Towards the linking of geospatial government data

A study on the semantic harmonisation between data from Dutch geo-registries

Gabriella Wiersma
4284423
m.g.wiersma@student.tudelft.nl

1st supervisor: Prof. dr. Jantien Stoter
2nd supervisor: dr. Linda van den Brink

January 6, 2020

1 Introduction

Interoperability solutions have long been a concern for the geospatial information community. According to the Open Geospatial Consortium (OGC), interoperability consists of "software components operating reciprocally (working with each other) to overcome tedious batch conversion tasks, import/export obstacles, and distributed resource access barriers imposed by heterogeneous processing environments and heterogeneous data." (2019).

There are mainly three types of interoperability issues: syntactic, structural or schematic and semantic. Syntactic interoperability issues are related to the exchange of data using agreed-upon data formats – it deals with GIS file format translations and the definition and use of geometric datatypes, for example. Structural issues, on the other hand, concern the application of data models and schemes. In the geospatial domain this is related to the use of different datums, map projections and coordinate systems. Finally, semantic heterogeneity deals with ambiguously defined concepts – it considers the content and meaning of data instances.

While it is possible to create converters and integrators that can solve differences in data coding, topology and data formats across databases, this approach cannot be used to resolve semantic issues. Semantic heterogeneity is especially challenging because it relates to the meaning of words across different systems. And since organizations understand the meaning of concepts in different ways – depending on the domain or application for which the data is primarily collected -, conflicts are most likely to occur.

Several approaches exist for semantic data integration, among which ontology-based methods. Ontologies are used to represent a shared understanding of a certain domain and consist of a set of concepts (mainly entities and attributes), definitions and relationships. Although ontologies can take up different forms, they will always include a vocabulary with definitions of terms. Depending on the degree of formality used to create the vocabularies, ontologies can be expressed in natural language, formally defined languages or with theorems and proofs (Uschold and Gruninger, 1996). In the context of information sciences, formal ontologies are mostly used as they give access to robust computational tools (such as inferencing and reasoning engines).

The employment of formal ontology structures for online data integration has been the subject of many studies and can be understood through the principles of the Web 3.0 – or Semantic Web. According to Berners-Lee et al. (2001), the Semantic Web extends the current Web by giving data well-defined meaning in machine-understandable language. One of the main goals of this vision is to enable systems to interpret the content of web resources. With increasing volumes of data being made accessible through the web, mechanisms and applications for search, retrieval and integration of this newly available information must be developed.

Therefore, the World Wide Web Consortium (W3C) has created a set of tools to implement the ideas of the Semantic Web. Some of the main technologies behind this idea are the Resource Description Framework (RDF) [1], the Web Ontology Language (OWL)[2] and the SPARQL query language [3]. RDF provides the graph-based data model to de-

[1] https://www.w3.org/RDF
[2] https://www.w3.org/OWL
[3] https://www.w3.org/TR/sparql11-overview

scribe all things on the Web. It allows structuring information about resources by using statements containing a subject, predicate and object – also referred to as triples (see Figure 1. Although the knowledge contained in RDF triples can be used to define simple relationships between concepts and resources, it is not rich enough to tackle semantic heterogeneity. To this end, formal and machine-understandable ontologies can be leveraged. These can be expressed using OWL, a formal and standardized ontology language based on description logics.

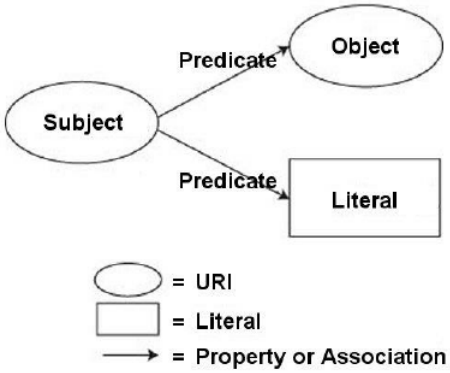


Figure 1: Representation of triple structure. Adapted from Žáček et al. (2015)

