Geospatial data and knowledge on the Web

Knowledge-based geospatial data integration and visualisation with Semantic Web technologies

Huang, Weiming

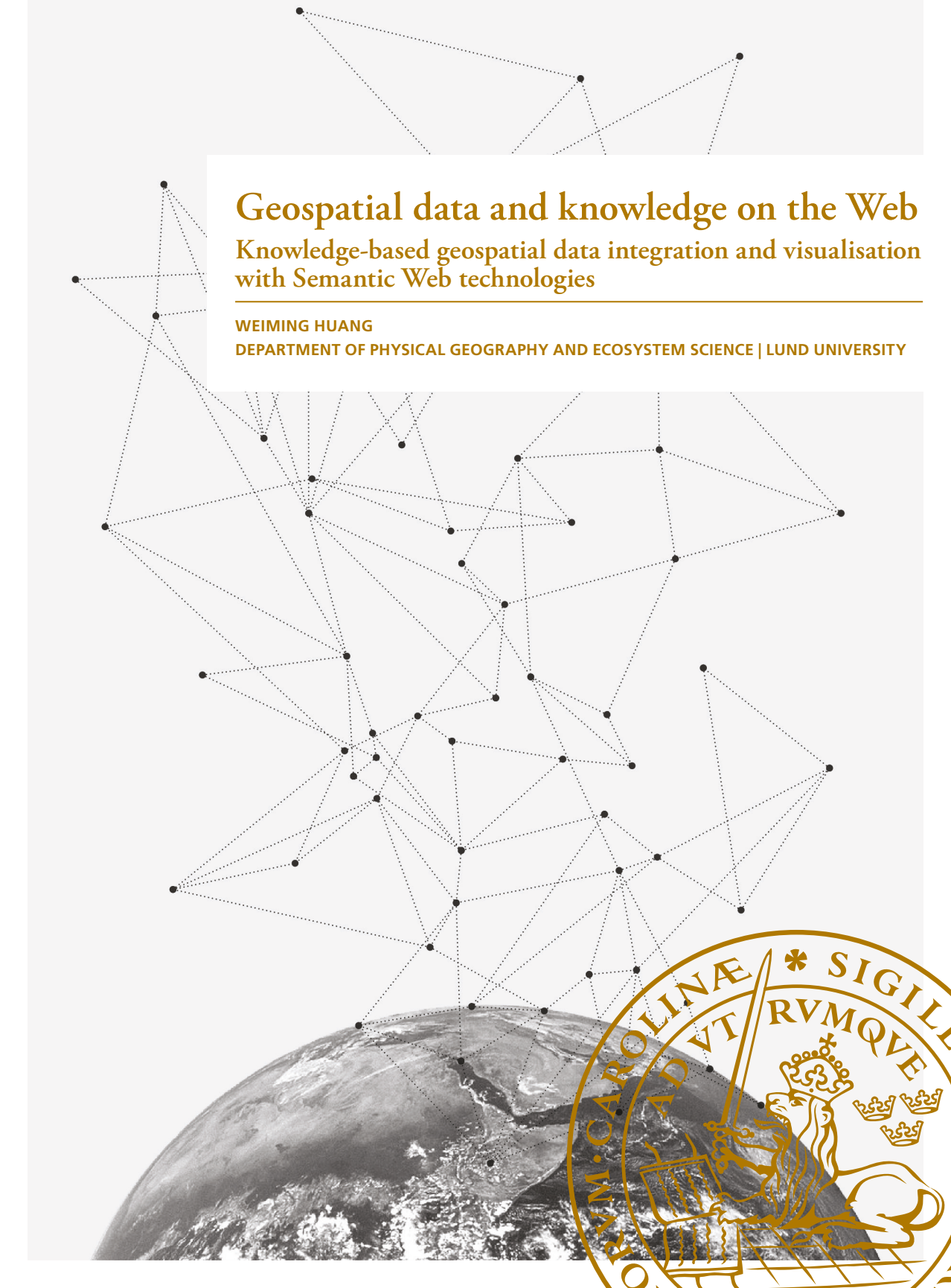
2020

Document Version:
Publisher's PDF, also known as Version of record

Geospatial data and knowledge on the Web

Knowledge-based geospatial data integration and visualisation with Semantic Web technologies

WEIMING HUANG
DEPARTMENT OF PHYSICAL GEOGRAPHY AND ECOSYSTEM SCIENCE | LUND UNIVERSITY



LUND
UNIVERSITY

Department of Physical Geography
and Ecosystem Science
Faculty of Science

ISBN 978-91-985015-1-3

Geospatial data and knowledge on the Web

Knowledge-based geospatial data integration and visualisation with Semantic Web technologies

Weiming Huang

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Science, Lund University, Sweden.
To be defended at Pangea auditorium, Geocentrum II, Sölvegatan 12, Lund.
Friday, January 10, 2020 at 10:00

Faculty opponent
Professor Lars Bernard
Technical University of Dresden

Organization	Document name
LUND UNIVERSITY	DOCTORAL DISSERTATION
	Date of issue January 10th, 2020
Author(s): Weiming Huang	Sponsoring organization

Title and subtitle: Geospatial data and knowledge on the Web: Knowledge-based geospatial data integration and visualisation with Semantic Web technologies

Abstract

Geospatial information is indispensable for various types of spatially informed analysis and decision-making, such as traffic analyses, and natural resource management. In addition, geospatial information is one of the most powerful information integrators to bridge diverse sources of information. Such natures of geospatial information entail the need of geospatial data integration and geospatial knowledge outreach.

In most of real-world applications, geospatial data from one single source can hardly suffice. Therefore, integrating multi-source geospatial data is a predominant need for a variety of applications. However, today's solutions for geospatial data integration in spatial data infrastructures (SDIs) are inadequate, and the data are often stored in the so-called "data silos", i.e. datasets are stored mostly isolated from each other.

Semantic Web technologies provide a promising way for geospatial data integration on the Web. In this thesis, Semantic Web technologies are utilised to integrate multi-source geospatial data or integrated geospatial data with data from other domains. Paper I leveraged Linked Data and ontologies to realise a relative positioning approach, which positions thematic data based on their relations with background data. The relatively positioned thematic data can be automatically synchronised in terms of their geometric representations in all scales to avoid substantial discrepancy. Paper II integrated distributed multi-scale building data and a heritage building dataset to accomplish a heritage building map with both fine geometries and thematic information of heritage building. Paper III identified that only using ontologies is inadequate for integrating geospatial data and data from other domains, where complex and subtle semantic relations often arise. Then, a knowledge-based framework coupling ontologies and semantic constraints is developed to tackle the complex semantic relations raised by multiple representations of geospatial data.

Besides data integration, there is another prominent need for utilising geospatial data for various applications, that is, the outreach of geospatial knowledge. Visualisation, as one of the most predominant ways of utilising geospatial data, pertains to a wide range of cartographic knowledge. Therefore, it is desirable to formalise the knowledge of geospatial data visualisation (geovisualisation) to facilitate its interpretation, transfer, and reuse.

In this context, Semantic Web technologies offer a framework to formalise and share geovisualisation knowledge, thanks to their knowledge representation capacity. To this end, Paper II and III formalised the knowledge of geovisualisation in knowledge bases with ontologies and semantic rules. Such knowledge bases are evaluated in two real-world applications, i.e. heritage building mapping, and urban bicycling suitability mapping. The knowledge bases for geovisualisation can be used as a visualisation enablement layer for geospatial Linked Data.

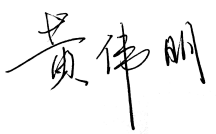
In addition, Paper IV performed a study for the technical environment of the support for geospatial Semantic Web (Linked Data). It assessed and benchmarked several well-known and mainstream Linked Data stores, mainly in terms of their spatial query capacities and standard compliance. The results demonstrated that the support for geospatial Linked Data and queries has becoming increasingly mature. Nevertheless, query correctness remained a challenge for cross-database interoperability.

In conclusion, this thesis provides insights into the potentials of Semantic Web technologies for geospatial data integration and knowledge outreach (sharing). The insights could benefit the development of the next generation of SDIs, in which Semantic Web and Linked Data will expectedly play a role. From the perspective of Semantic Web research, the thesis contributes to the modelling and representation of geospatial data and knowledge on the Semantic Web.

Key words Geospatial data integration, data visualisation, Semantic Web, Linked Data, ontology

Classification system and/or index terms (if any)

Supplementary bibliographical information		Language English
ISSN and key title		ISBN 978-91-985015-1-3
Recipient's notes	Number of pages 160	Price
	Security classification	

Signature  Date 2019-12-03

Geospatial data and knowledge on the Web

Knowledge-based geospatial data integration and visualisation with Semantic Web technologies

Weiming Huang

LUND
UNIVERSITY

Coverphoto © Weiming Huang, designed by Jingbei Zheng

Faculty of Science
Department of Physical Geography and Ecosystem Science

ISBN (print): 978-91-985015-1-3
ISBN (PDF): 978-91-985015-2-0
Printed in Sweden by Media-Tryck, Lund University
Lund 2020

且夫水之积也不厚，则其负大舟也无力。
覆杯水于坳堂之上，则芥为之舟，置杯焉则胶，水浅而舟大也。
风之积也不厚，则其负大翼也无力。
故九万里，则风斯在下矣，而后乃今培风；
背负青天，而莫之夭阏者，而后乃今将图南。

《庄子·逍遥游》

If there is not sufficient a depth, water will not float large ships. Upset a cupful into a hole in the yard, and a mustard-seed will be your boat. Try to float the cup, and it will be grounded, due to the disproportion between water and vessel. So with air. If there is not sufficient a depth, it cannot support large wings. And for this bird, a depth of ninety thousand li is necessary to bear it up. Then, gliding upon the wind, with nothing save the clear sky above, and no obstacles in the way, it starts upon its journey to the south.

Chapter Carefree Excursion in Zhuangzi

Translation from Lin Yutang

Abstract

Geospatial information is indispensable for various types of spatially informed analysis and decision-making, such as traffic analyses, and natural resource management. In addition, geospatial information is one of the most powerful information integrators to bridge diverse sources of information. Such natures of geospatial information entail the need of geospatial data integration and geospatial knowledge outreach.

In most of real-world applications, geospatial data from one single source can hardly suffice. Therefore, integrating multi-source geospatial data is a predominant need for a variety of applications. However, today's solutions for geospatial data integration in spatial data infrastructures (SDIs) are inadequate, and the data are often stored in the so-called "data silos", i.e. datasets are stored mostly isolated from each other.

Semantic Web technologies provide a promising way for geospatial data integration on the Web. In this thesis, Semantic Web technologies are utilised to integrate multi-source geospatial data or integrated geospatial data with data from other domains. Paper I leveraged Linked Data and ontologies to realise a relative positioning approach, which positions thematic data based on their relations with background data. The relatively positioned thematic data can be automatically synchronised in terms of their geometric representations in all scales to avoid substantial discrepancy. Paper II integrated distributed multi-scale building data and a heritage building dataset to accomplish a heritage building map with both fine geometries and thematic information of heritage building. Paper III identified that only using ontologies is inadequate for integrating geospatial data and data from other domains, where complex and subtle semantic relations often arise. Then, a knowledge-based framework coupling ontologies and semantic constraints is developed to tackle the complex semantic relations raised by multiple representations of geospatial data.

Besides data integration, there is another prominent need for utilising geospatial data for various applications, that is, the outreach of geospatial knowledge. Visualisation, as one of the most predominant ways of utilising geospatial data, pertains to a wide range of cartographic knowledge. Therefore, it is desirable to formalise the knowledge of geospatial data visualisation (geovisualisation) to facilitate its interpretation, transfer, and reuse.

In this context, Semantic Web technologies offer a framework to formalise and share geovisualisation knowledge, thanks to their knowledge representation capacity. To this end, Paper II and III formalised the knowledge of geovisualisation in knowledge bases with ontologies and semantic rules. Such knowledge bases are evaluated in two real-world applications, i.e. heritage building mapping, and urban bicycling suitability mapping. The knowledge bases for geovisualisation can be used as a visualisation enablement layer for geospatial Linked Data.

In addition, Paper IV performed a study for the technical environment of the support for geospatial Semantic Web (Linked Data). It assessed and benchmarked several well-known and mainstream Linked Data stores, mainly in terms of their spatial query capacities and standard compliance. The results demonstrated that the support for geospatial Linked Data and queries has becoming increasingly mature. Nevertheless, query correctness remained a challenge for cross-database interoperability.

In conclusion, this thesis provides insights into the potentials of Semantic Web technologies for geospatial data integration and knowledge outreach (sharing). The insights could benefit the development of the next generation of SDIs, in which Semantic Web and Linked Data will expectedly play a role. From the perspective of Semantic Web research, the thesis contributes to the modelling and representation of geospatial data and knowledge on the Semantic Web.

Sammanfattning

Geografisk information är oumbärlig för spatiala analyser och beslutsfattande, såsom trafik- och miljöanalyser. Dessutom är geografisk information i många fall mycket användbart för att integrera informationskällor. Därför är det viktigt att studera teorier och metoder för integration och representation av geografisk information.

I de flesta applikationer räcker inte geografiska data från en enda källa och därför behövs integrerring av data från flera källor. Dagens lösningar för dataintegration i infrastrukturer för geografiska data är emellertid otillräckliga, och informationen lagras ofta i så kallade "datasilos", dvs. datamängder lagras oftast isolerade från varandra.

Semantisk webbteknik har potential för integration av geografiska data på webben. I den här avhandlingen används semantisk webbteknik för att integrera geografiska data från flera källor och för att integrera geografiska data med data från andra domäner. Artikel I utnyttjar länkade data och ontologier för att realisera en relativ positioneringsmetod, som positionerar tematiska data baserat på deras relationer med bakgrundsdata. De relativt positionerade tematiska data synkroniseras då automatiskt med bakgrundsdata i alla skalor. Artikel II integrerar distribuerade byggnadsobjekt i flera skalor och en datamängd för kulturarv för att uppnå en kulturarvskarta med både detaljerad geometri och tematisk information om kulturarv. I artikel III identifieras att endast användning av ontologier är otillräcklig för att integrera geografiska data och data från andra domäner, där komplexa och subtila semantiska relationer ofta uppstår. I artikeln utvecklas ett kunskapsbaserat ramverk som kopplar samman ontologier och semantiska begränsningar för att hantera de komplexa semantiska relationer som uppstår genom att samanalysera flera representationer av geografiska data.

Det är också viktigt att utveckla teorier och metoder för att representera kunskap om geografisk information. Visualisering, som är ett av de mest dominerande sätten att använda geografiska data, är baserat på kartografiska teorier. Därför är det önskvärt att formalisera (den kartografiska) kunskapen om geografisk datavisualisering (geovisualisering) för att underlätta dess tolkning, överföring och återanvändning. Semantisk webbteknik erbjuder ett ramverk för att formalisera och dela kunskap om geovisualisering, tack vare deras kapacitet för kunskapsrepresentation. För detta ändamål formaliserade vi i artikel II och III kunskapen om geovisualisering med

stöd av ontologier och semantiska regler. Sådana kunskapsbaser utvärderas i två praktiska tillämpningar: visualisering av kulturminnesmärkta byggnader och visualisering av säkerhetsaspekter för cykling i urbana miljöer.

Artikel IV innehåller en studie av lagring och åtkomst av geografiska data i form av länkade data. I studier utvärderades och jämfördes flera välkända och vanliga databaser för länkade data (s.k. tripple stores), främst vad det gäller kapacitet för åtkomstfrågor och hur väl de implementerar internationella standarder. Resultaten visade att stödet för länkade geografiska data har blivit bättre jämfört med studier genomförda för några år sedan. Dock finns det fortfarande kompabilitetsproblem mellan databaserna.

Sammanfattningsvis ger denna avhandling fördjupad kunskap om potentialen för semantisk webbteknik för geografiska data. Kunskapen gynnar utvecklingen av nästa generation av infrastrukturer för geografiska data, där den semantiska webben och länkade data förväntas spela en viktig roll. Från perspektivet av forskning inom den semantiska webben bidrar avhandlingen med modellering och representation av geografiska data och kunskap om dessa data.

摘要

地理信息对于多种多样的空间相关的分析与决策至关重要，例如交通分析和自然资源管理。同时，地理信息可以作为一种桥梁来架通多种多样的数据。因此，地理数据的融合和地理知识的表示十分重要，可以促进地理数据被合理地应用在多种多样的场景中。

在很多现实的地理信息应用场景中，单一来源的地理数据很难满足需要。因此，对多源地理数据的融合已经成为了必不可少的前提条件。然而， 当前在空间数据基础设施中所提供的数据融合的解决方案远远不够。现在的解决方案造成了一个个“数据孤岛”，也就是说多种来源的数据相互之间是隔离的，很难相互沟通。

语义网技术（现又称为知识图谱技术）为我们提供了一个可以在网络上对多源地理数据融合的框架。在这篇博士论文中，语义网技术被应用于融合多源地理数据，与融合地理数据和其他领域的数据。在这个背景下，文章一利用关联数据和本体技术实现了一个相对定位方法。这个方法不使用传统的描述地理要素空间位置的方式，转而使用主题数据与背景数据的关系来对地理要素进行建模。这样的方法可以使得网络地图有更好的可视化效果，并在很大程度上避免不同尺度下的可视化的偏差。文章二利用了语义网技术融合了多尺度建筑物数据和文物建筑数据，来实现了一个既拥有精细几何图形又有主题属性（建筑年代）的文物建筑地图。文章三发现了单单使用本体无法表示一些在融合地理数据和其它领域的数据中出现的复杂的语义关系。因此，我们提出结合使用本体和语义约束来表示这样的复杂关系。

除了数据融合以外，应用地理信息于实际场景中还有一个难点，那就是怎样合理地使用地理数据。在这其中，地理可视化可以说是地理数据最为常见且广泛的应用方式之一，然而可视化并不简单，这个过程涉及到很多的地图可视化和制图知识。我们需要一种方法来对这样的知识进行形式化，来使得这样知识更容易被人类和计算机所使用和理解。

在这个背景下，语义网技术可以帮助我们对地理可视化的知识进行形式化，因为语义网有强大的知识表示的能力。文章二和文章三使用本体和语义规则对地图可视化的知识进行了形式化，并讲这样的知识封装在知识库中。这些知识库被应用于两个现实应用中：文物建筑地图的绘制与城市自行车适宜性的地图绘制。我们认为这样的知识库可以用作语义网中的一个对于其中地理数据的可视化层。

另外，文章四进行了一个对主流关联数据数据库的空间数据查询能力的评估。它对于五个数据库的评估展示出这些数据库对于地理数据的支持已经比过去更强大，且很多符合空间数据查询的标准。然而，这些数据对于同样的查询并不总是给出同样的答案，这在很大程度上给跨数据库的互操作造成了困难。

总的来说，这篇博士论文对使用语义网技术进行地理数据的融合和知识表示给出了新的见解。这些成果对于下一代空间数据基础设施的开发将会有一定帮助，因为语义网技术有一定可能会在其中扮演一个重要角色。从语义网技术研究的角度来讲，这篇博士论文对在语义网中地理数据和知识的建模与表示提出了新的思路。

Acknowledgment

For me (perhaps also for many others), the PhD study has been a both exciting and painful journey. There have been many ups and downs. For many times, I thought of my works as brilliant ideas, and then I shortly started to question the validity and usefulness of my ideas. In such cases, it is always worth digging into literature to find some arguments to persuade and comfort myself (this is particularly important when the doubts got stuck in my head at night). I remember the excitement when my first paper got accepted, and the disappointment when another paper got denied in its first form. These moments and all other pieces of life formed the very four years and three months in my PhD study, in the wonderful medieval town Lund.

