranged in working groups, and the products of the working group can only incorporate as much as the group members

are capable of producing. If certain knowledge or skills are not represented within the group, they cannot be incorporated in the outcome. For example, in the Spatial data on the Web Best Practices there are no guidelines about non-geospatial spatial data, although non-geospatial use cases were brought to the working group and were included in the Spatial Data on the Web Use Cases and Requirements (UCR) (Knibbe and Llaves, 2016), since there were no active participants in the working group who had expertise with spatial data other than geospatial.

8.3 Meaning of my work beyond the geospatial domain

Although geo-information is a specific research domain, several of the results of my study can be generalised and applied beyond it as they are not specifically dealing with geospatial aspects of data. I have shown how a standard, which is designed to be applicable in a specific context, for example a specific country and use case, can nevertheless be created based on a general international standard. I have also shown how semantic standards, which are already being used in practical use cases, can be harmonised by engaging their custodians in the tooling-aided mapping work. I have contributed to solving the puzzle of creating linked data ontologies based on UML models by showing how to enrich the UML with links to well-known linked data vocabularies. The spatial data on the web best practices contain not only guidance on geospatial aspects of data publishing, but also some on the publication of general data on the web.

8.4 Future work

Implementation and adoption of standards takes time (usually several years), making it difficult to evaluate if my solutions to some of the interoperability problems within the geospatial domain actually worked. Without having done extensive analysis, my observation is that the uptake of GML and CityGML, for example, has been somewhat disappointing. Most software systems which handle geospatial data either do not support these standards at all, only in a limited way, or only while requiring complex actions by the user. Within the geospatial domain, the interoperability problems I describe have thus been only partly solved, since some of the solutions depended on CityGML. Less complex data encodings, such as JSON, may improve adoption of geospatial data exchange standards — among new (data) users such

as web developers, but also in the geospatial community itself. This is one of the reasons why I shifted the focus of my research to general web standards.

The added value of geospatial data lies in its ability to play the role of linking pin between different datasets, thus allowing data integration based on location. Decades of work and a large, international standardisation effort have resulted in a lot of geospatial data now being available for reuse on the web. However, because the semantics, formats, and dissemination methods used are still very specific to the geospatial domain, a lot of this data is only findable and usable by geospatial ICT experts. And yet, there is a huge potential for the use of geospatial data in other domains. To achieve the wide reuse of geospatial data, i.e. beyond the geospatial domain, data users need to be able to find, access, and use geospatial data on the web without expert knowledge of geospatial standards. Geospatial data needs to become a common part of the web of data.

The Spatial Data on the Web Best Practices document gives guidance about how this can be achieved. Once these best practices are widely implemented, I believe a world of possibilities will open up. However, there is yet work to be done. First, after publication of a document like this, practical steps are needed to help its adoption and implementation. Besides proper communication to make sure a standard is well known, tools like conformance tests and automated validators, are helpfull implementation aids. Both adoption and implementation could also be promoted by the collection of (early) examples of web content containing spatial data, where the best practices are followed. From the implementation of the best practices, inevitably improvements and new developments will arise. These should be fed back into the document, thereby adding to it and keeping it up to date. Each recommendation in the Spatial Data on the Web Best Practices starts with a description of its rationale and intended outcome, followed by a description of the possible approach for implementation. The latter contains concrete examples and guidelines taken from practice and is therefore expected to need maintenance as the web continues to evolve, while the more general parts of the best practices can probably remain stable.

A major topic for future work is how the SDI as a standardised framework for geospatial data dissemination can evolve to apply and include the best practices for publishing spatial data on the web. This requires careful study, since there are many questions to consider and many possible solutions. For example, one proposal is to add an extra "proxy" layer on top of existing OGC web services that will publish geospatial dataset descriptions and object data on the fly, using general web standards. This is implemented in the experimental tool ldproxy (interactive instruments, 2018). Another approach is to develop major, best practice compliant revisions of OGC web service

standards, like is being done in the case of WFS 3.0 (Portele and Vretanos, 2018).

