Computers, Environment and Urban Systems 62 (2017) 233–242

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Towards a high level of semantic harmonisation in the geospatial domain

Linda van den Brink [a,*], Paul Janssen [a], Wilko Quak [a,b], Jantien Stoter [a,b,c]

[a] Geonovum, Netherlands
[b] TU Delft, Netherlands
[c] Kadaster, Netherlands

ARTICLE INFO

Article history:
Received 2 February 2016
Received in revised form 3 October 2016
Accepted 4 December 2016
Available online xxxx

Keywords:
Semantic harmonisation
Spatial data infrastructure
Ontology matching
Vocabulary
Community driven

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.

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) (Portele, 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

* Corresponding author.
E-mail addresses: l.vandenbrink@geonovum.nl (L. van den Brink), p.janssen@geonovum.nl (P. Janssen), w.quak@geonovum.nl, c.w.quak@tudelft.nl (W. Quak), j.stoter@geonovum.nl, j.e.stoter@tudelft.nl, Jantien.stoter@kadaster.nl (J. Stoter).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.12.002
0198-9715/

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 2 describes the background of the SDI approach in the Netherlands and the resulting harmonisation problems between domain models. Section 3 presents related work on harmonisation and ontology matching. Section 4 presents the overall methodology and tools that we developed to obtain a higher level of harmonisation between domain models. Section 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 6 presents part II of the research in which a further step was taken to resolve semantic discrepancies between information models where possible. Section 7 closes with conclusions and future work.

2. Background: model driven approach of the Dutch SDI

As explained in van den Brink, Stoter, and Zlatanova (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) (Group et al., 2009), UML is not widely different from these languages and provides sufficient means to record the meaning of objects (Kiko & Atkinson, 2008).

In the Netherlands' SDI, a Model Driven Approach (MDA) such as described in Gaševic, Djuric, and Devedžic (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, van Bennekom-Minnema, Van Oosterom, & Lemmen, 2008; OMG, 2003). As the UML models are conforming to an agreed meta model, i.e. the ISO 19109 (2015) 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) (Portele, 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 3610 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 (2015), ISO 19107 (2003), ISO 19109 (2015), ISO 19110 (2005), ISO 19131 (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 D2.5) 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 Fig. 1).

The abbreviations in the pyramid of Fig. 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 = T*opography*
- 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

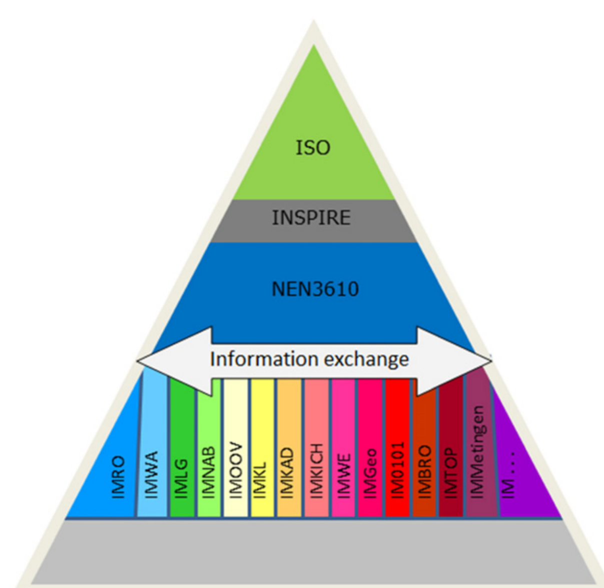

Fig. 1. The pyramid of domain information models.

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.

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 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, Shvaiko, Pane, Besana, and Marchese (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, Alani, & Smart, 2008). Their input can be collected using online surveys and serious games (Dewaraja, 2010) or using crowd sourcing (McCann, Shen, & Doan, 2008).

Nogueras-Iso, Zarazaga-Soria, Lacasta, Béjar, and Muro-Medrano (2004) studied the problem of harmonisation of metadata standards for the geographic domain and described transformations between ISO 19115-1 (2014), Dublin Core (DCMI, 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 19146, 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, Sunna, and Chaudhry (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, Reitz, & Gomarasca, 2008). They state that user interaction is required to overcome heterogeneity on the semantic 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. 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 was explained in Section 2. This section presents the methodology that we developed and applied to realise better harmonisation across the pyramid of information models.

Table 1
Methodology.