In The Netherlands as well as abroad there is a growing need for applications involving cross-domain data – especially considering the emergence of several web-based services in the context of open data developments. Therefore, the ideas behind the Semantic Web are being used to promote and stimulate the use of data made available through Spatial Data Infrastructures (SDIs) – this is referred to as 'linked data'. Linked data denotes the design principles used for the sharing of machine-readable and interlinked web data, and is summarized in four principles (Berners-Lee, 2019):

1. Use URIs to name (identify) things.
2. Use HTTP URIs so that these things can be looked up (interpreted, "dereferenced").
3. Provide useful information about what a name identifies when it's looked up, using open standards such as RDF, SPARQL, etc.
4. Refer to other things using their HTTP URI-based names when publishing data on the Web.

A recent survey conducted by the European Spatial Data Research (2018) showed linked data to be one of the most important research items and key factors moving SDIs toward the next generation. Thus, governments have been investing in linked data initiatives. A recent example is the Data Integration Partnership for Australia (DIPA), which created the 'Location Spine', a model for describing the links between objects from different datasets (Car et al., 2019). In Europe, several other initiatives can be found as well. The Italian Institute for Statistics and the Agency for Digital Italy have published public administration information as linked data. The United Kingdom Ordnance Survey has published three open data products as linked data – including the administrative

geography for Great Britain. And in The Netherlands, ongoing efforts from the Dutch Cadaster have led to the publication of several (geo)registries as linked data.

While publishing data online using formal ontology languages might be an important first step towards semantic interoperability, it does not lead to linked data. For data to be truly linked it does not suffice to publish it with linked data standards. It must also be possible to establish connections between instances from different datasets when they relate to the same real-world objects. Although progress has been made in this area (see the overview of currently linked datasets from the Linked Open Data Cloud initiative, in Figure 2), a lot can still be achieved. Therefore, the next section will give an overview of ontology-based data integration frameworks.