However, an SDI is more than just web services. In addition, there is a need for more accessible encodings for geospatial web data and a geometry format for the Web which addresses all requirements – a format similar to GeoJSON, but with support for different coordinate reference systems (CRS) and 3D geometry. Such encodings are already emerging: for example, CityJSON (Ledoux et al., 2017) is a proposal for a JSON encoding of OGC CityGML. This is a good candidate for standardisation and could play a role in the revised SDI, but also on the web in general, outside the geospatial domain. The need for working with 3D data in web browsers will probably grow in the next years, for example with the growth of augmented reality applications.

Moreover, encodings in the sense of technical exchange formats are not enough: there is a continuing need for vocabularies that describe or even formally define the semantics of the data. Agreed and shared vocabularies are an important factor in interoperability. These were always part of the SDI in the form of UML information models. In an SDI following web principles the semantics of data should be readily available on the web, and should therefore be described in vocabularies using web standards i.e. Simple Knowledge Organization System (SKOS), RDFS and/or OWL. While simple encodings usually name the elements of a data format, they do not describe or define their meaning in a machine-readable way. However, simple encodings based on, for example, JSON, can be linked to semantics defined in a web-based machine-readable vocabulary by using JSON-LD (Sporny et al., 2014). Some standardized vocabularies and ontologies for geospatial semantics are already available: the GeoSPARQL ontology and the Semantic Sensor Network ontology (SSN) (Haller et al., 2018), while the Time Ontology in OWL is also very relevant for geospatial data. More work is necessary to express additional geospatial semantics in ontologies making them part of the web of data. For example, a City ontology based on the data model of CityGML would be beneficial as this could be linked to CityJSON to provide explicit, machine readable semantics for 3D city models on the web.

Because the semantics are currently only available in UML, transformation rules are needed for creating ontologies on the basis of UML information models, taking into account the fundamental differences between UML and RDF. Once these differences are well understood, UML to RDF conversion is solvable with a combined approach of first automatic conversion followed by manual adjustments to create a good knowledge model of the domain. SKOS, RDFS and OWL can replace UML now that a new standard, SHACL is available, which provides a constraint language and can thus form the

mechanism for a basis of standardized and structured data within domains. Thus, RDFS and OWL can be seen as no longer complementary, but a replacement — in combination with SHACL, which has become available since writing my paper "linking geospatial data". RDFS/OWL are capable of modeling the real world much more closely than UML. However, good visualisation methods for RDFS/OWL models are still needed as explanatory aids for domain experts in order to engage them in the creation and use of semantic standards.

Another important part of SDIs is metadata: dataset descriptions which allow users to find datasets they are interested in, similar to searching for books in a library's catalogue. The need for metadata will lessen when the data itself is indexable, i.e. the use case of searching for data will be partly addressed by publishing the data itself using web principles. However, there is still a need for metadata in order to determine if a dataset is fit for a certain use. Important metadata elements for this use case are described in Best Practice 13 (Tandy et al., 2017). However, a web friendly metadata standard including these elements is not yet fully developed.

All this work assumes that applying general web standards will enlarge the reuse of spatial data. I have not addressed the obvious question: to what level will this actually happen? Further research is needed to answer this; a definitive answer is not possible until implementation and adoption of the best practices is achieved, and significant amounts of spatial data are available in the way I described. When published in accordance to the best practices, is spatial data actually indexed by search engines? And are web data users actually able to find, access, and use the data more easily, when compared to spatial data only available through an OGC-standards compliant SDI?

If this is the case, this will open up new ways for the reuse of geospatial data via the web across communities. While the geospatial web of data is growing, new technological developments will create unforeseen possibilities for its use. One such development is the availability of ever greater bandwith (5G) combined with more efficient handling of heavy dataloads and tasks in web browsers. New standards such as WebAssembly (Rossberg, 2018) will support this. Thus it will, for example, become possible in the near future to work with voluminous, 3D geospatial data natively in a browser — perhaps supported by CityJSON. Augmented and virtual reality applications, using 3D and other geospatial data, will be implemented in web browsers.