I feel extremely lucky and grateful that I have had so many extraordinary researchers to work with. My foremost gratitude goes to my main supervisor Prof. Lars Harrie. In fact, Lars profoundly shaped my modes of scientific thinking and developing academic capacity, particularly at the very beginning of the PhD study, when I even could barely speak decent English. Lars cared and stimulated me in an unselfish manner. He has been always supportive to me in pursuing the research topics that I am interested, patiently listening to my ideas, valid or invalid, and helping me develop my network with other researchers. Moreover, his passion and skills in teaching have affected me dramatically, for which I made a smooth transition to enjoy teaching in several courses and supervision of master theses.

My second supervisor, Dr. Ali Mansourian, has also been supportive throughout this journey. Ali has always been open-minded and constructively commenting on my ideas and works, which turned out to be useful additions. Ali and I have been collaborating quite much in teaching, and I believe that he is an excellent mentor in this respect. Working with Ali in both scientific research and teaching has been delightful and encouraging.

During the latter part of 2019, I had a splendid research visit at the Center for Spatial Studies, University of California, Santa Barbara (UCSB), where I was honoured to work with world-leading researchers in geospatial semantics/Semantic Web – Prof. Werner Kuhn and Prof. Krzysztof Janowicz. Werner and I had many interesting discussions in the topics of ontologies, semantics, core concepts of spatial information, and the Semantic Web, which were definitely enlightening. Jano is an extraordinary researcher, who is incredibly passionate, and energetic; I benefited quite much from discussions with Jano. I enjoyed every single group meeting in

both the Center for Spatial Studies and the STKO lab, which had led me to see many remarkable scientific works closely, and largely refreshed my understanding and thoughts. My research visit to UCSB was a fantastic chapter in this journey. I wholeheartedly appreciate the advice and guidance from Werner and Jano, and many other colleagues there.

During the course of my PhD study, I was luckily surrounded by quite a number of brilliant researchers and colleagues, which has not only benefited my studies, but also partially shaped my personality with open-mindedness, patience, and humbleness. Special thanks to my colleagues and friends at GIS Centre, INES, and Geocentrum II in Lund, including, but definitely not limited to, Andreas, Bingjie, Ehsan, Eva Andersson, Eva Kovacs, Feng, Finn, Hans, Hongxiao, Jing, Jonas Ardö, Karin, Minchao, Mitch, Olive, Per-Ola, Petter, Reza, Roger, Sha, Tomas, Torbern, Wenxin, Yao, Yanzi, Yixin, Zhanzhang, Zhao Li, Zhendong, Zheng, and Zhengyao. Special thanks are sent to Petter Pilesjö, who made the GIS Centre a lovely and wonderful place. Many thanks also go to my friends and colleagues out of the geo-side in Lund, to name just a few, Amir, Bofei, Khashayar, Shengnan, Zhao Liu, and Zhuo. Many faces and names are coming into my head as I am writing this passage. In fact, I have long realised that I would not be able to make it without the company and support from many others.

I would also like to express my appreciation to my fellow students and friends that I met during my visit at UCSB. It is always difficult for me to integrate into a completely new environment, for which I particularly thank the persons who made this process way easier than I thought: Behzad, Blake, Gengchen, Jianyu, Jingyi, Katja, Ling, Meilin, Mingyu, Rui, Sara, Thomas C, Thomas H, and Zhaodong. I enjoyed every moment at UCSB, and I sincerely hope I could meet every one of you again in the future.

I also very much appreciate the fabulous environments in both Lund and Santa Barbara, which helped me relax and reflex nearly every single day. In Lund, I usually take a certain path from the department to its east to have a stroll, which has been a great help to alleviate my pressure. On the campus of UCSB, I expected and enjoyed every afternoon-walk to the Campus Point. I believe I could never forget the spectacular scene of sunset in the sea there.

I sincerely appreciate all my other friends in my life: old friends from my childhood, from my study in my schools and collage, etc. They are one of the essential foundations of my life, and have motivated me throughout this journey.

Finally, I cannot express enough gratitude to my family: my grandma (from my mother's side), my parents, my sister, her husband, and their son, for their endless and tireless support for me to pursue what I am interested in. When I felt down (in fact for many times), I could always be stimulated by them and what they had to endure. My appreciation also goes to my grandparents (from my father's side). I

wish so hardly that you could know that your grandson tried his best to reach his next millstone in his life. I miss you so much all the time. I have been trying to become a person like my family members, who is open-minded, reasonable, loving, and reliable.

最后，我想要表达我对家人的感谢。是你们一直在背后的支持让我可以追求自己的梦想。也是你们一直在勉励我成为一个虚心，理性，且可靠的人。

Content

1 Introduction

1.1 Motivation

Over the last decades, massive use of geospatial information in various real-world applications (e.g. traffic analysis, and built environment processes) gradually reveals the indispensable role of geospatial information for spatially informed analysis and decision-making (Kuhn, 2012). At the meantime, geospatial information is one of the most powerful information integrators to bridge diverse sources of information (Janowicz et al., 2012). Such natures of geospatial information entail the need of geospatial data integration and geospatial knowledge outreach.

In most of spatially informed analyses and applications, geospatial data from a single source can hardly suffice. For instance, developing a Swedish heritage building map requires data from *Lantmäteriet* (Swedish National Mapping Agency) with detailed geometric representations of the building footprints and a base map, and from *Riksantikvarieämbetet* (Swedish National Heritage Board) with heritage information, e.g. construction years; the information from Wikipedia could also be a useful addition for users. These different sources of (geospatial) data should be properly integrated for visualisation and analyses. Another example of is the integration of authoritative geospatial data and volunteered geographic information (VGI), e.g. points of interest (POIs), in which each source has its own unique information for the applications such as wayfinding (Yang et al., 2014). Therefore, data integration (including the integration of multi-source geospatial data and the integration between geospatial data and other types of data that can be grounded geographically) plays a pivotal role in geospatial visualisation and analysis.

Today's geospatial data are mainly maintained and disseminated through spatial data infrastructures (SDIs) that aim to make geospatial data available for the benefit of the economy and the society (van den Brink et al., 2017). In Europe, the INSPIRE directive – a legal framework and standardisation body for SDI development – sets the data specifications, and mandates its member states to provide data mainly using Open Geospatial Consortium (OGC) Web services (INSPIRE, 2018). SDIs have partially achieved dissolving environmental and geospatial data held in silos, but the data are still largely isolated from other information domains (Schade and Smits,

2012). For example, the OGC web feature service (WFS) can make geospatial data available through its data query protocol, yet such data cannot be discovered by search engines or, more importantly, linked by other data resources. Another significant issue in SDIs is semantic heterogeneity, which is an impediment for data integration, as the semantics of metadata, schemas, and data content are usually not harmonised (Lutz et al., 2009). The above limitations undermine the discovery and (re-)usability of the data.

Moreover, the knowledge concerning how to appropriately use geospatial data is important. There have been many endeavours to develop theories for the visualisation and analysis of geospatial data, whereas few of them have been outreached and can be readily used, especially for experts from other domains. Today experts from other domains still often have to look into the literature, or cooperate with geospatial experts to accomplish meaningful use of geospatial data. Visualisation, as one of the most predominant ways of utilising geospatial data, is knowledge-intensive and entails many semantic intricacies (Scheider and Huisjes, 2019), as visualising geospatial data in a sense-making and cartographically satisfactory way pertains to a wide range of cartographic knowledge, which is difficult to transfer, interpret, and reuse, especially by non-geospatial experts (MacEachren, 2004). In this context, formalising the visualisation knowledge can potentially foster the knowledge outreach for wider users.

Over the last decade, Semantic Web technologies, particularly the parts relevant to Linked Data, have been increasingly adopted in the geospatial domain, which unveils a promising means to unravel the above discussed limitations of geospatial data integration and visualisation. Semantic Web technologies provision mechanisms for integrating and interlinking geospatial data on the Web in a distributed manner; they allow for lifting semantic harmonisation level with formally defined ontologies; and the knowledge representation capacity of this technology stack provides a promising way to represent and share geospatial (visualisation) knowledge on the Web to foster wider use of such knowledge and spatially enable the Web (Schade and Smits, 2012). A recent survey conducted in 2018 by EuroSDR (European Spatial Data Research) demonstrated that Linked Data is seen as one of the most important research issues and key factors moving SDIs toward the next generation (EuroSDR, 2019). Linked Data was also voted one of the most important SDI research topics during the AGILE 2018 workshop 'SDI research and strategies towards 2030'[1].

Promising methods and results have been delivered in geospatial data integration leveraging Semantic Web technologies. However, there are still some gaps that are

[1] https://kcopendata.eu/sdi2030/

necessary to fill to further unlock the potential of Semantic Web technologies for geospatial data integration and visualisation.

Multiple (geometric) representation is a special integration problem of geospatial data. Multiple representations delineate the geographic space with several abstraction levels (e.g. a building can be represented as a point or a polygon), and thereby enable visualisation and analysis at different scales. Integrating data with Semantic Web technologies can be problematic when the data have multiple representations, as the concepts used for data with different representations seem the same, but are not applied in the same way in data (van den Brink et al., 2017). Multiple representations sometimes arise difficulties when incorporating geospatial data for visualisation and analysis. For example, some certain analyses need geospatial data from different detailed levels, which embodies many semantic intricacies (c.f. Paper III). Therefore, how to organise and integrate multiple representation geospatial data with Semantic Web technologies needs to be further explored.

Moreover, Semantic Web technologies have potential to formalise knowledge concerning how to appropriately visualise geospatial data. The formalised knowledge can potentially compose knowledge bases to derive visualisations. Such knowledge bases can be readily shared on the (Semantic) Web to facilitate the knowledge transfer, interpretation, and reuse. This is a way to migrate niched geospatial knowledge into commonly-used and versatile information infrastructure, and it is promising to outreach geospatial knowledge to wider users. However, the studies utilising Semantic Web technologies for formalising geovisulisation (visualisation of geospatial data) is sparse, but worth exploring.

1.2 Research questions and objectives

The overarching research question of this thesis is *what are the benefits of Semantic Web technologies for geospatial data integration (in particular for multiple representation data) and the formalisation of geovisualisation knowledge?* This overall research question can also be phrased from a feasibility perspective: *is it possible to leverage Semantic Web technologies to accomplish the integration of geospatial data with multiple representation and geospatial data with other types of data?* and *is it possible to utilise Semantic Web technologies to formalise geovisualisation knowledge and thus share it on the Web?*

In this framework, we formulate several specific research questions focusing on real-world applications that can be potentially better addressed by Semantic Web technologies:

1. Geospatial data are often repetitively generated despite the intrinsic relations between objects. Is it possible to link geospatial objects to existing objects in the Semantic Web to diminish data repetition and inconsistency?
2. The knowledge concerning how to visualise geospatial data is important. Is it possible to use Semantic Web technologies to formalize such knowledge, and thus share it on the Web?
3. Multiple representation of geospatial data sometimes makes data integration complex and problematic. Is it possible to leverage Semantic Web technologies to formalize the knowledge of multiple representation and assist cross-detailed-level data integration?
4. The utilization of Semantic Web technologies and Linked Data for geospatial data entails the need of platforms for managing, storing, and querying such data (geospatial Linked Data). How is the support and performance of Linked Data stores for geospatial data, in particular for spatial queries and compliance to standards?

Therefore, the aim of this thesis is to develop methods based on Semantic Web technologies to facilitate geospatial data integration and visualisation, and thus strengthen the (re-)usability of geospatial data in real-world problem-solving and decision-making. The aim consists of four research objects, and each of them corresponds to one or more papers: (1) to develop methods for relatively positioning geospatial features based on their relations with background data using Linked Data (Paper I); (2) to exploit the knowledge representation capacity of Semantic Web technologies for formalising the knowledge of geospatial data visualisation (Paper II & Paper III); (3) to develop methods dealing with the cross-detailed-level data integration problem with Semantic Web technologies (Paper III); (4) to assess and benchmark widely-used and well-known Linked Data stores to understand where the methods can be better applied in, and bring insights to the (geospatial) Linked Data community at large (Paper IV).

1.3 Thesis organisation

Following the introduction, Chapter 2 provides a background and related works for this thesis. Section 2.1 is the preliminaries of Semantic Web technologies, providing the technical background for this thesis. Section 2.2 provides an overview and related works of geospatial data integration, with a focus on geospatial data integration using Semantic Web technologies. Section 2.3 describes relevant studies of geospatial knowledge representation, for both geovisualisation and geoprocessing. Chapter 3 summarises the papers that are included in this thesis. Chapter 4 concludes this thesis and provides outlooks in this research topic.

1.3.1 List of papers

I. Huang, W., Mansourian, A., Abdolmajidi, E., Xu, H., & Harrie, L. (2018). Synchronising geometric representations for map mashups using relative positioning and Linked Data. *International Journal of Geographical Information Science, 32*(6), 1117-1137. doi: 10.1080/13658816.2018.1441416

II. Huang, W., & Harrie, L. (2019). Towards knowledge-based geovisualisation using Semantic Web technologies: a knowledge representation approach coupling ontologies and rules. *International Journal of Digital Earth*, Advance online publication. doi: 10.1080/17538947.2019.1604835

III. Huang, W., Kazemzadeh, K., Mansourian, A., & Harrie, L. (under review). Towards knowledge-based geospatial data integration and visualization: a case of visualizing urban bicycling suitability.

IV. Huang, W., Raza, S. A., Mirzov, O., & Harrie, L. (2019). Assessment and Benchmarking of Spatially Enabled RDF Stores for the Next Generation of Spatial Data Infrastructure. *ISPRS International Journal of Geo-Information, 8*(7), 310. doi: 10.3390/ijgi8070310

1.3.2 List of contribution

I. **WH** conceived and designed the methodology mainly together with LH and AM; WH prepared the data, and implemented the method of the study together with HX and EA; **WH** interpreted the results together with the co-authors and led the writing.

II. **WH** led the study design, carried out the practical part of the study, interpreted the results together with the co-authors and led the writing.

III. **WH** led the study design and carried out the practical part of the study, interpreted the results together with the co-authors and led the writing.

IV. **WH** designed the study with OM and LH; **WH** carried out the practical part of the study together with SAR; **WH** interpreted the results together with the co-authors and led the writing.

1.3.3 Related papers

The author has also been involved in the following related papers.

- Huang, W., 2019. Knowledge-based geospatial data integration and visualization with Semantic Web technologies. In *International Semantic Web Conference 2019 doctoral consortium*, October 26-30, 2019, Auckland, New Zealand.
- Huang, W., 2018. Towards knowledge-based integration and visualization of geospatial data using Semantic Web technologies. In *Doctoral Consortium and Challenge at RuleML+ RR, RuleML+ RR-DCC 2018*, Luxembourg, September 18-21, 2018.
- Huang, W., Mansourian, A., Harrie, L., 2018. Geospatial data integration and visualisation using Linked Data. In: In *Proceedings of the 4th AGILE PhD School* (eds. Lex Comber & Nick Malleson), Leeds, UK, October 30 - November 02, 2017.

2 Background and related works

2.1 Preliminaries of Semantic Web technologies

As this thesis exploits Semantic Web technologies to facilitate geospatial data integration and visualisation, this section briefly introduces the preliminaries of Semantic Web technologies as part of the background.

2.1.1 Semantic Web and Linked Data

The development of the Web has come through several generations. In the first stage, Web 1.0, there were only few content creators with the huge majority of users who were only consumers of content. Afterward, Web 2.0 came on stage highlighting user-generated content, usability and interoperability for end users, which is also called participative social Web. With these two stages, the largest portion of information available online has been made available for human users, mostly in the form of hypertext augmented with images and other kinds of multimedia (Keßler, 2010). However, the content is mostly oriented to be understood by human users, but not machines.

The Semantic Web (also denoted Web 3.0 and Web of data) remedies the abovementioned limitation to allow machines to understand content on the Web and to enable meaningful communications between machines as well as between humans and machines (Berners-Lee et al., 2001). The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries (W3C, 2013). It is an extension of the current Web, not a replacement. The Semantic Web is built upon a stack of enabling technologies (see Figure 1).

Today's Semantic Web vision is mainly realised by building distributed data repositories following a number of recommended best practices for exposing, sharing, and connecting pieces of data, information, and knowledge across the Semantic Web using Uniform Resource Identifiers (URIs) and Resource Description Framework (RDF). The data published following the best practices are denoted *Linked Data*. Specifically, Berners-Lee (2009) established four best practices and principles of Linked Data:

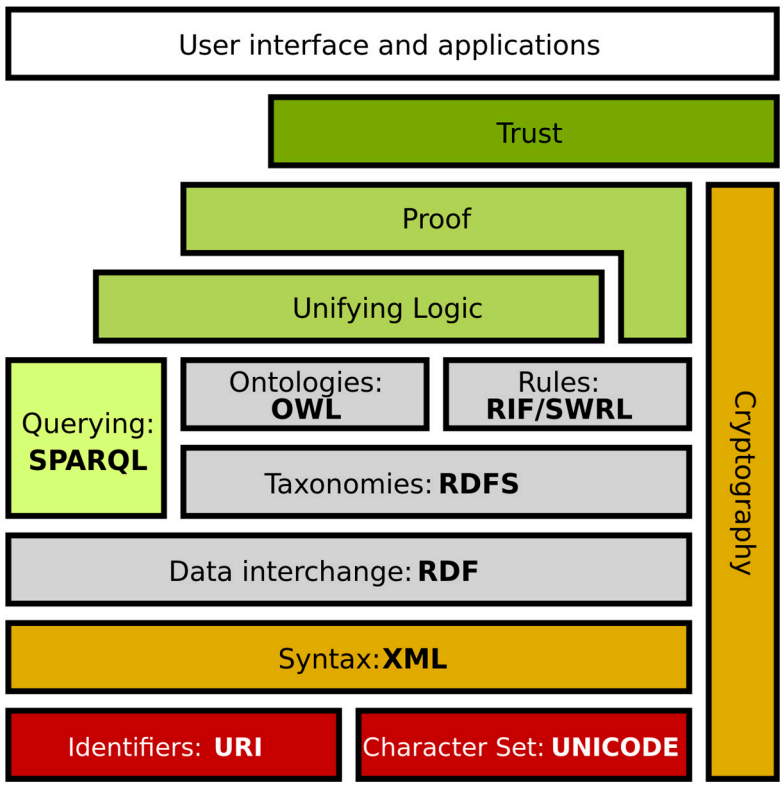


Figure 1. The Semantic Web stack, which illustrates the enabling technologies and how they build upon each other. Adopted from Semantic Web Stack (2019).