What	How	Main result
Part I: Identifying semantic overlap and resolving these (Section 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 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

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 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.

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 (Section 5.1), results (Section 5.2) and conclusions (Section 5.3).

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

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.

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.

5.2.1.1. 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).

5.2.1.2. 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.

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.

5.2.2.1. 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.

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 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.

Table 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

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.

5.2.2.2. 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.

5.2.3. Comparison IMGeo and IMLB

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

5.2.3.1. 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.

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.

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.

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.

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 2, 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.

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).

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 practices 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.

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 (Section 6.1) and finishes with results and conclusions (Section 6.2).

6.1. Methodology

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

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 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 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

Table 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	

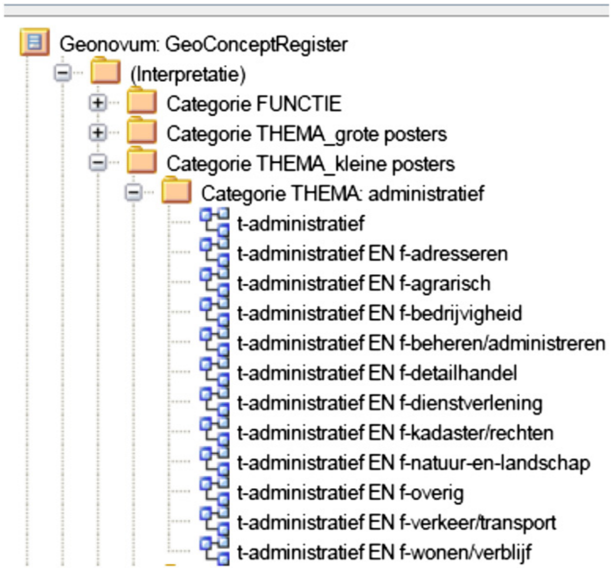

Fig. 2. Snapshot of the concept library prototype showing a list of diagrams with the results of combining the thematic keyword 'administrative' (prefix 't-' stands for 'thematic') with all functional keywords: 'addresses', 'agricultural', economic activity, administration, retailing, services, cadastre/rights, nature and landscape, other, traffic/transportation, living/residence. (prefix 'f-' stands for 'functional').

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 (Fig. 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 Fig. 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?
- 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.

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, Longley, Kellogg, Lanthaler, & Lindström, 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.

[1] See http://www.geostandaarden.nl/geoconceptregister/posters/index.htm.

[2] See http://www.geonovum.nl/melding/project/52/geosemantiek.