Another important development is the growing availability on the web of voluminous data created by different kinds of sensors. The data from these sensors provide information about numerous aspects of the world we live in: the quality of the air, water and soil; the physical objects that are around

us; windspeed, air temperature, tremors in the earth, and so on. Once such data, preferably combined and enhanced with semantically rich spatial data, becomes widely available and easily discoverable and usable, linked into the web of data, this will support the rise of smart energy grids, smart cities, smart homes, smart cars, smart farming; the richer the data that is available, the more smart applications people can think of and implement, assisting and improving our daily lives.

Bibliography

Filip Biljecki, Kavisha Kumar, and Claus Nagel. CityGML Application Domain Extension (ADE): overview of developments. *Open Geospatial Data, Software and Standards*, 2018.

Armin Haller, Krzysztof Janowicz, Simon JD Cox, Maxime Lefrançois, Kerry Taylor, Danh Le Phuoc, Joshua Lieberman, Raúl García-Castro, Rob Atkinson, and Claus Stadler. The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation. *Semantic Web Journal*, Pre-press:1—24, 2018.

interactive instruments. ldproxy. https://github.com/interactive-instruments/ldproxy (accessed 2018-04-18), 2018.

ISA. D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC. Available online: https://joinup.ec.europa.eu/sites/default/files/D7.1.3 - Study on persistent URIs%5F0.pdf (accessed 11 February 2014), 2012.

F. Knibbe and A. Llaves. Spatial data on the Web use cases & requirements. https://www.w3.org/TR/2016/NOTE-sdw-ucr-20161025/, 2016.

H. Ledoux, C. Nagel, and K. Arroyo. CityJSON Specifications. Available online: http://www.cityjson.org/en/0.5/specs/ (last accessed 2018-03-09), 2017.

H Overbeek and L van den Brink. Towards a national URI–Strategy for Linked Data of the Dutch public sector. Available online: http://www.pilod.nl/wiki/Bestand:D1-2013-09-19%5FTowards%5Fa%5FNL%5FURI%5FStrategy.pdf (accessed 3 June 2014), 2013.

Clemens Portele and Panagiotis (Peter) A. Vretanos. OGC® Web Feature Service 3.0 - Part 1: Core. Draft. Available online:

https://cdn.rawgit.com/opengeospatial/WFS%5FFES/master/docs/17-069.html (accessed 2018-04-18), 2018.

Andreas Rossberg. WebAssembly Core Specification. Available online: https://www.w3.org/TR/2018/WD-wasm-core-1-20180904/, 2018.

M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, and N. Lindström. JSON-LD 1.0: a JSON-based serialization for linked data, 2014.

J. Tandy, L. van den Brink, and P. Barnaghi. Spatial Data on the Web Best Practices. https://www.w3.org/TR/2017/NOTE-sdw-bp-20170928/, 2017.

L. van den Brink, P. Janssen, W. Quak, and J. Stoter. Linking spatial data: automated conversion of geo-information models and GML data to RDF. *International Journal of Spatial Data Infrastructures Research*, 9:59–85, 2014a.

Linda van den Brink, Jantien Stoter, and Sisi Zlatanova. Establishing a national standard for 3D topographic data compliant to CityGML. *International Journal of Geographical Information Science*, 27(1):92–113, 2013a.

Linda van den Brink, Jantien Stoter, and Sisi Zlatanova. UML-Based Approach to Developing a CityGML Application Domain Extension. *Transactions in GIS*, 17(6):920–942, 2013b.

Linda van den Brink, Jantien Stoter, and Sisi Zlatanova. Modeling an application domain extension of CityGML in UML. Available online: https://portal.opengeospatial.org/files/?artifact%5Fid=49000 (last accessed 2018-04-18), 2014b.

Linda van den Brink, Paul Janssen, Wilko Quak, and Jantien Stoter. Towards a high level of semantic harmonisation in the geospatial domain. *Computers, Environment and Urban Systems*, 62:233–242, 2017.