- URIs should be used to denote things.
- HTTP URIs should be used so that these things can be referred and dereferenced (looked up) by human users and software agents.
- W3C standards such as RDF and Web Ontology Language (OWL) should be used, so that useful information can be provided when looking up the URIs.
- Data should be interlinked using URIs to create a densely interconnected graph of knowledge (the Linked Open Data (LOD) cloud).

With the data model RDF, each piece of information (a statement) is constituted by three elements, i.e. a subject, a predicate, and an object. This simplest form, in which statements can be made in natural language, is an essential ingredient for linked data. Numerous such triples can express, share, and seamlessly integrate any data. Below is an excerpt of information regarding Lund Cathedral (*Lunds domkyrka* in Swedish)

from GeoNames[2] (a Linked Data gazetteer), describing the entity type, name, country, and geographic location. The RDF statements are in the syntax of Turtle.

```
<http://sws.geonames.org/8128831/> a <http://www.geonames.org/ontology#Feature>;

        <http://www.geonames.org/ontology#name> "Lunds domkyrka";

        <http://www.geonames.org/ontology#countryCode> "SE";

        <http://www.w3.org/2003/01/geo/wgs84_pos#lat> "55.70397";

        <http://www.w3.org/2003/01/geo/wgs84_pos#long> "13.19405".
```

This excerpt states that an entity (<http://sws.geonames.org/8128831/>) is a feature (<http://www.geonames.org/ontology#Feature>); its name is Lunds domkyrka; it belongs to Sweden (with the country code 'SE'); and it locates at a geographic position represented in longitude and latitude.

Linked Data (RDF data) has the standardised query language SPARQL (W3C, 2008). SPARQL can be used to express queries across diverse or distributed data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. Below is an example SPARQL query to retrieve the feature in GeoNames whose name is *Lunds domkyrka*.

```
SELECT ?feature

WHERE{

        ?feature a <http://www.geonames.org/ontology#Feature>;

                <http://www.geonames.org/ontology#name> "Lunds domkyrka".

        }
```

A SPARQL query engine will return the result <http://sws.geonames.org/8128831/> to this query according to the example RDF data above.

2.1.2 Knowledge representation in the Semantic Web

One prominent advantage of harnessing Semantic Web technologies is the inherent knowledge representation capacity equipped with the technology stack. Knowledge representation is a branch of symbolic artificial intelligence, which studies the formalisation of knowledge and its processing within machines (Grimm et al. 2007). Since 1960s, the focus of knowledge representation has evolved through several stages, including general problem solver, expert systems, frame based languages, and rule-based systems, and currently one of the most active areas of knowledge

[2] https://www.geonames.org/

representation research is the Semantic Web. The Semantic Web provisions us with the capacity for representing knowledge, supporting search queries on knowledge and inference. In the Semantic Web, knowledge is represented in different forms, and ontologies (descriptions logics) and rules (horn logic) are the two main paradigms for knowledge representation (Hitzler and Parsia, 2009).

The term ontology is borrowed from philosophy, where ontology is a systematic account of Existence. In computer and information science, ontologies are controlled vocabularies that describe concepts and relations between concepts using well-understood formal constructs; such constructs formalise the intended meaning of the vocabularies and capture background knowledge about the domain (Horrocks, 2008). Figure 2 illustrates the core part of GeoSPARQL ontology – an OGC standard for representing and querying geospatial Linked Data. This diagram shows that the classes *Feature* and *Geometry* that two subclasses of *SpatialObject*, and *Feature* and *Geometry* have the relation *hasGeometry* (the GeoSPARQL prefix is represented as *geo*).

In the Semantic Web, the most prominent and commonly used ontologies include RDF Schema (RDFS), and OWL. RDFS provides a data modelling vocabulary for RDF data and includes some basic semantics, e.g. *rdfs:Class*, *rdfs:subClassOf*, and *rdfs:property* (W3C, 2014). OWL is a Semantic Web language designed to represent rich and complex knowledge about things, and relations between things. OWL is based on computational logic language so that knowledge expressed in OWL can be exploited by computer programs (W3C, 2012). OWL has several sub-languages which corresponding to different languages from the family of Description Logics (DLs). Three sub-languages are OWL-Lite, OWL-DL and OWL-Full (Baader et al., 2003). OWL-Lite primarily supports the need of classification and simple constraints. OWL-DL supports the maximum expressiveness while retaining computational completeness (all conclusions are computable) and decidability (all computations can be finished in finite time); OWL-DL includes all OWL language constructs, but certain restrictions are

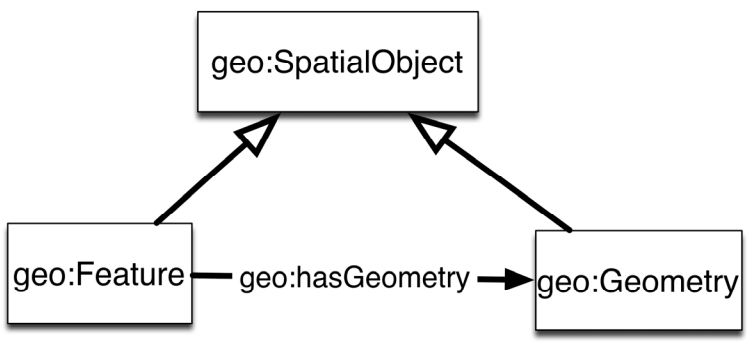


Figure 2. The core part of GeoSPARQL ontology. Adopted from Bermudez (2012).

imposed, e.g. a class may be a subclass of many classes, yet a class cannot be an instance of another class. OWL-Full is meant for users who want maximum expressiveness and the syntactic freedom of RDF but with no computational guarantee; it is unlikely that any reasoner will support complete reasoning for all the features of OWL-Full (W3C, 2004). In 2012, W3C introduced OWL 2 as a recommendation with several additions than its previous version such as property chains, and richer datatypes (W3C, 2012). OWL 2 inherits the sub-languages with different levels of expressivity. Example concepts or relations of OWL include *owl:Thing*, *owl:equivalantClass*, and *owl:ObjectProperty*.

Despite the generally powerful reasoning ability, OWL is still limited in expressing more complex inference. One commonly-used example of ternary relations is the *uncle* example – if *a* is a brother of *b*, and *b* is a parent of *c*, then *a* is *c*'s uncle. Another example that is more relevant to geospatial data and knowledge is that if A is a building and its age is more than 500 years, then use a certain symboliser (encapsulating a set of parameters for visualisation) to visualise the building on the map. Such knowledge and inference cannot be performed with OWL. Therefore, many approaches have been developed to lift the expressiveness by combining ontology languages with rules. One of the most prominent developments is the semantic web rule language (SWRL) that combining OWL and RuleML (Horrocks et al., 2004). SWRL extends OWL with horn-like rules. The SWRL rules have the form:

$$\text{antecedent} \Rightarrow \text{consequent}$$

The abovementioned uncle rule can be represented in SWRL as:

$$\text{brother}(?a,?b) \wedge \text{parent}(?b,?c) \Rightarrow \text{uncle}(?a,?c)$$

SWRL rules adopt the Open World Assumption (OWA; the assumption that what is not known to be true or false can possibly be true)[3] , and thus only support monotonic inference (the knowledge base grows with new facts in a monotonic fashion). Consequently, SWRL rules do not support including negation, e.g. SWRL cannot express the rule – if the age information of a building does not exist, then use a certain symboliser to render it on the map, as this rule entails non-monotonic reasoning.

Another type of commonly-used semantic rules is based on SPARQL, as SPARQL is able to deduce new statements from known facts. In this context, the object-oriented SPIN (SPARQL Inferencing Notation) rules, which combine concepts from object-oriented languages, SPARQL query language, and rule-based systems to model rules and data quality constraints in the Semantic Web, have been developed (Knublauch et al., 2011). In contrast to SWRL, SPIN has better

[3] https://en.wikipedia.org/wiki/Open-world_assumption

expressiveness in terms of handling non-monotonic semantics, and in principle it could readily allow spatial predicates (e.g. in GeoSPARQL) to be embedded in the condition of the rules within spatially enabled RDF stores (Linked Data stores). Below is an example SPIN rule deducing the area of a rectangle from its width and height.

```
ex:Rectangle

              spin:rule [

              rdf:type sp:Construct ;

              sp:text """

              CONSTRUCT {

                            ?this ex:area ?area .

                            }

              WHERE {

                            ?this ex:width ?width;

                                  ex:height ?height.

                            BIND ((?width * ?height) AS ?area) .

                            }""" ;] .
```

SPIN rules are associated with the class (*ex:Rectangle*) that they apply to due to the object-oriented nature of SPIN. The inferred statement(s) comes after the keyword *CONSTRUCT*, and the conditions are nested after the *WHERE* keyword. For details of modelling SPIN rules, see Knublauch (2011).

The upgrade of SPIN – SHACL (Shapes Constraint Language) has been developed and become a W3C recommendation in 2017 (Knublauch and Kontokostas, 2017). SHACL primarily is a language for validating RDF graphs, and can also be used for other purposes including, among others, data integration. The emergence and recommendation of SHACL came from the need for validating and imposing semantic constraints to RDF data (graphs). This need also stemmed from the limitations lie in OWL. Although OWL can express some certain restrictions, e.g. using *owl:maxCardinality* to represent the maximum cardinality constraint, the restrictions only describe the reasoning to be applied based on them (Knublauch, 2017). For example, assuming there is an *owl:maxCardinality 1* restriction stating that one person can only have one gender, and there is a person has two genders (*male* and *female*) due to input mistake, then an OWL reasoned will assume that the two values *male* and *female* must in fact have the same real-world meaning (if disjoint is not explicitly stated). Moreover, OWL adopts OWA so that some certain

constraints cannot be effectively imposed to RDF data. For instance, assuming there is an *owl:minCardinality 1* stating that a person must have one gender value, while such a value does not exist in the RDF data, the OWL reasoner will not report any issue here, because the value may appear at any time and any where to satisfy that restriction under the OWA. SHACL also has advanced features about rule-based reasoning, which are currently not included in the W3C recommendation (Knublauch et al., 2017). Below is an example stating that one person must only have exactly one family name, and the family name must be in the type of *xsd:string*.

```
ex:Person    a rdfs:Class, sh:NodeShape ;

             sh:property [

                          sh:path ex:lastName ;

                          sh:name "last name" ;

                          sh:datatype xsd:string ;

                          sh:maxCount 1 ;

                          sh:minCount 1 ;

             ] .
```

2.1.3 Geospatial Semantic Web and Linked Data

The appreciation of Semantic Web technologies and Linked Data has increased considerably in the geospatial domain in the last decade, and they have fostered a promising approach to connecting SDIs with mainstream IT to augment the application of geospatial data (Schade and Smits, 2012).

Several ontologies have been designed for geospatial data as Linked Data. The commonly used ones include the W3C Basic Geo Vocabulary [4] and the GeoSPARQL vocabulary as an OGC standard. The GeoSPARQL vocabulary has become increasingly popular, as it allows for embedding spatial predicates in queries. The GeoSPARQL vocabulary is lightweight and represents only some fundamental concepts – essentially the concepts of *feature* and *geometry* (cf. Figure 2). Moreover, Pilot studies have been performed releasing INSPIRE-compliant data as Linked Data, and draft guidelines and vocabularies have been developed (INSPIRE, 2017). The developed vocabularies are interoperable with GeoSPARQL. The development of INSPIRE Linked Data's URIs leveraged previous work on the standardisation of unique identifiers for geospatial objects (INSPIRE, 2013).

[4] https://www.w3.org/2003/01/geo/

An increasing amount of geospatial data have been delivered as Linked Data, mainly by governmental agencies and large-scale data infrastructures (Regalia et al., 2018). The UK is a pioneer to this end; Ordnance Survey, Great Britain's national mapping agency (NMA), released several geospatial datasets as Linked Data nearly a decade ago (Goodwin et al., 2008). However, the data relied on unstandardized methods to represent data semantics and thus lacked usability. In the Netherlands, Kadaster delivered several key geospatial datasets, e.g. building data and address data, as Linked Data on the Web, together with other governmental open data, e.g. statistical data (Folmer et al., 2018). In Finland, the National Land Survey piloted the delivery of geographic name data, authoritative data, and building data as Linked Data (Hietanen et al., 2016). In Norway, Kartverket also released some geospatial datasets as Linked Data (Shi et al., 2017). A recent report summarized and reflected on the development of geospatial Linked Data in the Netherlands, Finland, Norway, and Spain (Ronzhin et al., 2019). In the US, several geospatial Linked Data projects have been conducted: a pilot of design and development of Linked Data from The National Map was performed (Usery and Varanka, 2012); the Geographic Names Information System was served as Linked Data, and its geospatial visualization was enabled (Regalia et al., 2018); the GeoLink knowledge graph was published following Linked Data principles and served through a SPARQL endpoint, including Earth Science information captured by oceanographic cruises, physical sample metadata, etc. (Cheatham et al., 2018). Along with these Linked Data, development endeavours from authorities, crowd-sourcing projects have also produced several geospatial linked datasets, and some of them are serving as central hubs of the LOD cloud, e.g. GeoNames, and LinkedGeoData (a Linked Data distribution of OpenStreetMap (Stadler et al., 2012)). Figure 3 illustrates the geospatial (geography) part of the LOD cloud. Moreover, van den Brink et al. (2019) proposed the best practice of delivering geospatial Linked Data, and they bridged the OGC Web services and the Semantic Web.

In practice, Linked Data need to be managed, stored, and delivered by utilizing RDF stores (also known as triplestores), which are databases for storing and retrieving RDF data (Linked Data) through semantic queries (SPARQL queries). As the development of geospatial ontologies and the increasing amount of geospatial Linked Data, more and more RDF stores have supported spatial queries, e.g. with GeoSPARQL. This raises the need of assessing and benchmarking spatially enabled RDF stores. Battle and Kolas (2012) demonstrated the geospatial capacity of Parliament and successfully ran a number of GeoSPARQL-compliant queries. Garbis et al. (2013) presented the benchmark Geographica to assess several spatially enabled RDF stores in which spatial queries were written in both GeoSPARQL and

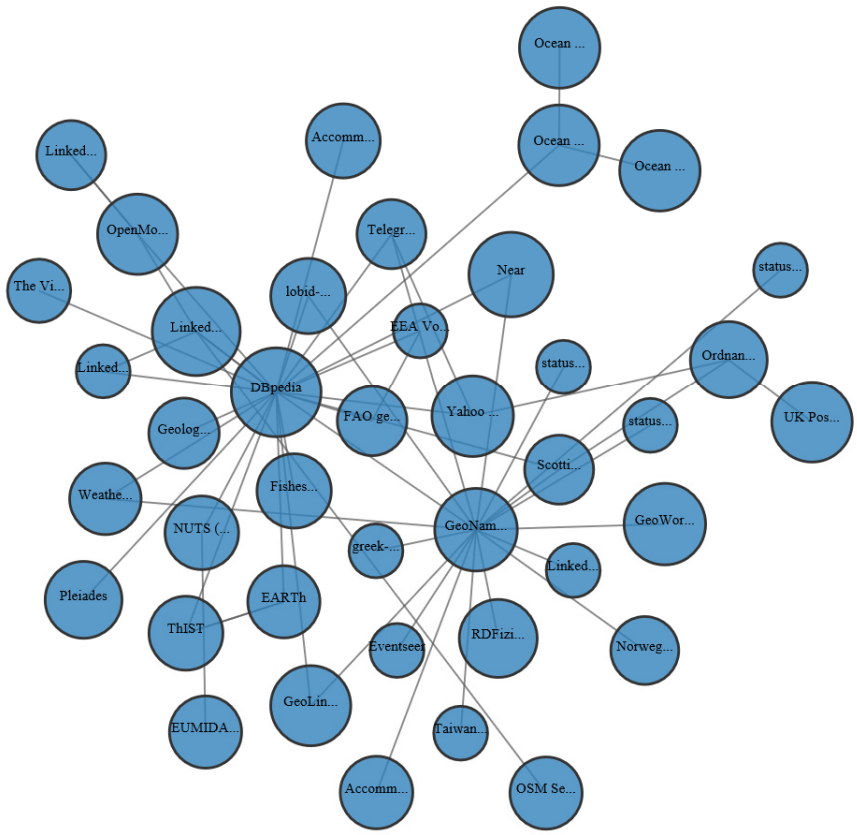


Figure 3. The Geography Linked Open Data Cloud. Adopted from Linked Open Data Cloud (2019).

stSPARQL (the spatiotemporal query language in the RDF store Strabon). In that benchmark, three RDF stores were evaluated, i.e. Strabon, uSeekM, and Parliament, in a micro-benchmark and a macro-benchmark. The micro-benchmark aims to test the efficiency of primitive spatial functions in spatially enabled RDF stores; the macro-benchmark aims to test the performance of the stores in some certain application scenarios, e.g. reverse geocoding, map search, etc. This benchmark's datasets and queries have been published online[5], and the benchmark was based on both real-world geospatial data (e.g. LinkedGeoData) and synthetic data. The GeoKnow project, which dealt with geospatial Semantic Web and Linked Data, released a thorough survey and evaluation of spatially enabled RDF stores, with a partial focus on GeoSPARQL compliance (Athanasiou et al., 2013). The stores evaluated in GeoKnow include Virtuoso, Parliament, OWLIM, uSeekM, and

[5] http://geographica.di.uoa.gr/

Strabon, as well as spatially enabled relational databases, i.e. Oracle Spatial and PostgreSQL with PostGIS extension. Bellini and Nesi (2018) assessed several well-known RDF stores, including Virtuoso, GraphDB, Oracle, and Stardog, for semantically enabled smart city services. The geospatial capacity of these RDF stores was one of the focuses of this study, as smart city services also have the need for capabilities such as temporal data query. The benchmark was based on the Florence Smart City model. Paper IV assessed and benchmarked the mainstream and well-known spatially enabled RDF stores RDF4J, GeoSPARQL-Jena, Virtuoso, Stardog, and GraphDB to provide an updated view of the development of the RDF stores in terms of spatial query capacity. The assessment and benchmarking results demonstrated that the GeoSPARQL compliance of the RDF stores has encouragingly advanced in the last several years. The query performances are generally acceptable, and spatial indexing is imperative when handling a large number of geospatial objects. However, query correctness remains a challenge for cross-database interoperability. The results indicate that the spatial capacity of the RDF stores has become increasingly mature, which could benefit the development of future SDIs.