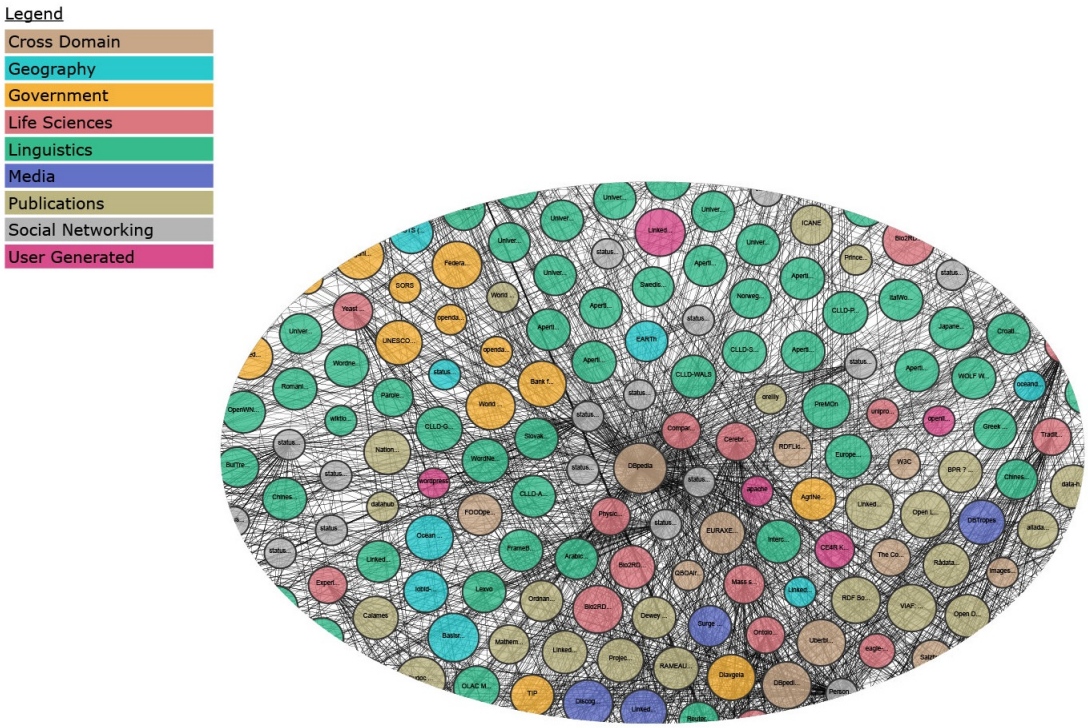


Figure 2: Fragment of Linked Open Data Cloud showing connections between datasets and ontologies. Adapted from The Linked Open Data Cloud (2019)

2 Background and related work

In the 1990s, the concept of ontologies was introduced into the field of geographic information science (Winter, 2001, as cited in Sun et al., 2019a). In order to understand the development and relationship between ontologies, Uschold (2000) classified them into global and local. Global ontologies represent a common and shared agreement between different groups or units and can be found at various levels of hierarchy – on the level of an organization as well as a domain. Local ontologies are then characterized by the fact that they have to be aligned to a global ontology, as they are not a primary source of reference (Janowicz, 2012). Although the distinction between global and local ontologies can be fluid, this classification will be used to understand research efforts in the devel-

opment of geo-ontologies.

Although global ontologies can be found within organisations and at smaller scales, the idea of creating one top-level or universally agreed upon ontology for the geo-information community has attracted some interest from researchers. Kuhn (2003) relates to this idea by introducing the concept of semantic reference systems. Similar to how spatial reference systems enable integration of spatial data across systems, semantic systems should allow integrating data cross heterogeneous semantics. A semantic reference system has a semantic reference frame and a semantic datum. The semantic datum allows projecting data models to simpler representations and translating data from different models. The reference frame consists of a conceptualization of a certain universe of discourse and can be formalized by using ontologies. To realize a semantic reference frame (and ultimately the system), the first step would be to extract semantic primitives and formalize their meaning. An experimental implementation of such a system was realized in Baglatzi and Kuhn (2013) by creating an ontology for the land cover domain based on the Conceptual Space Markup Language (CSML).

More recently, Sun et al. (2019a) attempted to create a comprehensive ontological framework named GeoDataOnt, which they believe could provide a standardized representation for semantic geo-information. Their research involved finding the main semantic issues related to geospatial data sharing and integration. The results were then used to identify relevant top-level terms based on essential, morphologic and provenance characteristics of geospatial data. Each type of characteristic was formalized in an ontology, represented in OWL. However, as the quality of framework has not been evaluated, there is no insight into whether the entities defined in the ontologies (classes, properties, relations and instances) are able to accommodate views from different local geo-ontologies. Furthermore, the ontologies were created manually. This method is not the most suited for such complex tasks, as it is error-prone and turns updating into a labour-intensive task.

While data integration might motivate the use of semantic web technologies, it is not an end on itself. Often, integration represents a pre-requisite to accomplish more advanced tasks. Therefore, another approach is to focus on the development and interlinking of local geo-ontologies, which are used to facilitate more specific tasks. Zhang et al. (2010), for example, explored the use of semantic web technologies in the context of disaster and emergency management. Their research focused on finding a solution for searching feature level geospatial data based on their content – instead of traditional metadata keyword searches -, by means of OWL ontologies. A more recent example is the research of Chen et al. (2018), which introduced a mapping mechanism and a semantic translation engine to generate domain ontologies that can be used for the computation of urban density indicators. They use OWL-DL to express their ontology. Wang et al. (2018) also developed an ontology-driven integration system that allows exploring information related to geology and palaeontology, with the goal of improving the compatibility between local and global geologic standards. Futia et al. (2017) used SPARQL queries to investigate inconsistencies in procurement data, and found problems related to incoherent payments under ongoing contract and multiple registered business names. Their research was motivated by developments in the publication of Open Government Data (OGD). More recently, the research of Homburg and Boochs (2019) emphasized the importance of data quality in geospatial linked data. The authors indicate that data

quality could be used by reasoners for decision making processes – to help assess the reliability of the information contained in the employed data sources. To this end, data quality requirement profiles were created, which define metrics and value ranges indicating the reliability of the data regarding a certain use case. The profiles were then converted into SWRL[4] reasoning rules. These rules are applied to the data stored in a triplestore, and through GeoSPARQL [5] queries it is possible to find out if the available data is suitable for the intended purpose.