Linda van den Brink, Payam Barnaghi, Jeremy Tandy, Ghislain Atemezing, Rob Atkinson, Byron Cochrane, Yasmin Fathy, Raúl Garcia Castro, Armin Haller, Andreas Harth, Krzysztof Janowicz, Sefki Kolozali, Bart van Leeuwen, Maxime Lefrançois, Josh Lieberman, Andrea Perego, Danh Le-Phuoc, Bill Roberts, Kerry Taylor, and Raphaël Troncy. Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web. *Semantic Web Journal*, Pre-press:1—20, 2018.

Abstract

Geospatial data is an increasingly important information asset for decision-making, from simple every day decisions like where to park your car, to national and international policy on topics like infrastructure and environment. Because of the location aspect, geospatial data is often the linking pin between different datasets and therefore important for data integration. A lot of geospatial data is created, for example, as part of governmental processes and nowadays, also disseminated as open data, traditionally through "Spatial data infrastructures" (SDIs).

There is a lot of potential for reusing this data in other domains than the domain and use case for which it was originally created. My main research question was: "How to reuse geospatial data, from different, heterogeneous sources, via the web across communities?" Several aspects of data dissemination must be addressed before open data is actually in a good position for getting reused. These aspects have been coined the "FAIR principles": findability, accessibility, interoperability, and reusability.

A general foundation of my work is the common knowledge in the geospatial domain that interoperability between systems is required to make reuse of data possible, and standards are able to realise this interoperability. Based on this, I have addressed several different problem areas where the potential to reuse geospatial data was there, but hampered in some way. These problems are introduced in chapter 1. All the research was exploratory in nature; the methodology a combination of desk study, analysis, literature study and experimentation.

Chapters 2 and 3 deal with the lack of a standard for describing three-dimensional (3D) geospatial data in the case of the Netherlands, hindering the reuse of 3D data. To solve this, a national standard for two-dimensional (2D) geospatial, topographic data, Information model Geography (IMGeo), was combined with an international 3D standard, CityGML. Both standards describe topographic objects that represent physical objects in the real world and are largely similar.

Chapter 2 describes how CityGML was selected as the basis for a na-

tional 3D standard; how IMGeo was semantically aligned with it; how it was formally defined as an extension of CityGML on the level of classes, properties and code lists; and how other interoperability aspects were addressed: geometry, topological structure, and reference system.

Chapter 3 describes in more detail how IMGeo was formally defined as an extension of CityGML. This addresses technical information modelling issues, related to the use of Unified Modeling Language (UML) and the specific extension mechanism defined by CityGML. Based on the IMGeo case, I defined a model-driven framework for developing CityGML application domain extensions.

In the case of IMGeo and CityGML, the semantic harmonisation, i.e. the alignment of concepts defined in both standards, was relatively straightforward. Both standards describe the same kinds of things; as a result, most IMGeo classes can be said to be the same as or a subclass of a class in CityGML. This is not always the case. Independently developed domain models model similar concepts in different ways, which makes reuse of data in other domains difficult.

This problem of semantic harmonisation is addressed in chapter 4. Different Dutch standards were examined to discover areas where they overlap and where semantic harmonisation can solve real world reuse issues. To aid this examination, a methodology was developed which combines human interaction with computer-aided analysis. In cases where information models were developed in cooperation between domains, the semantics were already harmonised. However, results showed that in the Netherlands, most domain models are developed independently, and reuse of concepts from other models occurs only on an ad-hoc basis. This is for a large part due to the lack of discoverability and accessibility of domain models.

Semantic harmonisation improves usability of data, but it is not enough to fully enable reuse of geospatial data outside the geospatial sector. Geospatial data disseminated via SDI methods is difficult to find, access and use for non-geospatial experts, who are familiar with more general data publication methods.