2.2 Geospatial data integration

This section presents an overview of geospatial data integration, which is one of the aims of this thesis, as well as previous works on geospatial data integration based on Semantic Web technologies.

2.2.1 Overview

The amount of geospatial data available is rapidly increasing on the Web. The data are mainly from governmental, volunteered, scientific and corporate initiatives and activities. In this context, geospatial data integration (also denoted data fusion or data conflation) has become increasingly important for many purposes, including preventing data isolation, enabling cross-dataset analysis and visualisation (Wiemann and Bernard, 2016).

Geospatial data integration syntheses geospatial data from multiple sources to extract meaningful information for applications. Geospatial data integration has been studied from different perspectives and for different purposes, including detecting differences and errors (Samal et al., 2004), transferring feature attributes in a cross-dataset manner (Zhang and Meng, 2007), and eliminating feature duplication (Samal et al., 2004).

Depending on the matching (integration) level, geospatial data integration can be categorised in three levels: representation (feature) level, data schema level, and ontology level (Wiemann and Bernard, 2016). Among them, the feature level mainly focuses on the geographic objects as data content, while data schema and ontology levels mainly deal with matching the conceptualisation of of the data.

The focus of geospatial data integration has partially lied in developing methods and algorithms to effectively discover the correspondence relations between features from different datasets with geometric, attribute, topological, context, and semantic information; see Xavier et al. (2016) for a survey. To name a few, Koukoletsos et al. (2012) developed an automated geospatial linear feature matching method for assessing data completeness of VGI with a multi-stage approach combining both geometric and attribute constraints. Ludwig et al. (2011) proposed a feature-based matching method for comparing street networks in OSM with authoritative data. The method initially creates a candidate list of matching pairs, then linked OSM data within the buffer of reference data and ranks the candidates by names and category attributes. Yang et al. (2014) developed a geometry-based approach for integrating POI with road networks. This method first creates a POI connectivity graph by mining the common linear cluster patterns from POIs; then it utilises probabilistic relaxation to fulfil the nodes matching between the connectivity graph and road network. The POIs are finally matched to the road network after an affine transformation. Kim et al. (2017) proposed a multi-strategy method for linear feature matching, in which the properties of distance, angle, topological relation are incorporated to have increasingly robust matching results; they adopted decision tree to derive thresholds during the matching process. Yang and Gidofalvi (2018) developed a fast matching algorithm to match trajectories with road network by combining hidden Markov model with precomputation, in which an upper bounded origin-destination table is precomputed to store all pairs of shortest paths within a certain length in a road network. Mustière and Devogele (2008) developed an approach for automatically matching networks with different levels of detail (with multiple representations). A multi-step matching process *NetMatcher* was proposed which relies on the comparison of geometric, attributive, and topological properties of objects in the networks. Abdolmajidi et al. (2015) employed two network matching strategies: segment-based and node-based matching for the purpose of evaluating the completeness of the OSM road network. The study demonstrated that the extended node-based approach is sufficiently accurate and efficient for the designed purpose.

Regarding data integration at the schema and ontology level, the purpose is different. The question lies in how interoperability between geographic information systems (GIS) and geospatial databases can be achieved to facilitate data exchange and reuse. In the database research domain, there are different types of techniques for schema matching. It can be merely schema-based, which only takes into account

schema-level information like names, descriptions, datatypes, relationship types (e.g. part-of, is-a), data constraints and schema structure; it can also be approaches incorporating instance-level data (Volz, 2005). For example, Volz (2005) developed a data-driven integration approach for geospatial database schemas, which exploits instance-level relations between multiple representations and their correlation.

Many approaches deal with schema integration issue with ontology matching or other types of semantic technology, and this type of research is common in the geospatial domain. This branch of research is partially the focus of this thesis, and is described in the following section 2.2.2, which illustrates geospatial data integration utilising Semantic Web technologies and within the Semantic Web.

2.2.2 Geospatial data integration with Semantic Web technologies

The studies concerning geospatial data integration using Semantic Web technologies, particularly with ontologies, can be traced back to about two decades ago. Since the seminal paper *Geospatial Semantic Web* (Egenhofer, 2002), the geospatial community and the Semantic Web community have witnessed many studies, approaches, and applications on integrating geospatial data with Semantic Web technologies and incorporating geospatial data in the Semantic Web (Schade and Smits, 2012).

Ontology, as one of the cornerstones of the Semantic Web stack (cf. Figure 1), has been extensively utilised for geospatial data integration, including integration at both instance level and conceptual/schema level. For example, Du et al. (2012) utilised an ontology-based approach to integrate authoritative and VGI road network data at the feature level. The approach first converts input datasets to ontologies, and then merges the ontologies into a new ontology, which is thereafter checked and modified to be consistent. Hong and Kuo (2015) developed a semi-automatic approach to determine the semantic relations of cross-domain concepts based on lightweight ontologies, in which an algorithm of comparing the structural information was proposed, and the measured relations were formally represented by a bridge ontology for further application needs. Fonseca et al. (2002) proposed an architecture for ontology-driven GIS, in which the ontologies were designed and could be browsed by users. Therefore, for each entity type, its roles, attributes, functions and sub-parts can be displayed. Once a user selects a certain ontological description and a geographic region, semantic mediators are initiated to retrieve data instances according to the semantic information. Métral et al. (2010) leveraged ontologies to achieve interoperability between three-dimensional (3D) geospatial city models. The need of this study stemmed from the interoperability issue between different 3D systems and models, thus they employed ontologies to mediate semantic heterogeneity. Zhao et al. (2017) argued that WFS does not provide direct

access to data distributed in multiple servers; thus they developed an RDF query interface to retrieve distributed WFS in a federated manner, and they showcased the approach in a Web-based prototype system.

Many studies of leveraging Semantic Web technologies for geospatial data integration have been conducted in the environment of SDIs. Janowicz et al. (2010) proposed a framework for semantically enabling SDIs, in which both the geospatial data and activities (discovery, registration, processing and visualisation) are semantically annotated; geospatial data are annotated with ontologies to facilitate data disambiguation and integration. Lutz et al. (2009) leveraged ontologies and logical reasoning for overcoming semantic heterogeneity in SDIs to foster better geospatial data exchange and reuse, where they showcased that users could use terms from a classification system they are familiar with and still find features with a different classification system. van den Brink et al. (2017) identified that many vocabularies (ontologies) have been defined within domains, whereas other domains are seldom taken into account; thus they proposed a methodology and tools for non-automatic, community driven ontology matching for data harmonisation to facilitate the data reuse between datasets in the geospatial domain; at the meantime, they also identified that some subtle semantic relations can hardly be represented using ontologies. Wiemann and Bernard (2016) investigated the possibilities for the combination of SDI and Semantic Web developments in terms of spatial data integration. In their prototype implementation, the spatial relations were explored by WFS and then explicitly stored separately in Linked Data, along with relation types, measurements and some other information. Afterward, Wiemann (2017) formalised the geospatial data integration processes on the Web, and argued that the formalisation could be transformed into ontologies.

The above studies mainly utilised ontologies to achieve semantic interoperability to foster better data integration, exchange, and reuse. The data are semantically annotated, nonetheless they are mainly in conventional data models (e.g. in relational databases) and disseminated through traditional SDIs or Web services (e.g. through WFS). There have been also some other studies investigating geospatial data integration in the Linked Data environment (Semantic Web), that is, underlying geospatial data are already Linked Data or transformed to Linked Data for data integration.

Linked Data has the intrinsic potential of interconnecting data on the Web, and this brings opportunities to integrate and link pieces of geospatial data on the Web, which can be otherwise siloed and isolated. Establishing links with data from other sources is one of the most primary aims and practices of delivering geospatial data as Linked Data. For example, GeoNames has been linked to DBpedia[6] (a Linked

[6] http://dbpedia.org

Data distribution of Wikipedia); LinkedGeoData is linked to DBpedia, GeoNames, and United Nations Food and Agriculture Organization geospatial data (Stadler et al., 2012); the Geographic Names Information System Linked Data of the US are linked to GeoNames and DBpedia (Regalia et al., 2018). Due to the nature of geospatial data that they can serve as nexuses to interconnect data from diverse sources, some of the geospatial Linked Data repositories have become central hubs in the LOD cloud, e.g. GeoNames.

As a results, studies have been performed investigating methodologies to establish links for geospatial Linked Data. Yu et al. (2018) argued that generic Linked Data alignment (matching) systems are inadequate in terms of geospatial Linked Data alignment, thereby they presented a holistic approach to aligning geospatial entities (including concepts, properties and instances) with spatial, lexical, structural, and extensional similarity metrics. The metrics were automatically aggregated by approval voting. Vilches-Blázquez et al. (2014) interlinked geospatial Linked Data from Spanish national datasets in three levels – intra-agency dataset interlinking using *owl:sameAs*, inter-agency data linking by adding geospatial locations to thematic data, and data linking with DBpedia. Zhu et al. (2017) presented an approach to interlink Linked Data of geospatial dataset metadata with eight metrics, i.e. theme, category, spatial topology, temporal topology, spatial precision, temporal granularity, type, and format. Zhang et al. (2019) performed geospatial Linked Data interlinking with spatial distance and topological relationships.

Several contributions are delivered in terms of geospatial data integration leveraging Semantic Web technologies in this thesis. Overall, they performed geospatial Linked Data integration on the Web to facilitate data visualisation and analysis at later stages.

Paper I developed an approach to integrating geospatial thematic data and background data on the Web. The purpose is to solve a long-standing visualisation issue in web maps, that is, the thematic web maps are generally created by spatially overlaying thematic information on top of various base maps despite the intrinsic relations between them, which often raises geometric inconsistences. Therefore, we proposed a relative positioning approach in which the thematic data are positioned based on shared geometries and relative coordinates. A Linked Data-based technical framework is used to realise the relative positioning approach, in which ontologies based on the GeoSPARQL vocabulary were designed. The approach can be used as a new way of modelling and integrating geospatial data on the Web, with merits in terms of both data visualisation and querying.

Paper II integrated several geospatial data sources for visualising data in a syncretic way. A Swedish heritage building map case study was demonstrated, in which three datasets were integrated. The integrated data and a knowledge base underpinned geovisualisation of distributed and integrated data.

Paper III identified the problem of complex and subtle semantic relations (differences) for data integration in an interdisciplinary geospatial study, namely visualising urban bicycling suitability. In this study, the intention is to evaluate the suitability of a road network for bicycling. However, the index that is used to evaluate the components of the road network views the geographic space differently than how the road networks are modelled in geospatial data. We identified that such complex semantic relations are raised by the multiple representations of geospatial data, and therefore cross-detailed-level data integration is necessary. Such semantic relations cannot be readily represented merely using ontologies. Therefore, we utilized semantic constraints (SHACL) to develop a framework for the cross-detailed-level data integration.

2.3 Geospatial knowledge representation

The importance of knowledge representation and formalisation have been recognised in the geospatial community for several decades. Schuurman (2006) stressed that geospatial experts and scientists have been at pains to address the same issues so that solutions can be incorporated into software technologies. Moreover, knowledge needs to be represented in a formalised way to facilitate geospatial knowledge sharing and exchange. This particularly matters in view of the inherent multi-disciplinary nature of geospatial information, which entails the need of geospatial knowledge sharing and outreach to various domains where geospatial information is incorporated.

The objective of knowledge representation is to make knowledge explicit and in a computer tractable form, so that it can be used to enable artificial intelligence agents (Bergmann et al., 2005). In fact, formalising data semantics for data integration with Semantic Web technologies, in particular ontology (as illustrated in Section 2.2), also falls under the umbrella of knowledge representation. Nevertheless, in this section we only focus on the knowledge formalisation concerning how geospatial information should be used in a meaningful way (semantics of data usage). This section describes related works in geospatial knowledge representation using Semantic Web technologies on the two aspects of geovisualisation and geoprocessing, which are two salient ways of utilising geospatial information and two actively investigated areas in terms of geospatial knowledge representation and formalisation.

2.3.1 Knowledge representation for geovisualisation

Geovisualisation, as an essential craft that aids users to gain insights from geospatial data, allows users to explore, synthesise, present, and analyse the underlying geospatial data in an interactive manner. Geovisualisation is a knowledge-intensive art, in which both its providers and users are required to possess substantial knowledge of how the geospatial data are visualised, and such knowledge pertains to a wide range of cartographic theories of scaling, portrayal (styles and symbols), etc. This need is important not only for the visualisation providers, who often have to endeavour to derive sensemaking and cartographically satisfactory visualisations, but also the users, who should interpret the presented information in a meaningful way. Geovisualisation is found tricky by non-geospatial experts as the cartographic theories often only lie in cartographic literature, complex program or cartographers' mind. The (digital) cartographic theories have been pursued and studied for decades, while only few are outreached and can be readily utilised by non-geospatial domain experts.

In cartography, it is commonly acknowledged that map making is an inherently human process that is difficult to automate as computers are usually not capable of handling perceptual properties of the data portrayal (Harrie and Weibel, 2007; Scheider and Huisjes, 2019). Nevertheless, cartographic knowledge can be formally represented to enhance computer aiding and propagation of such knowledge.

Semantic Web technologies have great, yet relatively unexplored, potential for formalising the knowledge of visualising geospatial data, and thus facilitate such knowledge to be readily interpreted, shared, expanded, and reused. The idea of a map as a knowledge base of logical representations is intuitive in view of the implicit concepts and rules inherent in the maps (Kavouras and Kokla, 2007). Varanka and Usery (2018) proposed to semantically represent map features using Semantic Web technologies to form the knowledge base of maps (the data were transformed to Linked Data). Grounded upon this idea, the knowledge concerning how the raw features are converted to visualisations can be formalised as another layer of the knowledge base. That is, the knowledge base not only contains geospatial data, but also the knowledge concerning how to visualise them for different applications.

Although it is intuitive to develop knowledge base with formalised geovisualisation knowledge on top of geospatial Linked Data, the studies concerning visualising geospatial Linked Data mainly leveraged hard-coded visualisation settings. The visualisation of linked data, in general, refers to techniques for visually presenting the links between entities to facilitate the intuitive discovery of underlying information and knowledge (Dadzie and Rowe, 2011). For geospatial data, the spatial context is crucial for easing this perception and discovery process. Therefore,

the visualisation of geospatial Linked Data is generally in the form of map mashups, in which the data are spatially presented as thematic data on top of various base maps. Nevertheless, the tools developed for visualising geospatial Linked Data do not employ a knowledge-based approach, e.g. LOD4WFS (Jones et al., 2014), and Map4RDF (Llaves et al., 2014).

There have been some studies formalising geovisualisation with Semantic Web technologies. For example, Scheider and Huisjes (2019) distinguished extensive and intensive properties using machine learning techniques and formalised different types of properties using ontologies to help map making, as the cartographic rules applied to the two types of properties are fundamentally different. Carral et al. (2013) designed an ontology for cartographic map scaling as scale resides in the very core of cartography and essential for geovisualisation; they formalised the cartographic scale information at dataset level for representing the scale knowledge associated with geospatial datasets. Gould and Mackaness (2016) formalised knowledge for on-demand map generalisation using ontologies to facilitate the knowledge to be shared, expanded, and reused in the mapping systems. Iosifescu-Enescu and Hurni (2007), Smith (2010), and Brus et al. (2010) designed cartographic ontologies for map making and "enable computers to learn cartography". Janowicz et al. (2010) advocated to semantically annotate the OGC standard for sharing and exchange cartographic data portrayal information – styled layer descriptor (SLD) (Lupp, 2007) in order to make the portrayal information formal and explicit. They regarded visualisation as a sink where semantics transferred through all the components of SDIs has to be aggregated, interpreted and visualised in a meaningful way. The OGC Testbeds 11, 12, and 13 developed ontologies to accomplish this vision (Fellah, 2015, 2017, and 2018). The OGC Testbeds designed ontologies for portrayal information with the initial purpose of semantic mediation of multi-source portrayal data. The ontologies evolved in an SLD-inclined manner through the Testbeds. They mainly modularised the theories into four micro-theories (style, symbol, symboliser and graphic ontologies) to avoid enormous ontology and to underlay better reusability.

This thesis contributes to geovisualisation knowledge formalisation in Paper I, Paper II, and Paper III.

Paper I formalised the knowledge for both visualisation scales for geospatial features and the relations between thematic data and base maps using ontologies, and structured the data accordingly in Linked Data to enable geometrically self-adapting thematic web maps.

Paper II proposed to formally represent the knowledge of context-aware geovisualisation in three aspects: cartographic scaling, data portrayal and geometry source, which are three prominent facets of geovisualisation knowledge in the contemporary web mapping environment. A Semantic Web technology-based

framework was employed, in which Linked Data was leveraged as underlying data model, and ontologies and semantic rules (SPIN rules) were used for formalising geovisualisation knowledge in a both human- and machine-readable manner.

Paper III enriched the geovisualisation knowledge base designed in Paper II with some high level cartographic knowledge to further automate the process of representing geovisualisation knowledge. A bicycling suitability map was developed fully based on such a knowledge base. Different than paper II, this knowledge base is capable of deducing colours for each type of geographic features depending on colour scales that are designed based on the high level cartographic theories.

2.3.2 Knowledge representation for geoprocessing

Geoprocessing is a core application of GIS and pertains to analysing geospatial data for knowledge discovery. Usually, a single geoprocessing operations can be hardly sufficient for some complex data analysis tasks, thereby composing geoprocessing workflows to chain several operations is often necessary. Such compositions of geoprocessing workflows require considerable knowledge from the users, despite the support provided from GIS tools which requires information about the data and operations from the system side (Hofer et al., 2016). Increasing tools and Web services have been available on the Web, and thus it is important to represent the knowledge concerning how to compose geoprocessing workflows to automate such processes. To realise this goal, one core requirement is the semantic interoperability between geoprocessing operations or services (Yue et al., 2015). In this context, many attempts have been made to leverage Semantic Web technologies into composition and formally representing of geoprocessing workflows.

