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

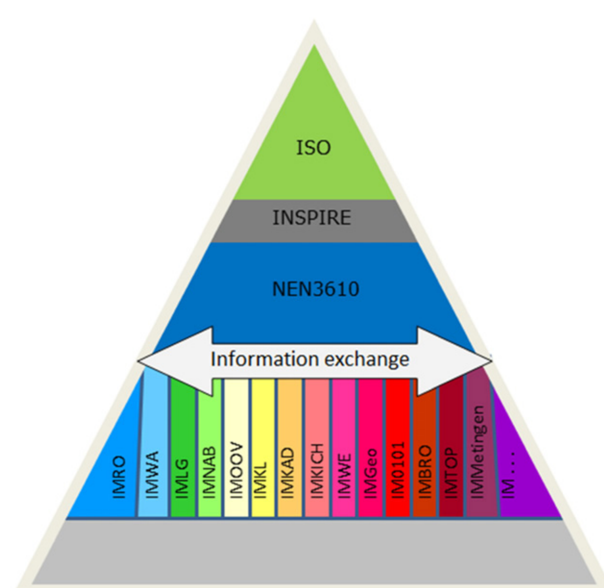

Fig. 1. The pyramid of domain information models.