Chapter 5 describes how general methods for data publication on the World Wide Web, Linked Data standards in particular, can be applied to geospatial data. Conversion of geospatial data formats such as Geography Markup Language (GML) to linked data standard Resource Description Format (RDF) is straightforward, but a choice between different ways to encode geometry in RDF is required, as is a URI strategy to ensure persistent and scalable URI identifiers for all data objects. Less straightforward is the conversion of geospatial UML-based data models to linked data models expressed in RDF Schema and Web Ontology Language (OWL), because of different

underlying paradigms. One aspect, the reuse of existing vocabularies, is addressed in detail.

Linked data, while broad in its applicability, is somewhat of a niche set of standards. When applied to geospatial data, it enables reuse of this data by linked data practitioners; however, it would still keep other potential users away, who experience linked data as an impediment to ease of use. To further improve reuse of geospatial data outside the geospatial sector, it is necessary to apply general web architecture principles and standards without mandating a specific metamodel such as linked data.

Chapter 6 describes a set of best practices for publishing geospatial data on the web, discovered in practice, and based on general web principles and standards. When implemented, these best practices make it easier to discover, interpret and use geospatial data for data users in general, e.g.web developers—not just for geospatial experts. In addition, some areas are identified where a best practice has not yet emerged.

Chapter 7 gives an overview of relevant developments since the research was carried out. 3D standardisation is ongoing in an international context. Semantic harmonisation in the Netherlands is progressing slowly but steadily. Geospatial linked data availability has grown significantly in the last few years, although there are still a few issues to solve. Several datasets implementing the Spatial data on the web best practices are available. These best practices have also triggered the further evolution of geospatial standards towards alignment with general web standards and principles.

Chapter 8 concludes the thesis and identifies the shift to (lighter) general web standards and principles as an important development and area for future work.

Samenvatting

Geodata is een steeds belangrijker wordende informatiebron bij het nemen van beslissingen, van eenvoudige keuzes zoals waar je je auto het beste kan parkeren, tot nationale en internationale beleidskeuzes over onderwerpen zoals infrastructuur en milieu. Vanwege het locatieaspect is geodata vaak de 'breinaald' die je door verschillende datasets heen kan steken om ze te integreren. Veel geodata wordt bijvoorbeeld geproduceerd in het kader van overheidsprocessen en wordt tegenwoordig gepubliceerd als open data, gewoonlijk via een "Geo-Informatie Infrastructuur" oftewel een "Spatial data infrastructure" (SDI).

Geodata heeft een hoge potentie voor hergebruik in andere domeinen dan het domein en het beoogde gebruik waarvoor het oorspronkelijk gemaakt is. Mijn onderzoeksvraag was: "Hoe kan geodata, die van verschillende, heterogene bronnen afkomstig is, worden hergebruikt via het web over domeinen heen?" Verschillende aspecten van datapublicatie moeten nog worden geadresseerd voordat open data daadwerkelijk goede kans maakt om te worden hergebruikt. Deze aspecten worden wel de "FAIR principles" genoemd: findability (vindbaarheid), accessibility (toegankelijkheid), interoperability (interoperabiliteit), en reusability (herbruikbaarheid).

Binnen het geo-domein is het algemeen bekend dat interoperabiliteit tussen systemen noodzakelijk is om hergebruik van data mogelijk te maken, en dat standaarden het mogelijk maken om deze interoperabiliteit te realiseren. Op basis hiervan heb ik een aantal verschillende probleemgebieden onderzocht, waar er potentieel was om geodata te hergebruiken, maar op de een of andere manier werd verhinderd. Deze problemen worden geïntroduceerd in hoofdstuk 1. Het onderzoek was exploratief van aard; de methodologie was een combinatie van bureaustudie, analyse, literatuurstudie en experimenten.

Hoofdstuk 2 en 3 richten zich op het ontbreken van een standaard voor drie-dimensionale (3D) geodata in Nederland, waardoor het hergebruik van 3D data verhinderd wordt. Als oplossing is een nationale standaard voor twee-dimensionale (2D) geografische, topografische data, het Informatiemodel Geografie (IMGeo), gecombineerd met een internationale 3D standaard, City-

GML. Beide standaarden beschrijven topografische objecten die fysieke objecten uit de werkelijkheid representeren en zijn voor een groot deel met elkaar te vergelijken.