The integration between local ontologies from government authorities has also attracted considerable attention. Years ago, Alani et al. (2007) were already exploring the benefits of using semantic web technology to enable better re-use of public sector information in the UK. The researchers collected data from several public sector organisations and designed OWL-DL ontologies for each dataset. Mappings were created between both concepts and instances of the datasets using CROSI, an alignment tool offering a wide range of mapping algorithms. Mapping on instance level was done by creating scripts searching for duplicates of specific instances, which were then connected through owl:sameAs links. Finally, all local ontologies were manually mapped to the best matched terms in the government's reference taxonomy (the Integrated Public Sector Vocabulary, or IPSV). This integration of data sources provided the researches with insights into the quality of datasets. For example, joining business information from different sources on their address coordinates revealed mismatching information.

Yu et al. (2017) explored the use of ontologies to avoid data duplication between Australian governmental authorities. The researchers build ontologies for Points of Interest datasets from different organizations. Then, the data integration tool Karma was used to convert the source data to RDF format. Finally, automated reasoning (through SWRL rules based on geometry, topology and policy rules) was proposed as a solution for finding the best location in the context of emergency response applications. The methodology was limited to handling point geometries.

In The Netherlands, Brink (2018) has explored cross-domain semantic harmonisation between different domain models within the Dutch SDI - starting from the Information model Geography. The method used in the research was based on manual matching, as the focus was to promote better data re-use by involving stakeholders personally. Firstly, the semantic overlap between models was found and published in a register using a domain independent classification. The visualisation of the register then exposed the semantic conflicts, which were discussed with the model-owners. Further efforts by the Dutch National Mapping Agency, Kadaster, have led to the publication and linking between three base registries: BAG, BRT and BRK. The research of Ronzhin et al. (2019) describes the process of building a knowledge graph for these and other official datasets. Most of the data was aligned by using spatial relations, by means of topological analyses based on GeoSPARQL queries.

Regardless of the approach chosen for integrating geo-ontologies, establishing correspondences between entities from different data sources remains a key research issue. The problem involves creating mechanisms for finding the correspondences (ontology matching), as well as deciding on how to express the results (the alignment) in a machine-processable way. To this end, many tools have been created and made available online. In order to evaluate the performance of such tools, the Ontology Alignment Eval-

[4]https://www.w3.org/Submission/SWRL

[5]https://www.opengeospatial.org/standards/geosparql

uation Initiative (OAEI) [6] organizes annual contests. The initiative publishes benchmark datasets from different domains, composed of two ontologies and a reference alignment developed manually by experts. The performance of the matching tools is measured by comparing the results of their alignments to the reference alignment. However, many of the alignment algorithms used by general matching tools do not account for the spatial characteristics of geo-information. Moreover, the OAEI does not provide benchmark datasets that are representative of the geospatial domain.