A few examples of studies employing Semantic Web technologies for geoprocessing are listed here. Hofer et al. (2016) developed a knowledge base to support the composition of geoprocessing workflow, in which ontologies were used to formalise the geooperators, and the SWRL was used to formulate the rules associated with geooperator chaining. Scheider and Ballatore (2018) proposed a method for semantically typing geoprocessing workflows using Linked Data. With the Linked Data paradigm, workflow resources are described in RDF and readily sharable on the Web. They utilised the core concepts of spatial information from Kuhn (2012) and principles of typed functional programming. Ontologies and SPARQL rules were leveraged to enrich the workflows with semantic types. Scheider et al. (2019) formalised both geoprocessing tools and the requirements from the users using ontologies and SPARQL CONSTRUCT queries, and they proposed an algorithm for computing query containment to match the formalised GIS tools with the questions that they can answer. Falquet et al. (2018) used

ontologies and SPARQL CONSTRUCT queries to provide an abstract description for the process of geospatial Linked Data publication.

Diverged from the abovementioned works, which mainly concentrate on the meta-level of geoprocessing operations for the purpose of workflow composition, Paper III in this thesis encapsulated a geoprocessing process in a knowledge base that comprises ontologies and semantic rules. The purpose here is to formalise the knowledge concerning how geospatial data should be processed to foster cross-disciplinary knowledge transfer and reuse.

3 Summary of papers

Chapter 3 summarises the papers that are the basis for the thesis.

3.1 Paper I: Synchronising geometric representations for map mashups using relative positioning and Linked Data

The aim of this study was to develop an approach to relatively position thematic data based on their relations with background geospatial data (base maps) on the Web for the purpose of synchronising geometric representations of web maps in a multi-scale environment (first objective in Section 1.2).

The starting point of this study was that map mashups (a type of web maps), as a common way of presenting spatial information and the most popular mashups on the Web, are generally created by spatially overlaying thematic information on top of various base maps despite the intrinsic connections that thematic data usually have with base maps. Such simple overlay approach often raises geometric inconsistencies between thematic data and base maps.

In this context, we developed a relative positioning method based on the Linked Data paradigm. In this approach, geometries of the thematic features are not represented with absolute coordinates; instead, the geometries are assembled based on shared geometries with background features and relative coordinates for the parts that have no correspondence relation with background features. For example, a feature representing a natural reserved area can be assembled by shared geometries with the background features of rivers, lakes, and cadastral units; the parts that do not have any correspondence relations with background features are represented with relative coordinates.

We designed ontologies with a set of competency questions for formalising the relations between thematic data and base maps. The two ontologies – thematic data ontology and base map ontology, are based on the GeoSPARQL vocabulary.

We realised our approach in a case of mapping several natural reserved areas in Northern Sweden. A multi-scale base map was used, where a multiple representation database (MRDB) was constructed and released as Linked Data. Thematic data were relatively positioned in a developed tool in this study. The tool enabled users to formulate relations between thematic and background data in a graphic interface. A backend server was developed to enable real-time generation and assembly of the relatively positioned thematic data, which were then transferred to frontend for visualisation. The results demonstrated that the proposed approach and architecture can effectively resolve geometric inconsistencies between thematic and background data in web maps.

The proposed approach can be used as a new way of modelling geospatial data on the Web. The benefits brought up by this approach includes not only in the respect of data visualisation as illustrated above; the approach also has unlocked potential for information retrieval and question answering. For example, answers to the following questions can be retrieved with low computational cost:

- Which feature type is most involved in the definition of the natural protected areas?
- Which national protected areas coincide completely with a single cadastral unit?

3.2 Paper II: Towards knowledge-based geovisualisation using Semantic Web technologies: A knowledge representation approach coupling ontologies and rules

The aim of this paper was to develop a formalised knowledge base for visualising geospatial data, i.e. how geospatial data are converted to visualisations. Such a knowledge base can be readily shared on the Web, and thus facilitate transfer, interpretation, and reuse of the geovisualisation knowledge (second objective in Section 1.2).

The starting point of this study was that geovisualisation is knowledge-intensive for both its providers and users. And the providers and users often should research a high level of cognitive consensus for better comprehension of the visualisations. Current approaches to representing the visualisation knowledge, i.e. how geospatial data are converted to visualisations, lack semantics on the one hand, and are too bespoke on the other hand. For example, OGC developed SLD, which is a syntactic standardisation of storing and sharing visualisation information, yet it lacks semantics and relies on ad-hoc parsers. These natures of such approaches impede

augmenting the transfer of visualisation knowledge, and its reuse. Knowledge transfer particularly matters, as geovisualisation is utilised in various domains.

In this paper, we designed a knowledge base for geovisualisation by encapsulating knowledge on the three aspects of cartographic scale, data portrayal, and geometry source, which, we believe, comprise the core of geovisualisation in the contemporary web mapping era from a visual representation perspective.

For cartographic scale, we represented and formalised the knowledge concerning in which scale range should the geospatial data be visualised in an ontology. The knowledge was modelled at dataset level, together with other types of metadata, i.e. data context. Geospatial data were organised and released as Linked Data in named graphs with multiple representations. We also formalised the knowledge of client context to enable context-aware web mapping.

For data portrayal, we revisited the ontologies designed from the OGC Testbeds (cf. Section 2.3.1). The data portrayal knowledge was represented using ontologies and semantic rules. Modularised ontologies for style, symbol, symboliser, graphic, and legend were designed. A rule base containing SPIN (SPARQL) rules was constructed to enable machines to derive corresponding symbolisers (the means of visualising features) for geospatial features under different conditions. Several rules were also formulated to enable context-aware visualisation, i.e. rendering geospatial data differently according to different visualisation contexts.

For geometry source, we also designed ontologies and semantic rules for rendering geospatial features with different types of geometries from several sources for varying visualisation purposes and scales. The rationale of formalising such knowledge was that, in the Linked Data environment, geospatial data are increasingly interlinked with each other and data from other domains. Hence, we can adopt an integrated visualisation strategy, that is, a strategy in which the visualisation of geospatial linked dataset relies on both the geometries modelled in its own dataset and geometries from other datasets.

The knowledge representation approach was realised in a case of visualising multi-scale heritage building maps of Stockholm, Sweden using multi-source data. The underlying geospatial data, geovisualisation knowledge base, and geovisualisation application are all distributed.

The experiment demonstrated several advantages of our approach compared to state-of-the-art approaches (e.g. SLD). First, the knowledge-based approach with Semantic Web technologies enables retrieving and visualising distributed multi-source data in a federated way. Second, it is less bespoke as it relies on several W3C standards. Third, our approach provides enriched semantics, which could lift semantic harmonisation from the data level to the visual level, and facilitate context-aware visualisation. The last but not the least, the enriched semantics is presented

to users through a semantically-enriched legend. The proposed approach can partially form the foundation for the vision of a Web of knowledge for geovisualisation.

3.3 Paper III: Towards knowledge-based geospatial data integration and visualization: A case of visualizing urban bicycling suitability

In this paper, a knowledge-based framework for geospatial data integration and visualization was proposed. The proposed framework was showcased in a spatially informed interdisciplinary study – visualizing urban bicycling suitability.

In this paper, we showcased a study, in which merely using ontologies for representing semantic relations between heterogeneous data sources was inadequate. The case study aimed at evaluating urban bicycling suitability by an index in this regard – level of traffic stress (LTS). This index implies that both links (segments) and nodes (junctions) are quantitatively evaluated to derive a comprehensive understanding of the network's suitability and connectivity. This index needs a comprehensive and detailed set of links, which corresponds to the links in a more detailed dataset; whereas this index views the junctions in the same way as they are modelled in the less detailed dataset (if present). That is, in the conceptualization of the road network from the traffic domain, road junctions correspond to the data modelling approach in the coarse level of geospatial data, while links correspond to the more detailed level. Therefore, this becomes a cross-detailed-level data integration task. Such complex and subtle semantic relations raised by multiple representations of geospatial data cannot be captured merely using ontologies. Therefore, we complemented ontologies with semantic constraints (SHACL) to formally represent such knowledge in the process of data integration, and to ensure that the data collected by traffic researchers are integrated with the geospatial data in a correct detailed level.

Furthermore, we enriched the knowledge base for geovisualisation developed in Paper II with high level cartographic knowledge to further automate the process of representing geovisualisation knowledge. A cartographic ontology was introduced into the knowledge base, and the high level cartographic knowledge such as measurement scale (e.g. ordinal data), and colour scale was modelled in the cartographic ontology. The high level cartographic knowledge was then transferred to lower level geovisualisation knowledge, i.e. style and symbol level. In addition, the derivation of LTS was also formally represented and encapsulated in the knowledge base.

The knowledge for geospatial data analysis and visualisation was modelled in three knowledge representation abstraction levels, and different types of data usage knowledge were modelled in different levels. The three levels are: (1) cartographic common knowledge; (2) the level of a type of indexes; and (3) the level of the particular index that was used in the study. Knowledge modelled in higher level can be automatically transferred to lower levels, so as to simplify the process of knowledge representation.

With this framework, a bicycling suitability map was developed fully based on integrated Linked Data and the knowledge base for data analysis and visualisation. The case study illustrated that the knowledge-based approach successfully overcame semantic heterogeneity for cross-domain data integration with subtle and complex semantic relations. In addition, the knowledge modelled for data analysis as well as visualization effectively empowered machines to derive desired outcomes. This work provides a methodological framework for the sharing and outreach of geospatial data and knowledge to a wider audience for interdisciplinary spatially informed studies.

3.4 Paper IV: Assessment and benchmarking of spatially enabled RDF stores for the next generation of spatial data infrastructure

The aim of this study was to evaluate and benchmark several well-known and mainstream RDF stores in terms of their spatial query capacities. Focuses of the evaluation included the GeoSPARQL compliance, query performance, and query correctness (fourth object in Section 1.2).

The starting point of this study was that current SDI solutions are facing a number of limitations, especially in terms of discovery, reuse, and integration of the data. Therefore, it is likely that Linked Data will be in the future path of SDI development, especially in view of the recent investigations, which revealed that Linked Data has been seen as one of the key factors moving SDIs to the next generation (cf. Section 1.1). In this context, one important question is how is the technical environment for delivering geospatial data using Linked Data, particularly in terms of the solutions for storage, querying, and analysis. Therefore, it is relevant to benchmark spatially enabled RDF stores (RDF stores supporting spatial queries).

Five mainstream and well-known RDF stores were chosen in this study – RDF4J, GeoSPARQL-Jena, Virtuoso, Stardog, and GraphDB. The stores were tested and benchmarks in two scenarios, that is, with two benchmarking datasets. The first scenario was benchmarking with ICOS Carbon Portal's metadata – a real-world

Earth Science Linked Data infrastructure, which has a mixture of geospatial and non-geospatial data. The rationale of this scenario was to investigate the performance of the spatially enabled RDF stores when operating and querying geospatial data out of a large amount of data containing both relevant and irrelevant data. The second scenario used benchmarking datasets from a previous benchmark Geographica with a large number of geospatial objects. The objective of having this scenario was that it is closer to real-world dedicated SDI, in which operating upon enormous geospatial objects is common. The benchmark queries were adopted or tailored from the Geographica benchmark.

The assessment and benchmarking results demonstrated that the GeoSPARQL compliance of the RDF stores has encouragingly advanced in the last several years. The query performances were generally acceptable, and spatial indexing was imperative when handling a large number of geospatial objects. Nevertheless, query correctness remained a challenge for cross-database interoperability. Precision setting was probably a major cause of inconsistent query results. There is a trade-off concerning whether we should perform spatial queries within RDF stores or leaving such operations to conventional geospatial tools (e.g. PostGIS), e.g. through pre-computation or federated queries.

In conclusion, the results indicated that the spatial capacity of the RDF stores has become increasingly mature, which could benefit the development of future SDIs.

4 Conclusions and outlook

4.1 Conclusions

Geospatial information has a salient inter-disciplinary nature, and thus it is important to share and outreach geospatial data and knowledge in a way that can be readily accessed, interpreted, and reused. Current solutions for delivering and disseminating geospatial data are facing apparent limitations in terms of data integration as well as semantic interoperability. In addition, geospatial knowledge needs to be formalised to foster better interpretation and reuse. In this context, Semantic Web technologies have been seen as a potential remedy for these limitations.

This thesis aims to investigate the potentials of Semantic Web technologies for geospatial data integration and knowledge formalisation of data visualisation, which is a predominating way of utilising geospatial data. Several research objectives are formulated according to the aim, which are respectively (partially) accomplished in this thesis:

Research objective 1: to develop methods for relatively positioning geospatial features based on their relations with background data using Linked Data

This objective is accomplished in Paper I, in which a relative positioning approach based on shared geometries and relative coordinates was developed. The approach was realised in a Linked Data framework so that geospatial data can be relatively positioning to background data with multiple representations on the Web. A use case of this approach in a web map was designed, implemented, and evaluated. The use of relative positioning in web maps indicated that the relatively positioned geospatial features are naturally integrated and synchronised with the multiple representation background data, i.e. the thematic data automatically obtain synchronised multiple scale representations, which is a prerequisite for proper visualisation. Therefore, the relatively positioned geospatial features avoided substantial visual deficiencies.

Research objective 2: to exploit the knowledge representation capacity of Semantic Web technologies for formalising the knowledge of geospatial data visualisation

This objective was investigated in Paper II and III. In Paper II, a knowledge base for geovisualisation was developed, which mainly covered the knowledge on the aspects of cartographic scale, data portrayal, and geometry source. These three aspects are important in the contemporary Web mapping era from a visual presentation perspective. Paper III enriched the knowledge base for geovisualisation with high level cartographic knowledge, and created different abstraction layers for geovisualisation knowledge representation, so as to further ease and automate the development of such knowledge bases. The formalised knowledge for geovisualisation was used in two case studies: heritage building mapping, and visualising urban bicycling suitability. In the two studies, several advantages of the knowledge-based approach for geovisualisation were unveiled. The knowledge-based methods are semantically enriched compared to syntactic standards such as SLD, and thus can facilitate the clarification of the meaning and selection of the styles and symbols. The enriched semantics can foster better interpretation, and reuse of the knowledge. Furthermore, such knowledge bases can be used as a visualisation enablement layer for geospatial data in the LOD cloud. In summary, we believe the knowledge-based approaches can facilitate the outreach and transfer of geovisualisation knowledge, in order to enable such knowledge to be readily utilized in different applications and domains.

Research objective 3: to develop methods dealing with the cross-detailed-level data integration problem with Semantic Web technologies

This objective is investigated in Paper III, in which semantic constraints (SHACL) were used to formally represent complex and subtle semantic relations in a data integration task between geospatial multi-scale road networks and field collected traffic data. A SHACL-based framework was developed to ensure that the field collected data were integrated with geospatial road network data in the appropriate level of detail. In fact, such semantic relations and constraints can hardly be represented only with ontologies, e.g. using OWL. The effectiveness of the proposed approach was evaluated and verified in the case study of evaluating urban road network's bicycling suitability. We believe this approach can be used to formally represent various complex semantic relations for geospatial data integration.

Research objective 4: to assess and benchmark widely-used and well-known Linked Data stores to understand where the methods can be better applied in, and bring insights to the (geospatial) Linked Data community at large

This objective is (partially) fulfilled in Paper IV, in which an assessment and benchmarking of five well-known and mainstream spatially enabled RDF stores were conducted in terms of their spatial query capacities, i.e. RDF4J, GeoSPARQL-Jena, Virtuoso, Stardog, and GraphDB. We assessed and benchmarked the stores in two scenarios. One scenario involves benchmarking the RDF stores with Integrated

Carbon Observation System Carbon Portal (ICOS CP) metadata, a large-scale Earth Science data infrastructure in which geospatial data are integrated with other types of data. The other scenario was in a dedicated SDI environment with a large amount of purely geospatial data, which is a mixture of crowd-sourced and authoritative geospatial data. The queries used in this study were mainly from the Geographica benchmark (Garbis et al., 2013). The results demonstrated that GeoSPARQL compliance had advanced dramatically in the last several years for the RDF stores compared to previous benchmarking results (see. Section 2.1.3), and query performances were generally acceptable. Furthermore, spatial indexing was important when querying a large number of geospatial objects. However, query correctness remained a challenge for cross-database interoperability.

In summary, this thesis has addressed several important issues spanning data modelling, data processing, data integration, data visualisation, knowledge formalisation, Linked Data for SDI development, and database benchmarking. This thesis lies in the conjunction of two research domains – Geographic Information Science (GIScience) and Semantic Web, and thus it contributes and provides insights to both of them. From a GIScience viewpoint, this thesis investigates and unveils several benefits of Semantic Web technologies for geospatial data integration and visualisation, which are long-standing research issues in GIScience. From the perspective of the Semantic Web, this thesis expands the border of its applications, and delivers insights concerning how to properly integrate and utilise geospatial data and knowledge on the Semantic Web.

4.2 Outlook

This thesis generates potentials for further studies in the areas of geospatial data integration, visualisation, and geospatial Semantic Web.

For geospatial data integration, further studies are desired to design complete ontologies representing correspondence relations between geospatial data. Currently, geospatial objects are linked via the relations such as *owl:sameAs*, *skos:closeMatch*, whereas none of them could comprehensively and precisely capture the various types of geospatial feature relations in real-world scenarios. For example, one object in one dataset could be semantically equivalent to the aggregation of several objects in another dataset. In addition, semantic rules could be formulated according to different types of spatial relations. Such ontologies would matter, e.g. for federated spatial querying, geospatial data update, and multi-scale spatial analysis.

Another key for geospatial data integration, especially in the environment of Linked Data, is entity alignment, namely discovering correspondence relations between different (linked) datasets. In this thesis, this was accompanied mainly in semi-automated rule-based approaches (Paper I, II, and III). However, such rule-based approaches usually are inefficient, not scalable, impractical, and inferior, due to the limited and often unrealistic perspectives informing the rules. In this end, it is worth to explore machine learning methods, in particular the machine learning methods for Linked Data (knowledge graph) – knowledge representation learning (e.g. knowledge graph embedding techniques), to improve geospatial entity alignment.

The Linked Data based relative positioning approach developed in this thesis (Paper I) has yet unlocked potential in spatial data retrieval and question answering. This is because the spatial relations are explicitly represented. Further studies concerning benefits and potential problems of this approach are desired. And more automated approach of constructing relatively positioned geospatial data is anticipated (than the tool for data creation that we developed in Paper I).