Hoofdstuk 2 beschrijft hoe CityGML is geselecteerd als de 3D standaard die als basis voor de nationale 3D standaard kon dienen; hoe de inhoud van IMGeo met CityGML is afgestemd; hoe IMGeo als een formele uitbreiding op CityGML is gedefinieerd op het niveau van klassen, eigenschappen en codelijsten; en hoe andere interoperabiliteitsaspecten zoals geometrie, topologische structuur en referentiesysteem zijn geadresseerd.

Hoofdstuk 3 beschrijft in meer detail hoe IMGeo als een formele uitbreiding van CityGML is gedefinieerd. Dit is een technisch modelleervraagstuk, dat te maken heeft met het gebruik van Unified Modeling Language (UML) en het specifieke extensiemechanisme dat door de CityGML standaard wordt voorgeschreven. Op basis van het specifieke geval, IMGeo, is een modelgedreven raamwerk voor het maken van CityGML uitbreidingen beschreven.

In het geval van IMGeo en CityGML was de semantische harmonisatie, dat wil zeggen het afstemmen van de inhoud van beide standaarden, relatief eenvoudig. Beide standaarden beschrijven dezelfde soorten dingen; daardoor zijn de meeste klassen uit IMGeo op te vatten als hetzelfde of als een meer specifieke variant van een klasse in IMGeo. Dit is echter niet altijd zo eenvoudig. Het is onvermijdelijk dat domein-specifieke informatiemodellen, die onafhankelijk van elkaar ontwikkeld zijn, vergelijkbare inhoud op een verschillende manier modelleren, wat het hergebruik van data conform deze informatiemodellen hindert.

Dit probleem van semantische harmonisatie wordt geadresseerd in hoofdstuk 4. Verschillende Nederlandse standaarden werden onderzocht om die onderdelen te vinden waar overlap zat en waar inhoudelijke afstemming kon helpen bij het oplossen van praktische hergebruikproblemen. In het kader van dit onderzoek werd een methodologie ontwikkeld die menselijke interactie combineert met computer-ondersteunde analyse. Het onderzoek bevestigde dat semantische harmonisatie verbetert als informatiemodellen in samenwerking tussen domeinen ontwikkeld worden. Dit wordt echter in de praktijk doorgaans niet gedaan, en hergebruik van concepten uit andere standaarden gebeurt voornamelijk op een ad hoc manier. Voor een groot deel ligt dit aan de slechte vindbaarheid en toegankelijkheid van bestaande domeinmodellen.

Semantische harmonisatie verbetert de herbruikbaarheid van data, maar is niet de enige barrière die hergebruik van geodata in andere domeinen verhindert. Als geodata uitsluitend via een traditionele SDI wordt verspreid, kan deze data niet gemakkelijk worden gevonden en gebruikt door gebruikers van buiten het geo-domein, die wel bekend zijn met meer algemene data publicatie methodes.

Hoofdstuk 5 beschrijft hoe algemene methodes voor datapublicatie op het World Wide Web, in het bijzonder Linked Data standaarden, kunnen worden toegepast op geodata. Conversie van geo-dataformaten zoals Geography Markup Language (GML) naar de linked data standaard Resource Description Format (RDF) is niet gecompliceerd, maar vereist wel een keuze tussen verschillende manieren om geometrie op te nemen in RDF, en een URI strategie om te zorgen voor persistente, schaalbare URI identificaties voor alle dataobjecten. De conversie van in UML uitgedrukte geo-informatiemodellen naar linked data modellen die zijn uitgedrukt in RDF Schema en Web Ontology Language (OWL) is problematisch, omdat deze verschillende onderliggende paradigma's hebben. Eén aspect hiervan, het hergebruik van bestaande vocabulaires, is in detail uitgewerkt.