3 Research objectives

3.1 Objectives

Much work has been done on creating a Dutch system of registries based on standardized information models. Nonetheless, more research is needed to determine how data from different models can be better combined and to what extend integration can aid more specific applications. Currently available datasets are not completely linked according to the fourth principle of linked data. Therefore, this research aims to explore the ways in which further interlinking of data - trough the use of semantic web technologies - leads to integration and coherence in the system of registries. //

The main question of the research then becomes: "To what extent can ontology-based alignment using semantic web technologies contribute to the integration and use of data from geo-spatial registries?". //

To answer this question, it is important to consider the different ways in which alignments between ontologies can be made. Moreover, integration must be seen as a means to an end. Therefore, the extent to which integration is accomplished will be measured by analyzing how the alignments can help infer new and useful knowledge. Thus, the main question can be broken down into the following sub-questions:

- What type of alignment and mapping techniques are best suited in the case of integrating data from the geo-registries? How does the semantic overlap and differences between the data affect the alignments?
- To what extent can class-based manual alignment and inference rules be used for generating new knowledge? Can this knowledge be used to detect logical inconsistencies?
- To what extent are current (semi)automated alignment tools equipped for dealing with geospatial data and what are their limitations? Could (semi)automated instance-based alignment lead to more findings regarding data inconsistencies?

3.2 Scope of research

This research will focus on the alignment of data from geo-registries, based on the semantics of classes, properties and their values. There are many techniques available for the alignment and mapping of ontologies, and this thesis will only consider those that can be used in conjunction with semantic web technologies. The alignment itself will be performed manually. However, qualitative research will be carried out to address the

[6] http://oaei.ontologymatching.org/

usefulness of (semi) automated methods. If the manual alignments do not lead to satisfactory results, it could be possible to investigate the use of instance-based alignment tools - this has been expressed in the last sub-question.

4 Methodology

A general overview of the methodology is given below, in Figure 3. It consists of the basic steps needed to answer the research question. The first steps regard data preparation, as both the chosen geo-registry datasets must be made available as linked data. Afterwards, the actual alignment and evaluation of the integration can take place. This follows a heuristic method, were alignment rules might be readjusted and improved to test knowledge inference. Qualitative research to assess adequate alignment techniques - both manual and (semi) automated - will be a part of the process.