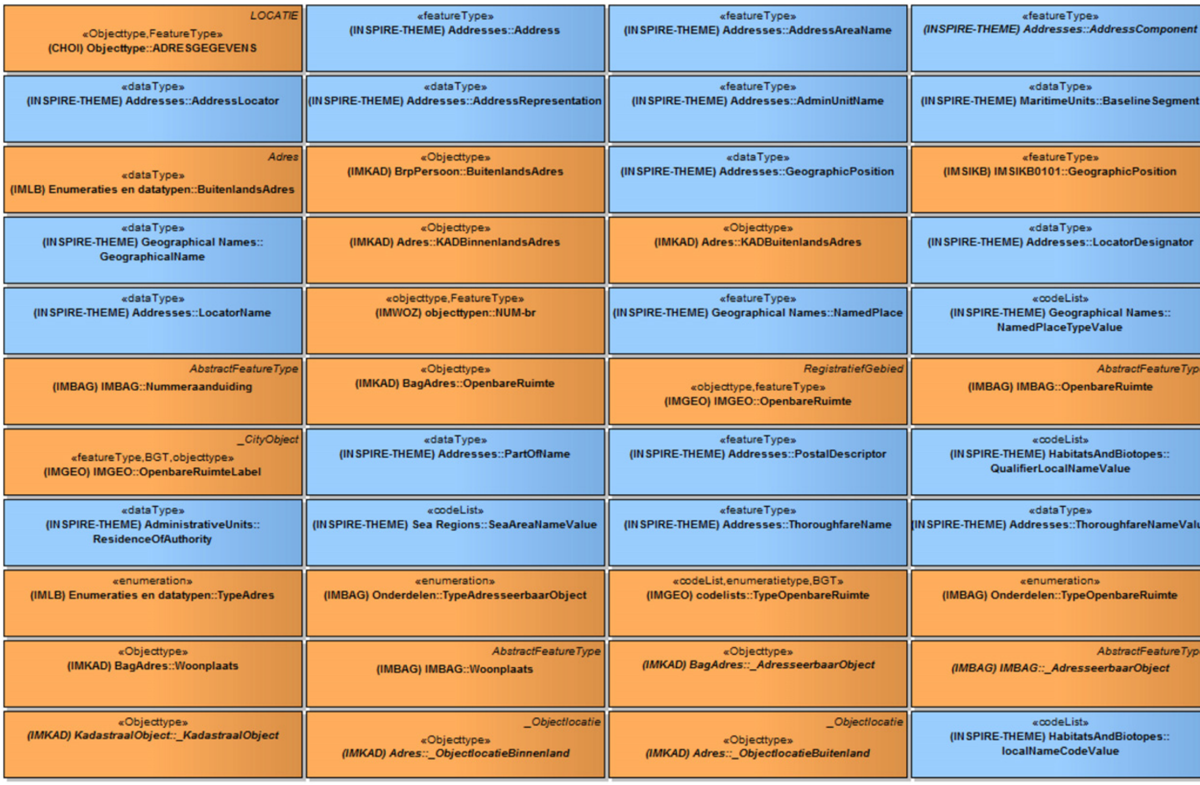

Fig. 3. Example poster: administrative & addresses. Concepts from Dutch models are in orange, INSPIRE concepts in blue. The figure shows that related concepts such as 'Woonplaats' (residence) from different models, IMKAD, IMBAG, INSPIRE, IMGEO, IMSIKB, IMLB are presented in one view to facilitate discussion on semantics. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 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

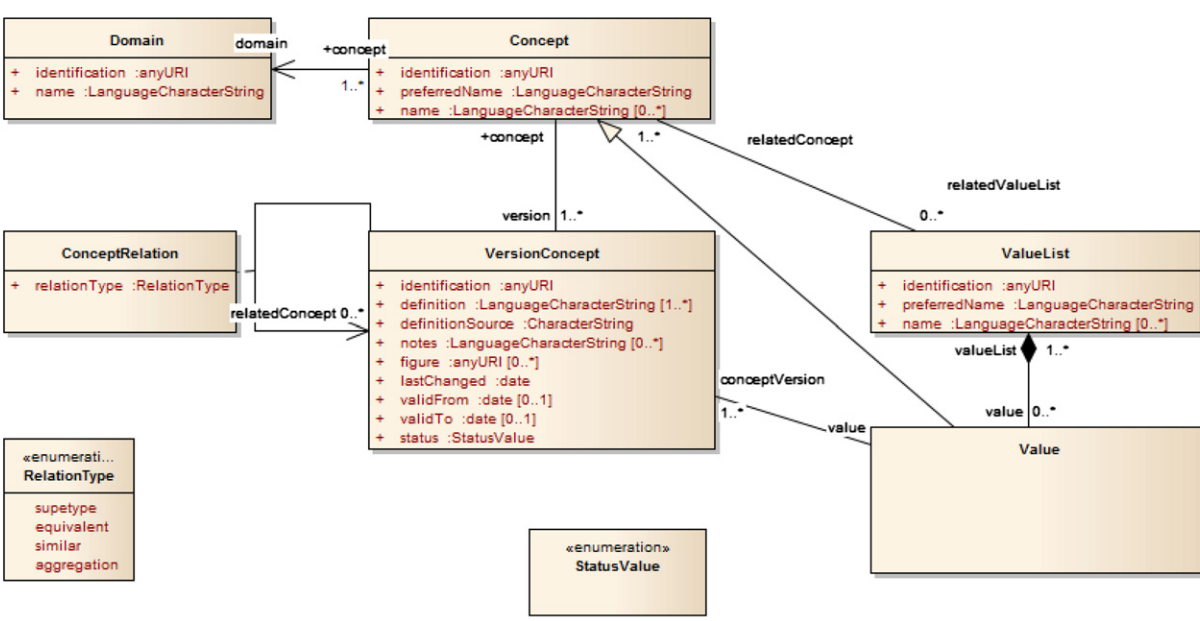

Fig. 4. Metamodel of the concept library.

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 & 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, Duchene, & Mackaness, 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 & 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 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.

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.

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 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 5.2).

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.