Linked data is, hoewel breed toepasbaar, toch in zekere zin te beschouwen als een niche standaard. Het publiceren van linked geodata maakt hergebruik mogelijk door linked data beoefenaars, maar een grote groep potentiële gebruikers ervaart linked data als een barrière voor eenvoudig hergebruik. Om het hergebruik van geodata verder te bevorderen, is het nodig om algemene architectuurprincipes en standaarden van het World Wide Web toe te passen, zonder een specifiek metamodel zoals dat van linked data te vereisen.

Hoofdstuk 6 beschrijft een verzameling aanbevelingen voor het publiceren van geodata op het web, die gedistilleerd zijn uit de praktijk, en gebaseerd zijn op de algemene architectuurprincipes en standaarden van het web. Als deze richtlijnen gevolgd worden zorgt dit ervoor dat geodata beter te vinden en gemakkelijker te interpreteren en te gebruiken wordt voor data gebruikers in het algemeen, bijvoorbeeld voor web developers - in plaats van alleen voor geo-experts. Ook worden een aantal deelaspecten van geodatapublicatie op het web geïdentificeerd waar nog geen goede richtlijn voor te vinden was.

Hoofdstuk 7 geeft een overzicht van relevante ontwikkelingen sinds ik mijn onderzoek heb uitgevoerd. 3D standaardisatie is nog gaande, niet alleen in Nederland maar ook internationaal. De semantische samenhang tussen standaarden is in Nederland langzaam maar zeker aan het verbeteren. Het aanbod aan geo-linked datasets is significant gegroeid in de laatste paar jaar, hoewel er nog een paar problemen op te lossen zijn. Meerdere datasets die de richtlijnen voor geodata op het web implementeren zijn inmiddels ook beschikbaar. Deze richtlijnen hebben een evolutie van de huidige geostandaarden in gang gezet naar het toepassen van algemene webstandaarden en -principes.

Hoofdstuk 8 sluit dit proefschrift af en wijst de overstap naar (lichtere) webstandaarden en -principes aan als een belangrijke ontwikkeling en onderwerp voor toekomstig werk.

Curriculum vitae

Linda van den Brink was born in Heerhugowaard, the Netherlands (1972). An affinity with languages and literature led to her choice to study "Algemene Letteren" at Utrecht University, specialising in "Algemene Literatuurwetenschap" (literary science). She obtained her doctorate in 1996. Also interested in computers and an early adopter of email and the (then just emerging) World Wide Web, she followed courses and obtained an internship at the department of "Computer en letteren" (humanities computing). Her doctorate thesis was concerned with the use of computer technology, in particular Standard Generalized Markup Language (the predecessor of XML), for literary text analysis tasks.

Linda went to work at Baan Company (1997-2003), where she wrote help texts and user manuals for enterprise resource planning software, and set up an SGML/XML-based authoring environment for user documentation. She attempted to combine this job with a PhD position at the Informatiekunde (previously Computer en Letteren) department at Utrecht University (2001 - 2004), but when she moved from Baan Company to her next job at Salience, this became problematic.

At Salience (2004-2007), a company specialised in implementing solutions based on XML, Linda worked as an expert consultant performing information analysis, implementing information models in XML Schema and transformations using XSLT, as well as giving training in XML technology.

After switching to a small outsourcing company (2007 - 2009), Linda went to work at the Kadaster as an information analist. She implemented an information modelling methodology based on UML, GML and XML Schema and created the 1.0 version of the information model Kadaster.

Next, Linda moved on to Geonovum (2009 - present day) where she works as an advisor on geo-standards and has worked on numerous national standards and information models for e.g. 2D/3D topography, soil and subsurface, and utility networks as well as, for example, a national URI strategy. In addition, she contributes to international standards developments and has been part of working groups for GML and CityGML, is chair of

the GeoSemantics working group at Open Geospatial Consortium, editor of the W3C/OGC Spatial Data on the Web Best Practice and co-chair of the W3C/OGC spatial data on the web interest group.