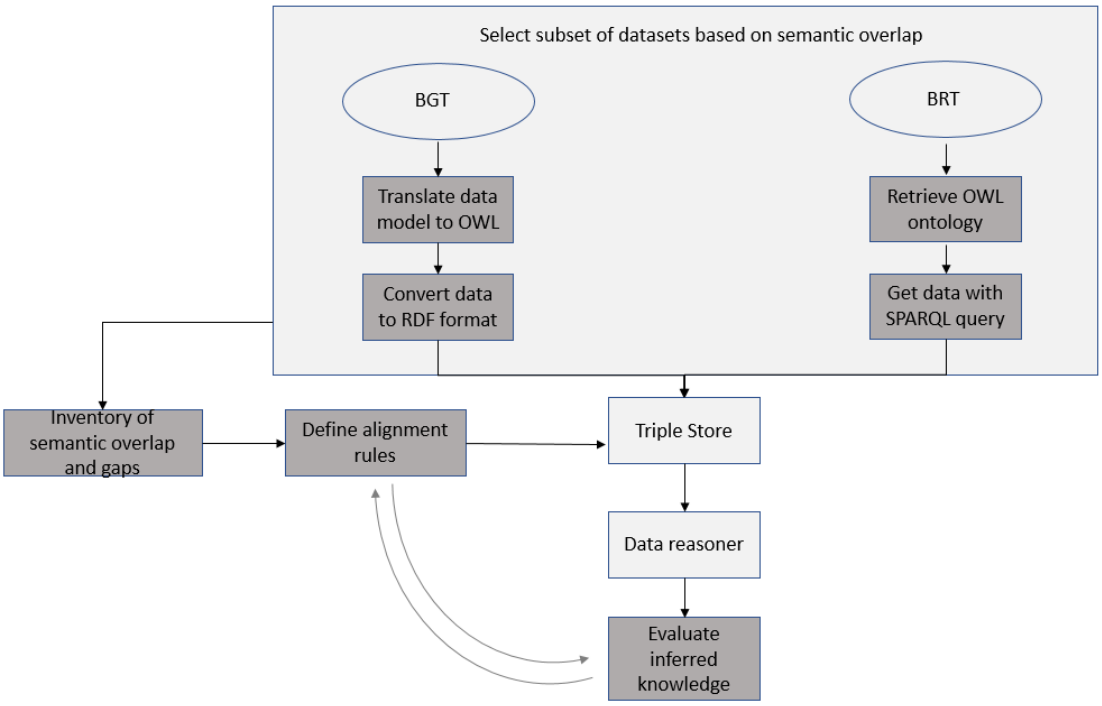


Figure 3: Overview of the main steps in the methodology

4.1 Dataset selection, conversion and storage

To answer the research questions a subset of data from minimally two geo-registries will be used - Basisregistratie Grootschalige Topografie (BGT) and the Basisregistratie Topografie (BRT - Top10NL) are the starting point, as shown in Figure 3. These datasets have been selected for several reasons. Firstly, there is a significantly high semantic overlap in their domains, meaning there is a high potential for integration (see for comparison between both datasets the fragments shown in Figure 4 and Figure 5. Secondly, BGT contains more detailed information and could be used to automatically extract the Top10NL in the future. Knowledge on the semantic links between their data objects

could provide valuable insights. Finally, there are currently no efforts into integrating both models by using semantic web technologies - as the BGT has not yet been published using linked data standards. Therefore, the first step will be to transform a subset of BGT data to linked data. As a vocabulary for the data model is already available, only scripts to translate the data to RDF format will have to be produced. The BRT data, on the other hand, can be downloaded as linked data through the PDOK portal [7]. After obtaining the linked data 'snapshot' of both models, it can be loaded into a triple store - to be used the next steps.

Figure 4: Example of BGT data from Rotterdam

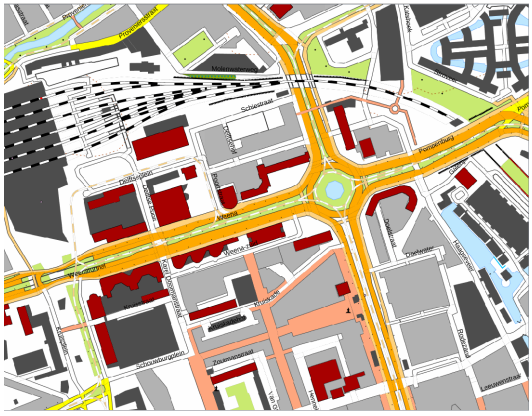


Figure 5: Example of Top10NL data from Rotterdam

4.2 Finding semantic overlap and defining alignment rules

In this step the actual analysis of the data models takes place, as it is necessary to identify the overlap between entities before alignment is possible. Initially, only overlap on the data model level (concepts and properties) will be taken into consideration. The work of Brink (2018) on the concept register for Dutch information models will be used as a starting point. Manual alignments will then be created for both data models. The alignments can be expressed by different means - with additional OWL assertions, SPARQL construct queries, SWRL rules or Jena rules, for example. These rules can then be added