For geospatial data visualisation, further formalising its knowledge with Semantic Web technologies is worth exploiting. We believe such work should be grounded upon the development of cartographic ontologies. Nevertheless, how much knowledge can be formalised, and is it at all possible to formalise cartographic knowledge in an analytical way is questionable. One possible direction is combining the knowledge that can and cannot be formalised using Semantic Web technologies.

From the viewpoint of SDI development, there is certainly a long way ahead for deploying Linked Data solutions at large, especially in view of the scepticism toward it in the geospatial domain. Tools that could facilitate geospatial Linked Data generation and publication are needed. The trade-off concerning whether we should perform spatial queries within RDF stores or leaving such operations to conventional geospatial tools also deserves exploring.

Overall, as Semantic Web technologies have been adopted in many research domains and evolved into the mainstream of the Web, we expect that they can act as a reinforced bridge between the geospatial domain and other domains where geospatial information plays a role.

References

ABDOLMAJIDI, E., MANSOURIAN, A., WILL, J. & HARRIE, L. 2015. Matching authority and VGI road networks using an extended node-based matching algorithm. *Geo-spatial Information Science,* 18**,** 65-80.

ATHANASIOU, S., BEZATI, L., GIANNOPOULOS, G., PATROUMPAS, K. & SKOUTAS, D. 2013. GeoKnow – Making the Web an Exploratory for Geospatial Knowledge: Deliverable 2.1.1 Market and Research Overview.

BAADER, F., CALVANESE, D., MCGUINNESS, D., PATEL-SCHNEIDER, P. & NARDI, D. 2003. *The description logic handbook: Theory, implementation and applications*, Cambridge university press.

BATTLE, R. & KOLAS, D. 2012. Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web,* 3**,** 355-370.

BELLINI, P. & NESI, P. 2018. Performance assessment of RDF graph databases for smart city services. *Journal of Visual Languages & Computing,* 45**,** 24-38.

BERGMANN, R., KOLODNER, J. & PLAZA, E. 2005. Representation in case-based reasoning. *The Knowledge Engineering Review,* 20**,** 209-213.

BERMUDEZ, L. 2012. *Is the OGC playing with Linked Data?* [Online]. Available: http://www.opengeospatial.org/blog/1673 [Accessed August 19 2019].

BERNERS-LEE, T. 2009. *Linked Data - Design Issues* [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html [Accessed June 1 2016].

BERNERS-LEE, T., HENDLER, J. & LASSILA, O. 2001. The semantic web. *Scientific american,* 284**,** 28-37.

BRUS, J., ZDENA, D., KANOK, J. & PECHANEC, V. Design of intelligent system in cartography. Roedunet International Conference (RoEduNet), 2010 9th, 2010. IEEE, 112-117.

CARRAL, D., SCHEIDER, S., JANOWICZ, K., VARDEMAN, C., KRISNADHI, A. A. & HITZLER, P. An ontology design pattern for cartographic map scaling. Extended Semantic Web Conference, 2013. Springer, 76-93.

CHEATHAM, M., KRISNADHI, A., AMINI, R., HITZLER, P., JANOWICZ, K., SHEPHERD, A., NAROCK, T., JONES, M. & JI, P. 2018. The GeoLink knowledge graph. *Big Earth Data,* 2**,** 131-143.

DADZIE, A.-S. & ROWE, M. 2011. Approaches to visualising linked data: A survey. *Semantic Web,* 2**,** 89-124.

DU, H., ANAND, S., ALECHINA, N., MORLEY, J., HART, G., LEIBOVICI, D., JACKSON, M. & WARE, M. 2012. Geospatial information integration for

authoritative and crowd sourced road vector data. *Transactions in GIS,* 16**,** 455-476.
EGENHOFER, M. J. Toward the semantic geospatial web. Proceedings of the 10th ACM international symposium on Advances in geographic information systems, 2002. ACM, 1-4.
EUROSDR 2019. EuroSDR Annual Report 2018 [Online]. Available: http://www.eurosdr.net/sites/default/files/images/inline/eurosdr_annual_report_2018.pdf [Accessed October 12 2019].
FALQUET, G., METRAL, C., OZAINNE, S. & GIULIANI, G. An Abstract Specification Technique for the Publication of Linked Geospatial Data. 21th AGILE conference on Geographic Information Science, June 12–15 2018 Lund, Sweden.
FELLAH, S. 2015. OGC Testbed-11 Symbology Mediation Engineering [Online]. Available:https://portal.opengeospatial.org/files/?artifact_id=64385 [Accessed July 20 2019].
FELLAH, S. 2017. Testbed-12 Semantic Portrayal, Registry and Mediation Engineering Report [Online]. Available: http://docs.opengeospatial.org/per/16-059.html [Accessed July 20 2019].
FELLAH, S. 2018. Testbed-13: Portrayal Engineering Report [Online]. Available: http://docs.opengeospatial.org/per/17-045.html [Accessed July 20 2019].
FOLMER, E., BEEK, W. & RIETVELD, L. 2018. Linked Data Viewing as part of the Spatial Data Platform of the Future. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* 42**,** 49-52.
FONSECA, F. T., EGENHOFER, M. J., AGOURIS, P. & CÂMARA, G. 2002. Using ontologies for integrated geographic information systems. *Transactions in GIS,* 6**,** 231-257.
GARBIS, G., KYZIRAKOS, K. & KOUBARAKIS, M. Geographica: A benchmark for geospatial rdf stores (long version). International Semantic Web Conference, 2013. Springer, 343-359.
GOODWIN, J., DOLBEAR, C. & HART, G. 2008. Geographical linked data: The administrative geography of great britain on the semantic web. *Transactions in GIS,* 12**,** 19-30.
GOULD, N. & MACKANESS, W. 2016. From taxonomies to ontologies: formalizing generalization knowledge for on-demand mapping. *Cartography and Geographic Information Science,* 43**,** 208-222.
HARRIE, L. & WEIBEL, R. 2007. Modelling the overall process of generalisation. *Generalisation of geographic information: cartographic modelling and applications***,** 67-87.
HIETANEN, E., LEHTO, L. & LATVALA, P. 2016. Providing Geographic Datasets as Linked Data in Sdi. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences***,** 583-586.
HITZLER, P. & PARSIA, B. 2009. Ontologies and rules. *Handbook on Ontologies.* Berlin, Heidelberg: Springer.

HOFER, B., MÄS, S., BRAUNER, J. & BERNARD, L. 2016. Towards a knowledge base to support geoprocessing workflow development. *International Journal of Geographical Information Science,* 31**,** 694-716.
HONG, J.-H. & KUO, C.-L. 2015. A semi-automatic lightweight ontology bridging for the semantic integration of cross-domain geospatial information. *International Journal of Geographical Information Science,* 29**,** 2223-2247.
HORROCKS, I. 2008. Ontologies and the semantic web. *Communications of the ACM,* 51**,** 58-67.
HORROCKS, I., PATEL-SCHNEIDER, P. F., BOLEY, H., TABET, S., GROSOF, B. & DEAN, M. 2004. SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member submission.*
HUANG, W. & HARRIE, L. 2019. Towards knowledge-based geovisualisation using Semantic Web technologies: a knowledge representation approach coupling ontologies and rules. *International Journal of Digital Earth***,** 1-22.
HUANG, W., RAZA, S. A., MIRZOV, O. & HARRIE, L. 2019. Assessment and Benchmarking of Spatially Enabled RDF Stores for the Next Generation of Spatial Data Infrastructure. *ISPRS International Journal of Geo-Information,* 8**,** 310.
INSPIRE 2013. Guidelines for the encoding of spatial data [Online]. Available: https://inspire.ec.europa.eu/file/1412/download?token=lHYwKEk5 [Accessed August 18 2019]
INSPIRE. 2017. *Linking INSPIRE data: draft guidelines and pilots* [Online]. Available: https://inspire.ec.europa.eu/news/linking-inspire-data-draft-guidelines-and-pilots [Accessed April 8 2018].
INSPIRE. 2018. Available: https://inspire.ec.europa.eu/ [Accessed 2 December 2018].
IOSIFESCU-ENESCU, I. & HURNI, L. Towards cartographic ontologies or how computers learn cartography. Proceedings 23rd International Cartographic Conference, 2007. 4-10.
JANOWICZ, K., SCHADE, S., BRÖRING, A., KEßLER, C., MAUÉ, P. & STASCH, C. 2010. Semantic enablement for spatial data infrastructures. *Transactions in GIS,* 14**,** 111-129.
JANOWICZ, K., SCHEIDER, S., PEHLE, T. & HART, G. 2012. Geospatial semantics and linked spatiotemporal data–Past, present, and future. *Semantic Web,* 3**,** 321-332.
JONES, J., KUHN, W., KEßLER, C. & SCHEIDER, S. 2014. Making the web of data available via web feature services. *Connecting a Digital Europe Through Location and Place.* Springer.
KAVOURAS, M. & KOKLA, M. 2007. *Theories of geographic concepts: ontological approaches to semantic integration*, CRC Press.
KEßLER, C. 2010. *Context-aware semantics-based information retrieval*, IOS Press.
KIM, I.-H., FENG, C.-C. & WANG, Y.-C. 2017. A simplified linear feature matching method using decision tree analysis, weighted linear directional

mean, and topological relationships. *International Journal of Geographical Information Science,* 31**,** 1042-1060.
KNUBLAUCH, H. 2011. SPIN-modeling vocabulary. *W3C Member Submission.*
KNUBLAUCH, H. 2017. SHACL and OWL compared. Technical report, W3C.
KNUBLAUCH, H., ALLEMANG, D. & STEYSKAL, S. 2017. SHACL Advanced Features. W3C Working Group Note.
KNUBLAUCH, H., HENDLER, J. A. & IDEHEN, K. 2011. SPIN-Overview and Motivation. *W3C Member Submission.*
KNUBLAUCH, H. & KONTOKOSTAS, D. 2017. Shapes constraint language (SHACL). *W3C Recommendation*.
KOUKOLETSOS, T., HAKLAY, M. & ELLUL, C. 2012. Assessing data completeness of VGI through an automated matching procedure for linear data. *Transactions in GIS,* 16**,** 477-498.
KUHN, W. 2012. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science,* 26**,** 2267-2276.
LINKED OPEN DATA CLOUD. 2019. Available: https://lod-cloud.net/ [Accessed December 2 2019].
LLAVES, A., CORCHO, O. & FERNANDEZ-CARRERA, A. Map4rdf-ios: a tool for exploring linked geospatial data. Proceedings of Workshop on Linked Geospatial Data, 2014.
LUDWIG, I., VOSS, A. & KRAUSE-TRAUDES, M. 2011. A Comparison of the Street Networks of Navteq and OSM in Germany. *Advancing Geoinformation Science for a Changing World.* Springer.
LUPP, M. 2007. OGC Implementation Specification 05-078r4: Styled Layer Descriptor profile of the Web Map Service Implementation Specification. *Open Geospatial Consortium, Wayland, USA*.
LUTZ, M., SPRADO, J., KLIEN, E., SCHUBERT, C. & CHRIST, I. 2009. Overcoming semantic heterogeneity in spatial data infrastructures. *Computers & Geosciences,* 35**,** 739-752.
MACEACHREN, A. M. 2004. *How maps work: representation, visualization, and design,* New York, Guilford Press.
MÉTRAL, C., BILLEN, R., CUTTING-DECELLE, A.-F. & VAN RUYMBEKE, M. 2010. Ontology-based approaches for improving the interoperability between 3D urban models. *Journal of Information Technology in Construction,* 15**,** 169-184.
MUSTIÈRE, S. & DEVOGELE, T. 2008. Matching networks with different levels of detail. *GeoInformatica,* 12**,** 435-453.
REGALIA, B., JANOWICZ, K., MAI, G., VARANKA, D. & USERY, E. L. GNIS-LD: Serving and Visualizing the Geographic Names Information System Gazetteer As Linked Data. European Semantic Web Conference, June 3–7 2018 Heraklion, Crete, Greece. Springer, 528-540.
RONZHIN, S., FOLMER, E., MELLUM, R., VON BRASCH, T. E., MARTIN, E., ROMERO, E. L., KYTÖ, S., HIETANEN, E. & LATVALA, P. 2019. Next

Generation of Spatial Data Infrastructure: Lessons from Linked Data implementations across Europe.
SAMAL, A., SETH, S. & CUETO, K. 2004. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science,* 18**,** 459-489.
SCHADE, S. & SMITS, P. Why linked data should not lead to next generation SDI. Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International, 2012. IEEE, 2894-2897.
SCHEIDER, S. & BALLATORE, A. 2018. Semantic typing of linked geoprocessing workflows. *International Journal of Digital Earth,* 11**,** 113-138.
SCHEIDER, S., BALLATORE, A. & LEMMENS, R. 2019. Finding and sharing GIS methods based on the questions they answer. *International journal of digital earth,* 12**,** 594-613.
SCHEIDER, S. & HUISJES, M. D. 2019. Distinguishing extensive and intensive properties for meaningful geocomputation and mapping. *International Journal of Geographical Information Science,* 33**,** 28-54.
SCHUURMAN, N. 2006. Formalization matters: Critical GIS and ontology research. *Annals of the Association of American Geographers,* 96**,** 726-739.
SHI, L., SUKHOBOK, D., NIKOLOV, N. & ROMAN, D. Norwegian State of Estate Report as Linked Open Data. OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", October 23-27 2017 Rhodes, Greece. Springer, 445-462.
SMITH, R. A. Designing a cartographic ontology for use with expert systems. A special joint symposium of ISPRS Technical Commission IV & AutoCarto in conjuction with ASPRS/CaGIS, 2010.
SEMANTIC WEB STACK. 2019. Available: https://en.wikipedia.org/wiki/Semantic_Web_Stack [Accessed July 16 2019].
STADLER, C., LEHMANN, J., HÖFFNER, K. & AUER, S. 2012. Linkedgeodata: A core for a web of spatial open data. *Semantic Web,* 3**,** 333-354.
USERY, E. L. & VARANKA, D. 2012. Design and development of linked data from the national map. *Semantic Web,* 3**,** 371-384.
VAN DEN BRINK, L., BARNAGHI, P., TANDY, J., ATEMEZING, G., ATKINSON, R., COCHRANE, B., FATHY, Y., CASTRO, R. G., HALLER, A. & HARTH, A. 2019. Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web. *Semantic Web,* 10**,** 95-114.
VAN DEN BRINK, L., JANSSEN, P., QUAK, W. & STOTER, J. 2017. Towards a high level of semantic harmonisation in the geospatial domain. *Computers, Environment and Urban Systems,* 62**,** 233-242.
VARANKA, D. E. & USERY, E. L. 2018. The map as knowledge base. *International Journal of Cartography,* 4**,** 201-223.
VILCHES-BLÁZQUEZ, L. M., VILLAZÓN-TERRAZAS, B., CORCHO, O. & GÓMEZ-PÉREZ, A. 2014. Integrating geographical information in the Linked Digital Earth. *International Journal of Digital Earth,* 7**,** 554-575.

VOLZ, S. Data-driven matching of geospatial schemas. International Conference on Spatial Information Theory, 2005. Springer, 115-132.
W3C. 2004. *OWL Web Ontology Language Overview* [Online]. Available: https://www.w3.org/TR/owl-features/ [Accessed February 21 2017].
W3C. 2008. *SPARQL Query Language for RDF* [Online]. Available: https://www.w3.org/TR/rdf-sparql-query/ [Accessed March 21 2016].
W3C. 2012. *Web Ontology Language (OWL)* [Online]. Available: https://www.w3.org/OWL/ [Accessed January 26 2016].
W3C. 2013. *W3C SEMANTIC WEB ACTIVITY* [Online]. Available: https://www.w3.org/2001/sw/ [Accessed].
W3C. 2014. *RDF Schema 1.1* [Online]. Available: https://www.w3.org/TR/rdf-schema/ [Accessed July 8 2016].
WIEMANN, S. 2017. Formalization and web-based implementation of spatial data fusion. *Computers & geosciences,* 99**,** 107-115.
WIEMANN, S. & BERNARD, L. 2016. Spatial data fusion in spatial data infrastructures using linked data. *International Journal of Geographical Information Science,* 30**,** 613-636.
XAVIER, E., ARIZA-LÓPEZ, F. J. & UREÑA-CÁMARA, M. A. 2016. A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys (CSUR),* 49**,** 39.
YANG, B., ZHANG, Y. & LU, F. 2014. Geometric-based approach for integrating VGI POIs and road networks. *International Journal of Geographical Information Science,* 28**,** 126-147.
YANG, C. & GIDOFALVI, G. 2018. Fast map matching, an algorithm integrating hidden Markov model with precomputation. *International Journal of Geographical Information Science,* 32**,** 547-570.
YU, L., QIU, P., LIU, X., LU, F. & WAN, B. 2018. A holistic approach to aligning geospatial data with multidimensional similarity measuring. *International journal of digital earth,* 11**,** 845-862.
YUE, P., BAUMANN, P., BUGBEE, K. & JIANG, L. 2015. Towards intelligent giservices. *Earth Science Informatics,* 8**,** 463-481.
ZHANG, M. & MENG, L. 2007. An iterative road-matching approach for the integration of postal data. *Computers, Environment and Urban Systems,* 31**,** 597-615.
ZHANG, Y., LI, C., CHEN, N., LIU, S., DU, L., WANG, Z. & MA, M. 2019. Semantic web and geospatial unique features based geospatial data integration. *Geospatial Intelligence: Concepts, Methodologies, Tools, and Applications*. IGI Global.
ZHAO, T., ZHANG, C. & LI, W. 2017. Adaptive and Optimized RDF Query Interface for Distributed WFS Data. *ISPRS International Journal of Geo-Information,* 6**,** 108.
ZHU, Y., ZHU, A.-X., SONG, J., YANG, J., FENG, M., SUN, K., ZHANG, J., HOU, Z. & ZHAO, H. 2017. Multidimensional and quantitative interlinking approach for Linked Geospatial Data. *International Journal of Digital Earth,* 10**,** 923-943.