During her work at Geonovum Linda got the opportunity to publish several research articles in peer-reviewed journals. In 2014, she became a PhD candidate at TU Delft under the supervision of Prof.Dr. Jantien E. Stoter. Linda's dissertation is based on the articles she wrote and published. All the research described in these articles was performed in the context of her work at Geonovum.

List of publications

- Stoter, J. E., Morales, J. M., Lemmens, R. L. G., Meijers, B. M., Van Oosterom, P. J. M., Quak, C. W., Uitermark, H.T. & van den Brink, L.. A data model for multi-scale topographical data. In Headway in spatial data handling (pp. 233-254). Springer, Berlin, Heidelberg, 2008.

- Stoter, J., van den Brink, L., Vosselman, G., Goos, J., Zlatanova, S., Verbree, E., Klooster, R., van Berlo, L., Vestjens, G., Reuvers, M. & Thorn, S. A generic approach for 3D SDI in the Netherlands. In Proceedings of the Joint ISPRS Workshop on 3D City Modelling & Applications and the 6th 3D GeoInfo Conference Wuhan, China (pp. 26-28), 2011.

- van den Brink, L., Portele, C. & Vretanos, P.A. Geography Markup Language (GML) simple features profile (with Corrigendum). OpenGIS® Implementation Standard, 2012.

- van den Brink, L., Stoter, J. & Zlatanova, S. Establishing a national standard for 3D topographic data compliant to CityGML. International Journal of Geographical Information Science, 27(1):92-113, 2013.

- van den Brink, L., Stoter, J. & Zlatanova, S. UML-Based Approach to Developing a CityGML Application Domain Extension. Transactions in GIS, 17(6):920-942, 2013.

- Overbeek, H. & van den Brink, L. Towards a national URI Strategy for Linked Data of the Dutch public sector, 2013. Available online: `http://www.pilod.nl/wiki/Bestand:D1-2013-09-19%5FTowards%5Fa%5FNL%5FURI%5FStrategy.pdf`.

- Stoter, J., van den Brink, L., Beetz, J., Ledoux, H., Reuvers, M., Janssen, P., Penninga, F., Vosselman, G. & Oude Elberink, S. Three-dimensional modeling with national coverage: case of The Netherlands. Geo-spatial information science, 16(4), 267-276, 2013.

- Stoter, J., Beetz, J., Ledoux, H., Reuvers, M., Klooster, R., Janssen, P., Penninga, F., Zlatanova, S. & van den Brink, L.. Implementation of a national 3D standard: Case of The Netherlands. In Progress and New Trends in 3D Geoinformation Sciences, Lecture Notes in Geoinformation and Cartography, pages 277-298. Springer, 2013.

- van den Brink, L., Janssen, P., Quak, W. & Stoter, J. Linking spatial data: automated conversion of geo-information models and GML data to RDF. International Journal of Spatial Data Infrastructures Research, 9:59-85, 2014.

- van den Brink, L., Stoter, J. & Zlatanova, S. Modeling an application domain extension of CityGML in UML. OGC® Best Practice, 2014. Available online: `https://portal.opengeospatial.org/files/?artifact%5Fid=49000`.

- van den Brink, L., Janssen, P., Quak, W. & Stoter, J. Towards a high level of semantic harmonisation in the geospatial domain. Computers, Environment and Urban Systems, 62:233-242, 2017.

- Tandy, J., van den Brink, L. & Barnaghi, P. Spatial Data on the Web Best Practices, 2017. Available online: `https://www.w3.org/TR/2017/NOTE-sdw-bp-20170928/`.

- van den Brink, L., Barnaghi, P., Tandy, J., Atemezing, G., Atkinson, R., Cochrane, B., Fathy, Y., Garcia Castro, R., Haller, A., Harth, A., Janowicz, K., Kolozali, S., van Leeuwen, B., Lefrançois, M., Lieberman, J., Perego, A., Le-Phuoc, D., Roberts, B., Taylor, K. & Troncy, R. Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web. Semantic Web Journal, Pre-press:1-20, 2018.