[7] https://www.pdok.nl/

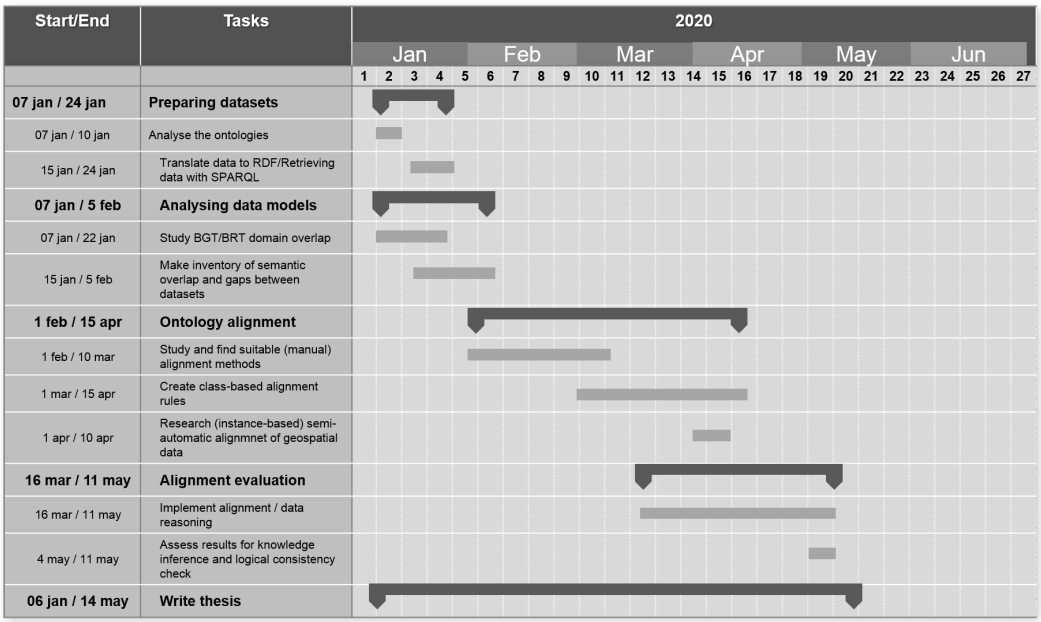
to the triple store containing the linked data and ontologies. It must be ensured that the alignment rules contain enough connecting statements that would allow inconsistencies between the data to be found, if these exist. Thus, an inventory of situations with possibly contradicting BGT/BRT data has to be made beforehand. Then, the reasoner can be employed to check for these inconsistencies.

4.3 Overview of automated alignment tool capabilities

The alignment performed in the previous step relies on manually defined rules for classes and instances. However, automated alignment algorithms based on instance-level data could generate more insights - leading to better data integration and identification of inconsistencies. Therefore, it might be useful to provide an overview of the current state of automated ontology matching tools/systems with regards to how they handle geospatial alignment. This step would consist mostly of qualitative research. The workflow provided by Sun et al. (2019b) to describe entity alignment in the geospatial domain can be used to guide the evaluation of the tools. Although the OAEI does not provide a comprehensive standard for geospatial data, the performance of the systems on these gold standards can be used to select the tools most likely to produce good alignment outcomes. One of the possible limitations of current frameworks is that matching techniques focus on establishing equivalence relations alone, and fail to account for other possibilities - such as 'is-a' and 'part-of' relations. These limitations should be kept in mind.

5 Time planning

The following schedule indicates the activities necessary to meet the research objectives, and gives an estimate of the time that is needed for each task.



6 Tools used

For visualization and study of the ontologies and converted RDF data the Protégé ontology editor will be used, as it supports many plugins and tools for this purpose. The RDF data and alignment rules can be stored by using Apache Jena[8], a java framework that allows building semantic web and link data applications. Jena provides support for several inference engines, and can be used for the evaluation of the alignment. The scripts necessary for conversion of BGT data will be written in Python, using the RDFLib module[9].

[8] https://jena.apache.org/
[9] https://rdflib.readthedocs.io/en/stable/

References

H. Alani, D. Dupplaw, J. Sheridan, K. O'Hara, J. Darlington, N. Shadbolt, and C. Tullo. Unlocking the potential of public sector information with semantic web technology. 01 2007.

A. Baglatzi and W. Kuhn. On the formulation of conceptual spaces for land cover classification systems. In *Geographic Information Science at the Heart of Europe*, pages 173–188. Springer International Publishing, 2013. doi: 10.1007/978-3-319-00615-4_10.

T. Berners-Lee. Linked data - design issues. `https://www.w3.org/DesignIssues/LinkedData.html`, 2019. Accessed 10 Nov 2019.

T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *ScientificAmerican.com*, 05 2001.

L. v. d. Brink. *Geospatial Data on the Web*. PhD thesis, 2018.

N. J. Car, P. J. Box, and A. Sommer. The location index: A semantic web spatial data infrastructure. In P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, and K. Hammar, editors, *The Semantic Web*, pages 543–557, Cham, 2019. Springer International Publishing. ISBN 978-3-030-21348-0.