Paper I

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2018
VOL. 32, NO. 6, 1117–1137
https://doi.org/10.1080/13658816.2018.1441416

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Synchronising geometric representations for map mashups using relative positioning and Linked Data

Weiming Huang, Ali Mansourian, Ehsan Abdolmajidi, Haiqi Xu and Lars Harrie

Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden

ABSTRACT

Map mashups, as a common way of presenting geospatial information on the Web, are generally created by spatially overlaying thematic information on top of various base maps. This simple overlay approach often raises geometric deficiencies due to geometric uncertainties in the data. This issue is particularly apparent in a multi-scale context because the thematic data seldom have synchronised level of detail with the base map. In this study, we propose, develop, implement and evaluate a relative positioning approach based on shared geometries and relative coordinates to synchronise geometric representations for map mashups through several scales. To realise the relative positioning between datasets, we adopt a Linked Data–based technical framework in which the data are organised according to ontologies that are designed based on the GeoSPARQL vocabulary. A prototype system is developed to demonstrate the feasibility and usability of the relative positioning approach. The results show that the approach synchronises and integrates the geometries of thematic data and the base map effectively, and the thematic data are automatically tailored for multi-scale visualisation. The proposed framework can be used as a new way of modelling geospatial data on the Web, with merits in terms of both data visualisation and querying.

ARTICLE HISTORY
Received 19 July 2017
Accepted 13 February 2018

KEYWORDS
Map mashups; geometry synchronisation; multiple representation; relative positioning; Linked Data

Introduction

Map mashups, as a common way of presenting spatial information and the most popular mashups on the Web (Fichter 2009), are generally created by spatially overlaying thematic information on top of various base maps (Moseme and Van Elzakker 2012). However, most commonly, the thematic data have no explicit link to the base map, although there are often intrinsic connections between the features in thematic data and the base map. For example, in a postcode area thematic map, the boundaries of the postcode areas often coincide with e.g. the road, river and administrative border features represented in the base map. This simple overlay approach often raises geometric inconsistencies between thematic data and the base map due to geometric uncertainties in the data. This problem is particularly apparent in a multi-scale context, as the thematic data seldom have synchronised level of detail with the base map. The

CONTACT Lars Harrie lars.harrie@nateko.lu.se; Weiming Huang weiming.huang@nateko.lu.se

level of detail synchronisation of geospatial data from several data sources implies that the data providers of the thematic information and the base map have to follow an agreement in terms of map generalisation rules. However, such cross-organisational agreements can be hardly realised in practice. Alternatively, the synchronisation of level of detail can be realised by a relative positioning approach, that is, modelling the geometries of thematic data (partly) by their relations with geometries of the features in the base map. Essentially, this type of relative positioning pertains multi-scale data integration and the sharing of geometric elements between features. This technique is sometimes used within a single geospatial dataset where instances in the feature level can share elements in the geometry level. This type of sharing geometric elements is, however, uncommon between datasets.

The relative positioning is common in other spatial data domains, such as Building Information Modelling (BIM). Generally, the geometries in BIM models are constructed using parametric modelling where the locations of the geometric objects are defined relative to each other (see e.g. Eastman *et al*. 2011). This is an intuitive approach for buildings because they have a well-defined hierarchical structure (a building contains floors, each floor contains rooms, etc.). However, the relative positioning could also be exploited in geospatial applications, in particular for geospatial data integration. In this study, we explore two types of relative positioning: sharing geometries between thematic data and the base map; and relative coordinates used when there is no correspondence relation between thematic data and the base map.

The definition, realisation and use of relative positioning for geospatial data are non-trivial because geographic features generally do not have a well-defined hierarchical structure. The emerging Semantic Web technology stack, particularly the part concerning Linked Data, provides a promising technical framework that can be used to establish explicit links between thematic data and the base maps to enable geometrically self-adapting (synchronised) thematic maps. A prerequisite of this is that it is envisioned that the base map in different scales would be available as Linked Data, then thematic data could be created by linking to base map data using the Linked Data paradigm.

This article proposes, implements and evaluates a relative positioning approach using Linked Data technologies for the purpose of synchronising geometric representations in map mashups. Following the introduction section, Section 2 provides a brief overview of previous relevant studies on geometry synchronisation and Linked Data. In Section 3, a case study on the use of relative positioning for creating geometrically synchronised map mashups is described. Finally, the paper ends with a discussion and conclusions.

Previous studies

2.1. *Synchronisation of geometric representations in map mashups*

There are several techniques proposed for creating map mashups in a multi-scale context. Most of those studies have concentrated on thematic point data; see Korpi and Ahonen-Rainio (2013) for an overview. One example is Bereuter and Weibel (2013), who proposed a quad-tree based method for real-time generalisation supporting progressive levels of detail in the zooming process. The thematic features in most of these studies only have loose connections with the base maps; therefore, the synchronisation

and integration of the geometric representations of the thematic features and the base maps through scales were not a significant problem. The synchronisation problem was encountered by Stern and Sester (2013) when they studied map mashups of natural protected areas on top of a base map, where the protected areas often have common geometry elements with the base map. To overcome the problem of the inconsistencies in the multi-scale representation, they argued that the base map should act as constraints for generalising the thematic data. Furthermore, Toomanian *et al*. (2013) defined the semantic relations between feature types in the thematic data and the base map in the map mashups using ontologies. These semantic relationships together with map matching were then used to enable real-time adjustment of the thematic features to the base map. However, their study concentrated on the integration between thematic data and a single-scale base map, not applying a multi-scale context.

2.2 *Linked Data*

'Linked Data' is a term for the collection of design principles and technologies centred around a paradigm to publish, retrieve, reuse, and integrate data on the Web (Kuhn *et al*. 2014). Linked Data are encoded as Resource Description Framework (RDF) in triples, where a triple is composed of a subject, a predicate and an object; to express these triples, the World Wide Web Consortium (W3C) has adopted several serialisation syntaxes. The applications of Linked Data have developed considerably in geospatial domain in recent years because of its significant advantage in terms of data integration and it fosters a promising approach to connect spatial data infrastructures with the mainstream IT to augment the application of geospatial data (Schade and Smits 2012).

Linked Data are usually built upon defined ontologies as vocabularies for organising data. Ontologies are agreements about shared conceptualisations in corresponding domains (Fonseca *et al*. 2006) and play an important role in terms of knowledge sharing. They are formally described in ontology languages, e.g. Web Ontology Language (OWL)[1]. In recent years, ontology design has become popular for formalisation of geospatial concepts on different aspects. For example, Carral *et al*. (2013) designed an ontology for cartographic map scaling, which formalised the scale information on the dataset level. Ontologies can reduce costs and improve the accuracy of integration by making the semantic differences of geospatial data explicit (Hart and Dolbear 2013). As a result, they have been used for facilitating the representation and integration of geospatial data in a number of studies (e.g. Couclelis 2010, Farnaghi and Mansourian 2013, Hong and Kuo 2015).

Some techniques within the framework of Linked Data have been extended in order to improve the handling of geospatial Linked Data. SPARQL, the query protocol for RDF, has a standardised geospatial extension – GeoSPARQL (Perry and Herring 2012). GeoSPARQL defines an ontology to provide a common representation model and a standardised exchange basis for geospatial Linked Data. And it also provides a number of SPARQL query predicates and functions to facilitate the queries using geometric and topological relations between geospatial entities (Battle and Kolas 2012). Currently, there are a few implementations of GeoSPARQL in RDF triple stores, e.g. Parliament[2] and Stardog[3].

For the use of Linked Data in the geospatial domain, another important issue is how to create URIs for geospatial data. In Europe, the INSPIRE directive provided a guideline

for the design of URIs for environmental geospatial data (INSPIRE 2014). Ordnance Survey (OS), the national mapping agency (NMA) of the UK, has a set of organisation-specific IDs for each geographic feature, and the OS reused them when creating URIs for their geospatial Linked Data sets (Goodwin *et al*. 2008; Ordnance Survey 2016). *Geonovum* in the Netherlands has proposed a URI design strategy for geospatial data, in which they took into account the criteria of e.g. persistence, scalability and trust (van den Brink *et al*. 2014, URI-strategie linked open data 2018).

There have been some projects publishing geospatial data as Linked Data. Stadler *et al*. (2012) described the LinkedGeoData project that transformed OpenStreetMap (OSM) into Linked Data; the transformed data were linked to other data sources, e.g. DBpedia[4] and GeoNames[5]. The US Geological Survey (USGS) developed ontologies for *The National Map* and converted certain datasets to RDF so that these data can be downloaded and queried as Linked Data (Usery and Varanka 2012). Patroumpas *et al*. (2015) exposed the INSPIRE-compliant data and metadata as Linked Data by transforming them into RDF using XSLT transformations and then exposing them through (Geo) SPARQL endpoints. Hietanen *et al*. (2016) implemented a prototype service to provide geospatial data as Linked Data, and they employed the GeoSPARQL vocabulary; the data were first retrieved in Geography Markup Language (GML), and subsequently transformed to RDF on-the-fly. These studies brought a number of geospatial datasets which are capable of serving as base maps in map mashups into Linked Data, and it is envisioned that more geospatial datasets will become parts of the Linking Open Data cloud (Abele *et al*. 2017).

Case study

In this case study, in order to synchronise the geometric representations of thematic data and the base map through different scales, we explore the use of relative positioning for creating, storing and visualising thematic data. We have chosen to work with the feature type natural protected areas because the objects of this type often have intrinsic connections with the features in the base map (denoted *background features* hereinafter). Specifically, the extents of natural protected areas are commonly defined by background features such as lakes, rivers, roads, and cadastral or other administrative units. For instance, in Figure 1, a part of the boundary of *Sillmansåsen* is defined by a lane as a background feature. However, their geometries are not synchronised at this part, and as the visualisation scale of the map mashup changes from (a) to (b) (1: 4,000 to 1: 8,000) through a zoom out operation, the changes of the corresponding geometries are also unsynchronised, which is likely to cause geometric conflicts or other types of visualisation deficiencies.

A diagram of our prototype system architecture is shown in Figure 2. The base map is a multiple representation database (MRDB, see e.g. Jones *et al*. 1996), i.e. there are links between the features in different levels of detail that represent same physical entities. In this study, the base map is accessible through a Web Map Service (WMS), a SPARQL endpoint (for the base map in Linked Data) and a download service (in shapefile). The thematic data provider can download the base map to serve as context data and then use a specific editing tool for generating thematic data using the relative positioning approach. The created thematic data are released in RDF through a SPARQL endpoint.

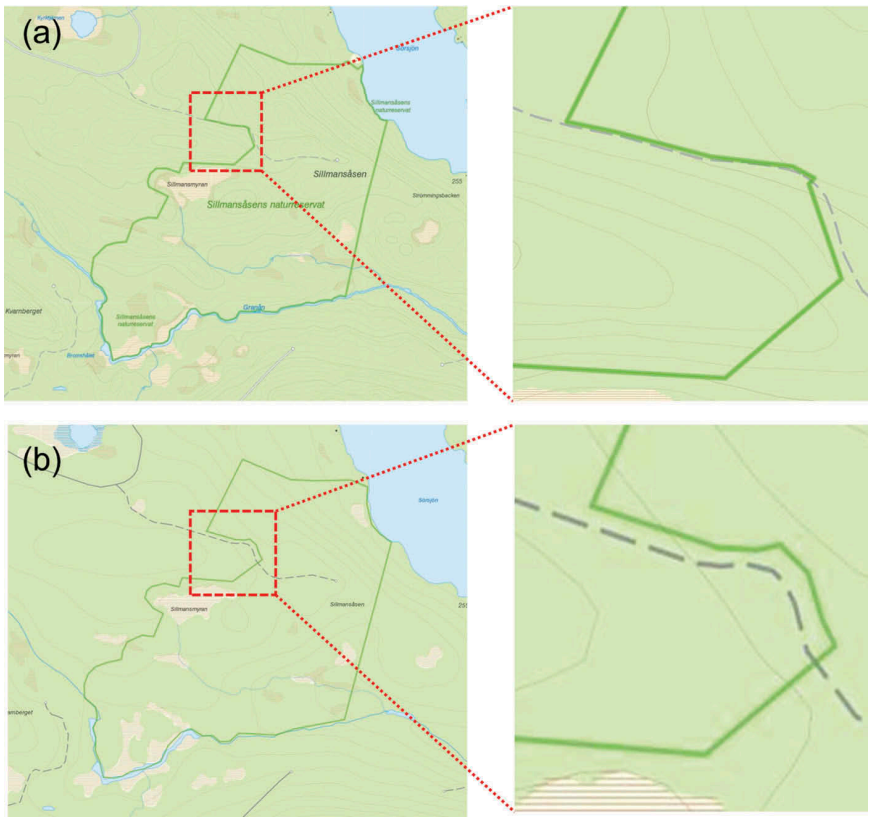


Figure 1. The natural protected area *Sillmansåsen* in northern Sweden. The figure shows the base map and thematic information in two scales (1: 4,000 and 1: 8,000). It illustrates the lack of synchronisation of geometries through scales, and the correspondence relations between the protected area and the background features are vague as shown in the maps. The maps are from *Lantmäteriet* (Swedish NMA) (© *Lantmäteriet*, Dnr: I2014/00579).

A real-time process serves the end users with the map mashups. In this real-time process, the geometrically synchronised thematic data are constructed and added on top of a base map, where the base map is provided by a WMS service for efficiency reasons. Since the geometries of the thematic data are defined relative to the base map, the geometric representations of thematic data and the base map are automatically synchronised in all scales.

Geometric representation using relative positioning

The relative positioning is implemented based on shared geometries and relative coordinates. Specifically, we decompose the geometry of each thematic feature into a set of geometric components; some of these components (denoted as *matched components* hereinafter) are defined by background features, and others (denoted as *independent components* hereinafter) do not have any counterpart in the base map (cf. Figure 6). For matched components, the relations of correspondence are stored. For independent

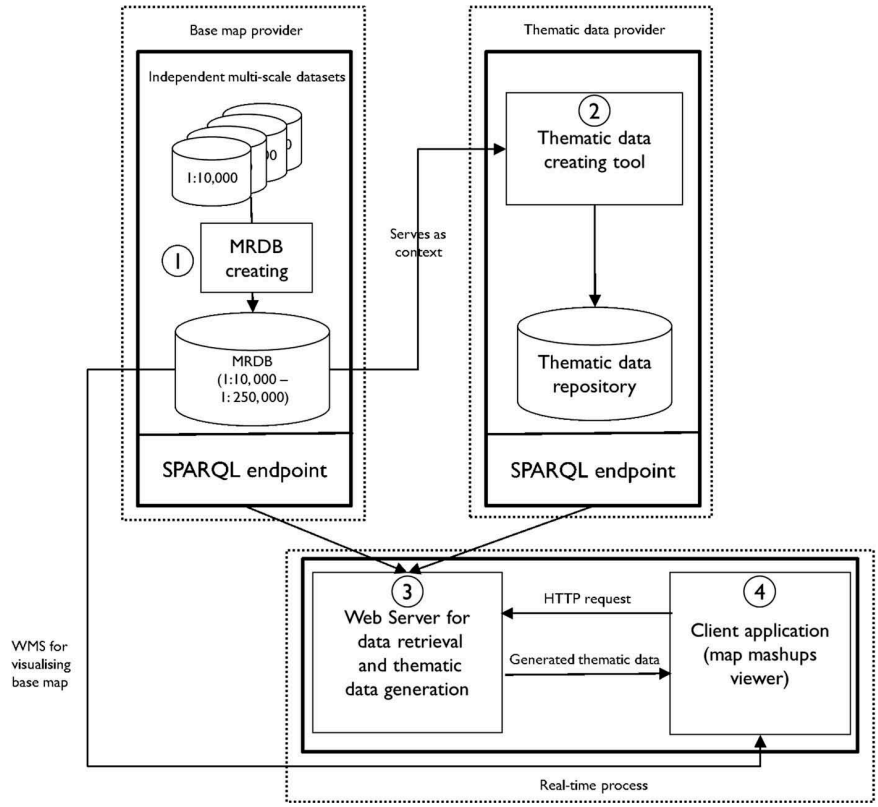


Figure 2. Architecture of the prototype system.

components, the geometries are stored using relative coordinates for adapting to the displacements of their contiguous matched components. When a thematic feature needs to be geometrically represented, the geometry is assembled by combining its components, and the assembled thematic features inherit the coordinate system from the base map. The assembled thematic features share geometric elements with the base map at the matched components, and the integration and synchronisation through scales in map mashups are thereby enabled.

In order to locate the matched components, the starting and ending points of each matched component are necessary. The starting and ending points define the part of the background feature used for assembling the thematic data geometry at a later stage; they are created in the most detailed scale, thus they can only be approximate values for other scales because the geometries of the background features vary in different scales. When a thematic feature needs to be generated, its matched components need to be fetched by splitting the respective background features using the nearest points of the starting and ending points in the scale that the thematic feature will be represented. In addition, for closed geometries, we utilise a stored indicator of the direction of the matched component. The geometries of a background feature in some scales, especially

coarser scales, are possibly not available; in this case, the geometry of the matched component is fetched from the closest scale where a geometry is available.

The independent components, however, are located in a different way. In principle, the previous and next components of each independent component are matched ones, thereby an independent component needs to be adjusted to the displacement of its contiguous matched components as the hosting background features displace during the change of visualisation scale. Therefore, we use relative coordinates in a local coordinate system in which the origin is the last vertex of the previous matched component to store the location of each independent component. This ensures that the head end of its geometry remains at the same position relative to the rear end of its previous matched component. During the feature assembly, linear scaling is applied to coordinates separately in the *X*- and *Y*-directions of each vertex to adjust the rear end of the independent component to the displacement of the head end of its next matched component, and the coordinates are then transformed back to absolute coordinates. For a feature that is entirely independent (i.e. contains no matched components), the location of the first vertex is stored in absolute coordinates. In short, relative coordinates are used for the independent components (instead of absolute coordinates) for avoiding potential geometric deficiencies due to the displacement of the contiguous matched components in coarser scales.

In this study, we work with the assumption that the spatial extents of the natural protected areas have formal definitions and the features will be defined accordingly using relative positioning. The matched components coincide with the corresponding parts of the background features; whereas in reality, it is also possible that a part of boundary of a thematic feature is defined as 'some metres away from a background feature', and in that case the overall workflow is generally the same except more semantic information needs to be stored and extra geometric operations are needed during the feature assembly phase.

Ontology design

For the ontology design, there are two major types of design patterns: logical patterns and content patterns (Hu *et al*. 2013). This paper adopts the content pattern to address the design of classes and properties for the formalisation of spatial concepts and relations in the relative positioning approach, and facilitating the data retrieval for the purpose of synchronising and integrating geometric representations in map mashups. We conceptualise the ontology design pattern using competency questions (Gruninger and Fox 1994) such as:

(1) Which components does the thematic feature have, and which of them are matched/independent components?
(2) For a matched component, what are the coordinates of its starting/ending point?
(3) For a matched component, what is its hosting background feature?
(4) How does the geometry change for a feature at different levels of detail (scales)?
(5) For an independent component, what are its coordinates?
(6) In what order should all the components be assembled?