References

Burghardt, D., Duchene, C., & Mackaness, W. (Eds.). (2014). *Methodologies and applications of map generalisation. Series: Publications of the International Cartographic Association (ICA).*

Correndo, G., Alani, H., & Smart, P. R. (2008). *A community based approach for managing ontology alignments.*

Cruz, I. F., Sunna, W., & Chaudhry, A. (2004). Semi-automatic ontology alignment for geospatial data integration. *Geographic information science* (pp. 51–66). Springer.

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

Dublin Core Metadata Initiative (DCMI) (2012). DCMI metadata terms. (Online; accessed 2-February-2016) http://dublincore.org/documents/dcmi-terms/

Euzenat, J., & Shvaiko, P. (2013). *Ontology matching* (2nd ed.). Springer.

Gaševic, D., Djuric, D., & Devedžic, V. (2006). *Model driven architecture and ontology development.* Springer Science & Business Media.

Group, W. O. W., et al. (2009). *Owl 2 web ontology language document overview.*

Hespanha, J., van Bennekom-Minnema, J., Van Oosterom, P., & Lemmen, C. (2008). 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. *Fig working week'* (pp. 19).

INSPIRE (2007). 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. (Online; accessed 2-February-2016) http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002

INSPIRE (2014). D2.5: Generic conceptual model. (Version 3.4) http://inspire.ec.europa.eu/documents/Data_Specifications/D2.5_v3.4.pdf

ISO 19103 (2015). *Geographic information – Conceptual schema language. ISO standard, ISO.*

ISO 19107 (2003). *Geographic information – Spatial schema. ISO standard, ISO.*

ISO 19109 (2015). *Geographic information – Rules for application schema. ISO standard, ISO.*

ISO 19110 (2005). *Geographic information – Methodology for feature cataloguing. ISO standard, ISO.*

ISO 19115-1 (2014). *Geographic information – Metadata – Part 1: Fundamentals. ISO standard, ISO.*

ISO 19131 (2007). *Geographic information – Data product specifications. ISO standard, ISO.*

ISO 19146 (2010). *Geographic information – Cross-domain vocabularies. ISO standard, ISO.*

Kiko, K., & Atkinson, C. (2008). *A detailed comparison of UML and OWL. Reihe Informatik TR-2008-004.*

McCann, R., Shen, W., & Doan, A. (2008). Matching schemas in online communities: A web 2.0 approach. *'Data engineering, 2008. ICDE 2008. IEEE 24th International Conference on', IEEE* (pp. 110–119).

Miles, A., & Bechhofer, S. (2009). SKOS simple knowledge organization system reference. *W3C recommendation. 18.* (W3C).

Miller, G. (1995). *WordNet: a lexical database for English. Commun. ACM 38(11), 39–41.NEN 3610:2011 nl Basismodel geo-informatie – Termen, definities, relaties en algemene regels voor de uitwisseling van informatie over aan de aarde gerelateerde ruimtelijke objecten. NEN norm, NEN.*

Nogueras-Iso, J., Zarazaga-Soria, F. J., Lacasta, J., Béjar, R., & Muro-Medrano, P. R. (2004). Metadata standard interoperability: Application in the geographic information domain. *Computers, Environment and Urban Systems*, *28*(6), 611–634.

OMG, M. (2003). Guide version 1.0. 1. *Object management group. 62.*

Perry, M., & Herring, J. (2010). *OGC GeoSPARQL - A geographic query language for RDF data.* (At: http://www.opengeospatial.org/standards/geosparql, online; accessed 2 February 2016).

Portele, C. (2007). Opengis geography markup language (gml) implementation specification v3.2.1. *OGC document ref. no.: OGC.*

Reitz, T., & Kuijper, A. (2009). Applying instance visualisation and conceptual schema mapping for geodata harmonisation. *Advances in GIScience* (pp. 173–194). Springer.

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., & Lindström, N. (2014). *Json-ld 1.0-a json-based serialization for linked data. w3c recommendation.*

Vaccari, L., Shvaiko, P., Pane, J., Besana, P., & Marchese, M. (2012). An evaluation of ontology matching in geo-service applications. *GeoInformatica*, *16*(1), 31–66.

van den Brink, L., Stoter, J. E., & Zlatanova, S. (2013). Uml-based approach to developing a citygml application domain extension. *Trans. GIS*, *17*(6), 920–942.

Villa, P., Reitz, T., & Gomarasca, M. A. (2008). 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.