Y. Chen, S. Sabri, A. Rajabifard, and M. E. Agunbiade. An ontology-based spatial data harmonisation for urban analytics. *Computers, Environment and Urban Systems*, 72:177–190, nov 2018. doi: 10.1016/j.compenvurbsys.2018.06.009.

European Spatial Data Research. Annual report. `http://www.eurosdr.net/sites/default/files/images/inline/eurosdr_annual_report_2018.pdf`, 2018. Accessed on 7 Nov 2019.

G. Futia, A. Melandri, A. Vetrò, F. Morando, and J. C. D. Martin. Removing barriers to transparency: A case study on the use of semantic technologies to tackle procurement data inconsistency. In *The Semantic Web*, pages 623–637. Springer International Publishing, 2017. doi: 10.1007/978-3-319-58068-5_38.

T. Homburg and F. Boochs. Situation-dependent data quality analysis for geospatial data using semantic technologies. In *Business Information Systems Workshops*, pages 566–578. Springer International Publishing, 2019. doi: 10.1007/978-3-030-04849-5_49.

K. Janowicz. Observation-driven geo-ontology engineering. *Transactions in GIS*, 16(3): 351–374, may 2012. doi: 10.1111/j.1467-9671.2012.01342.x.

W. Kuhn. Semantic reference systems. *International Journal of Geographical Information Science*, 17(5):405–409, jun 2003. doi: 10.1080/1365881031000114116.

Open Geospatial Consortium. Glossary of terms - i. `https://www.opengeospatial.org/ogc/glossary/i`, 2019. Accessed 1 Dec, 2019.

Ronzhin, Folmer, Maria, Brattinga, Beek, Lemmens, and van't Veer. Kadaster knowledge graph: Beyond the fifth star of open data. *Information*, 10(10):310, oct 2019. doi: 10.3390/info10100310.

K. Sun, Y. Zhu, P. Pan, Z. Hou, D. Wang, W. Li, and J. Song. Geospatial data ontology: the semantic foundation of geospatial data integration and sharing. *Big Earth Data*, 3 (3):269–296, jul 2019a. doi: 10.1080/20964471.2019.1661662.

K. Sun, Y. Zhu, and J. Song. Progress and challenges on entity alignment of geographic knowledge bases. *ISPRS International Journal of Geo-Information*, 8(2):77, feb 2019b. doi: 10.3390/ijgi8020077.

The Linked Open Data Cloud. `https://lod-cloud.net/`, 2019. Accessed 10 Dec 2019.

M. Uschold. Creating, integrating and maintaining local and global ontologies. In *Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI-2000)*. Citeseer, 2000.

M. Uschold and M. Gruninger. Ontologies: Principles, methods and applications. *KNOWLEDGE ENGINEERING REVIEW*, 11:93–136, 1996.

C. Wang, X. Ma, and J. Chen. Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. *Computers & Geosciences*, 115:12–19, jun 2018. doi: 10.1016/j.cageo.2018.03.004.

F. Yu, D. A. McMeekin, L. Arnold, and G. West. Semantic web technologies automate geospatial data conflation: Conflating points of interest data for emergency response services. In *Lecture Notes in Geoinformation and Cartography*, pages 111–131. Springer International Publishing, dec 2017. doi: 10.1007/978-3-319-71470-7_6.

M. Žáček, R. Miarka, and O. Sýkora. Visualization of semantic data. In *Advances in Intelligent Systems and Computing*, pages 277–285. Springer International Publishing, 2015. doi: 10.1007/978-3-319-18476-0_28.

C. Zhang, T. Zhao, and W. Li. Automatic search of geospatial features for disaster and emergency management. *International Journal of Applied Earth Observation and Geoinformation*, 12(6):409–418, dec 2010. doi: 10.1016/j.jag.2010.05.004.