Based on the competency questions, the base map ontology and the thematic data ontology are designed as OWL ontologies, and both are based on the GeoSPARQL ontology. Below we use the prefix *geo* to represent the namespace of the GeoSPARQL ontology[6], and the prefix *sf* to represent the namespace of simple feature geometries in GeoSPARQL ontology[7]. The GeoSPARQL ontology is selected because it provides general concepts for geospatial data, and the data can then be shared more readily with other geospatial datasets in RDF following the same standard.

Figure 3 shows the key concepts within the designed ontologies and their connections with each other and with the GeoSPARQL ontology. For the base map ontology, a class *Background_Feature* is created as a subclass of *geo:Feature*, and each instance of this *Background_Feature* is connected to one or more *sf:Geometry* instance(s) using the object property *geo:hasGeometry*. Furthermore, to serve the competence question 4, we introduce a class *Scale* and two datatype properties *hasUpperBound* and *hasLowerBound* to model the scales; the concept *scale* in this study refers to the visualisation (rendering) ranges of the geometries adopted in the multi-scale base map (the map served from *Lantmäteriet*'s WMS in this study, cf. Figure 2). Each *Geometry* instance is then associated with a *Scale* instance through an object property *hasScale* to indicate the scale range of its visualisation. The coordinates are stored in literal Well-Known Text (WKT), and the literals are associated with corresponding geometry instances through the *geo:AsWKT* property.

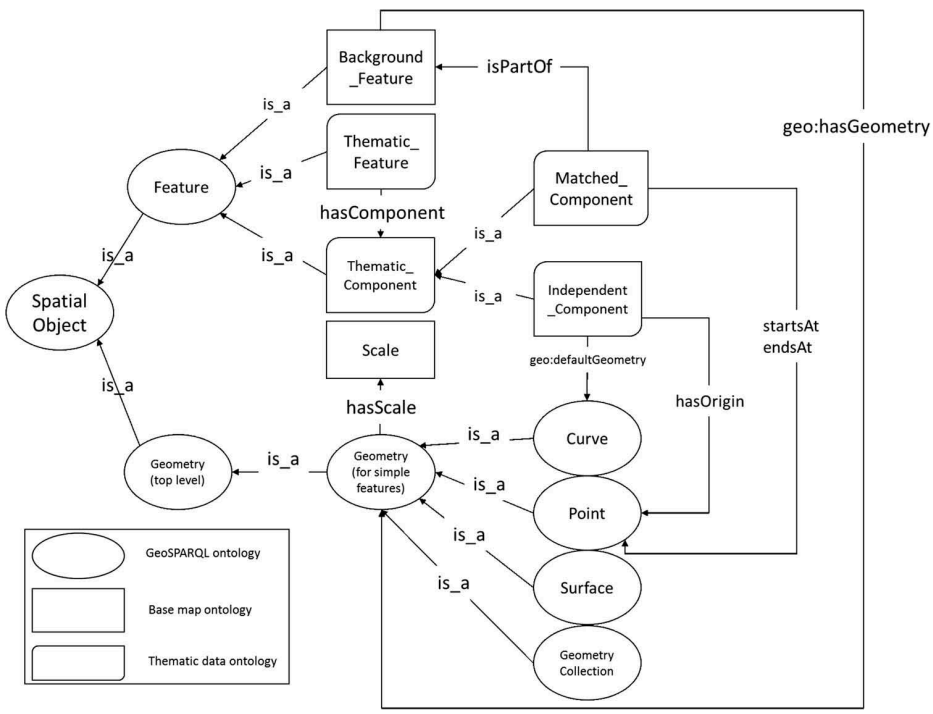


Figure 3. Diagram of the two created ontologies, including their connections with each other and with the GeoSPARQL ontology (datatype properties and several classes in GeoSPARQL ontology are not shown).

For the thematic data ontology, a class *Thematic_Feature* is created as a subclass of *geo:Feature*. We also create a sibling class of *Thematic_Feature* – *Thematic_Component* and its two subclasses *Matched_Component* and *Independent_Component*. Five object properties are defined: *hasComponent* is used for connecting a thematic feature with its components, *isPartOf* is used for connecting a matched component with its hosting background feature, *startsAt* is used for connecting a matched component with the point where it starts, *endsAt* is used for connecting a matched component with the point where it ends, and *hasOrigin* is used for connecting an independent component with the origin (in absolute coordinates) of its relative coordinates. A datatype property *verticesOrder* is defined for indicating the direction of a given component that is matched to a closed geometry. Another datatype property *componentOrder* is defined to indicate in which order these components should be assembled. For the situations in which some of the components are used to compose interior rings, a datatype property *innerRingNo* is defined for distinguishing their corresponding interior rings.

Implementation

The implementation is released under a GPL license and distributed through GitHub (https://github.com/RightBank/Relative-positioning-implementation/). For licensing reasons, we are not allowed to add the geospatial data used for this case study to GitHub. The description of the implementation below is structured according to the numbers given in Figure 2.

(1) MRDB publishing

The MRDB as Linked Data was published based on the base map ontology (Figure 3) that was created in the open-source ontology editor Protégé[8]. Protégé enabled us to graphically view relations that existed between classes or properties in ontologies and manually add the necessary axioms and restrictions. In our study, the MRDB was stored in shapefiles and we developed a Python script to convert it to RDF according to the base map ontology. This convertor was developed using GDAL 2.1.3[9] and RDFLib 4.2.1[10], in which GDAL was used for reading geospatial data from shapefiles, and RDFLib was used for writing geospatial data into RDF; the created RDF triples were then added to the triple store Stardog.

(2) Creation of thematic data

The thematic data ontology was created in the same way as the base map ontology. An ArcGIS Python add-in tool using Arcpy[11] was developed to act as a digitalisation tool that enables users to create thematic data using relative positioning (for details, see Xu 2017). In this case, natural protected areas were created fully/partially relying on the base map. During the creation of thematic data, for the matched components, the user needs to specify the hosting background features, starting and ending points, etc. according to the thematic data ontology; for the independent components, the user needs to digitise them manually, then the relative coordinates are generated by the tool. A snap functionality ensures that the starting point of next component coincides exactly

with the ending point of the previous component. The created thematic features were finally exported to Stardog as RDF following the thematic data ontology.

(3) Retrieval and generation of thematic data

The process of generating thematic features, in particular their geometries, was implemented in a Python backend server. The free and open source Web framework Django[12] was employed. The Python server sends HTTP requests wrapping SPARQL queries to the SPARQL endpoints provided by Stardog to retrieve data in real-time. GDAL was used in the Python server for parsing WKT and conducting necessary spatial operations. The thematic data are wrapped into GeoJSON[13] objects and sent to the client side as the server receives certain HTTP requests.

(4) Client application

A client application was developed to enable users to explore the web map using relative positioning. The client side was implemented in HTML and JavaScript. For the retrieval of thematic features from server side, AJAX calls (wrapping the current rendering scale and other parameters) are used to fetch GeoJSON objects that are then processed by a callback function. In order to obtain the current rendering scale, the standardised rendering pixel size is defined to be 0.28 mm×0.28 mm, unless the information of actual pixel size of the final display device is available. The base map was retrieved through the WMS service provided by *Lantmäteriet*. The JavaScript library Leaflet[14] was used to handle the GeoJSON objects and visualise the features contained in them, and to connect to the WMS server and visualise the base map.

Study area and data

Figure 4 shows the test area used in this case study. This area is in *Västernorrland* County, Sweden, and it contains six natural protected areas (including *Sillmansåsen* in Figure 1). Its area is approximately 436 km^2.

Base map – multiple representation database

The original base map includes independent topographic and cadastral datasets in several scales that use absolute coordinates to locate each geographic object. From these datasets, we manually created an MRDB (even though automatic methods do exist, see e.g. Harrie and Hellström 1999, Bobzien *et al*. 2008). Note that we created links between geometries in different scales only for the background features (by introducing URIs on feature level) that are of concern in this case study, i.e. the background features that have connections to the thematic features. We use file repositories (shapefiles) to serve as the created MRDB.

The geometric representations of the topographic features in the base map vary as the scale changes. For example, as demonstrated in Figure 5, the geometric representation of a part of a river *Granån* in this test area in the most detailed scale is a polygon, and it changes to a combination of three polylines and two polygons as the map zooms out to the next two coarser levels of detail, and finally it becomes a single polyline in the least detailed

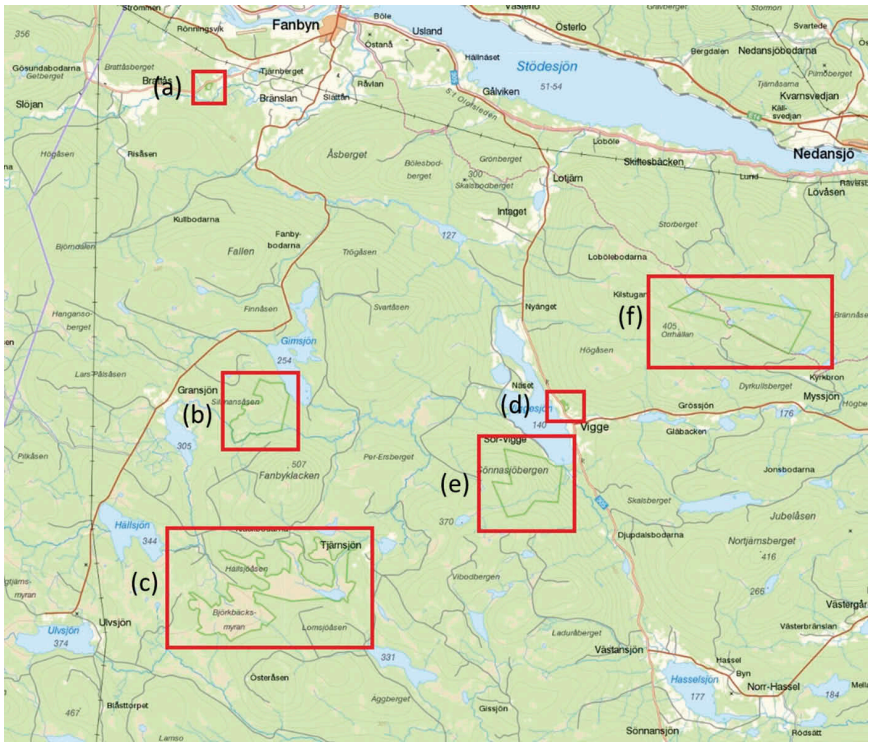


Figure 4. The study area. The six natural protected areas are depicted in red solid rectangles. (© *Lantmäteriet*, Dnr: I2014/00579).

Figure 5. Illustration of a part of the river *Granån* in the test area from original independent multi-scale datasets from *Lantmäteriet*.

scale. The issue here is to define what we mean by a *feature* to enable linkage through scales. Our approach is to define and assign a URI to each feature in the most detailed level.

The reason for this is that we always link the thematic features to the base map in the most detailed scale. As the base map is zoomed to less detailed scales, the corresponding matched components of the thematic features change synchronously. An implication of this is that the definition of each feature should be consistent in all scales; for example, in Figure 5, all parts of the given feature (a river) are treated as one single feature in all scales (i.e. the feature has the same URI in different scales).

Listing 1 partially shows how the information in MRDB corresponding to Figure 5 is released in RDF. In this example, when the scale is larger than 1: 10,000, this feature has one geometry (a polygon instance) as its geometric representation; in scale ranges 1: 10,000 to 1: 50,000 and 1: 50,000 to 1: 100,000, this feature has a geometry collection (three line string instances and two polygon instances) as its geometric representation;

```
@prefix : <[namespace of the base map MRDB in RDF]>
@prefix bm_ontology: <[namespace of the base map ontology]>
[other prefix definitions, e.g. xsd for XML schema]
# the geometries in four levels of detail of this background feature
:[feature_id] a bm_ontology:Background_Feature ;
              geo:hasGeometry :[geometry_id_0],
                              :[geometry_collection_id_0],
                              :[geometry_collection_id_1],
                              :[geometry_id_1].
# the geometry in most detailed level is a polygon and should be rendered in a certain scale
# range
:[geometry_id_0] a sf:Polygon;
                 bm_ontology:hasScale :[scale_id_0];
                 geo:asWKT [WKT literal]^^sf:wktLiteral.
:[scale_id_0] a bm_ontology:Scale;
              bm_ontology:hasUpperBound [scale upper bound value]^^xsd:float;
              bm_ontology:hasLowerBound [scale lower bound value]^^xsd:float.
```

Listing 1 : A snippet of RDF Turtle representation of linking geometries in the MRDB in Figure 5.

and in the scale range of 1: 100,000 to 1: 250,000, its geometric representation returns to one single geometry (a line string instance).

Thematic data

The thematic data were generated by the aforementioned digitalisation tool (2 in Figure 2). Listing 2 is an RDF representation snippet of the created thematic data. In

```
@prefix : <[namespace of thematic data in RDF]>
@prefix td_ontology: <[namespace of thematic data ontology]>
@prefix bm_data: <[namespace of the base map MRDB in RDF]>
[other prefix definitions, e.g. xsd for XML schema]
# the thematic feature has two components
:[feature_id] a td_ontology:Thematic_Feature ;
              td_ontology:hasComponent  :[component_id_0],
                                        :[component_id_1];
# the first component - matched component
:[component_id_0] a td_ontology:Matched_Component;
                  td_ontology:startsAt  :[starting_point_id];
                  td_ontology:endsAt  :[ending_point_id];
                  td_ontology:isPartOf bm_data:[feature_id];
                  td_ontology:componentOrder "0"^^xsd:integer;
                  td_ontology:verticesOrder "reverse"^^td_ontology:[enumeration_datatype_id].
# the second component - independent component
:[component_id_1] a td_ontology:Independent_Component;
                  geo:defaultGeometry :[geometry_id];
                  td_ontology:componentOrder "1"^^xsd:integer.
```

Listing 2 : RDF Turtle representation snippet of thematic data.

this case, a thematic feature has two components: the first component is an instance of the *Matched_Component* class, and it is defined by a part of a background feature and has a starting point and an ending point; the second component is an instance of the *Independent_Component* class, thus its geometry is digitalised and stored as WKT.

Evaluation

The approach was evaluated in the environment that both the client and server were running on the same machine. The experimental computer contains an Intel Core i7-6600U CPU at 2.6GHz, 16GB of memory and a solid-state drive (SSD), running a 64-bit Windows 10 operating system with Python 2.7.8. As a performance test, we generated the six natural protected areas simultaneously in different scales for more than 1,000 times. The generation took 0.45 s in average; the generations in coarse scales took slightly less time than those in detailed scales. To manifest the visual improvement by our method, comparatively to using independent thematic data, we utilised the natural protected areas in the original datasets (cf. Figure 4): these features are referred as *reference features* below. The geometries of the reference features are defined using absolute coordinates and not linked to the base map.

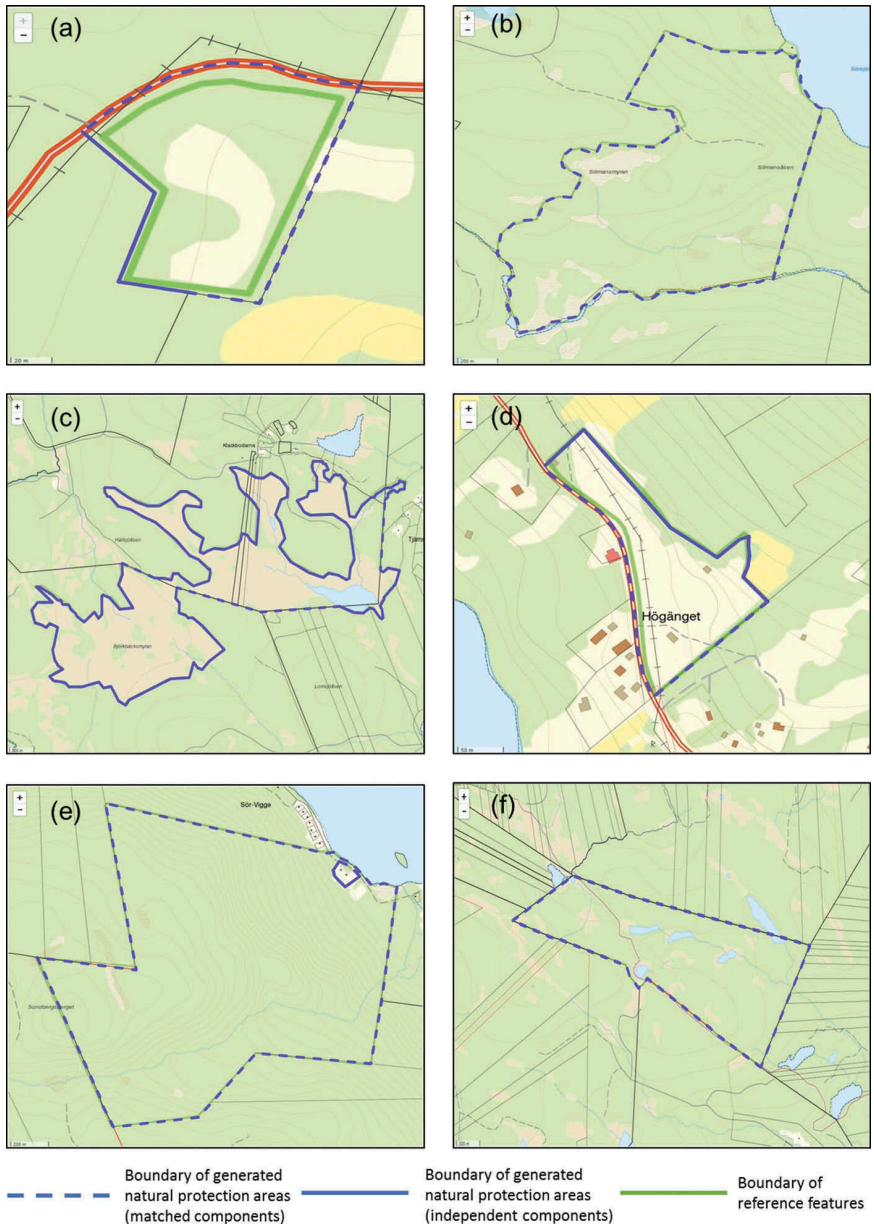


Figure 6. Illustration of each generated thematic feature. The letters a–f correspond to Figure 4.

Figure 6 shows each generated thematic feature overlaying the base map, as well as the reference natural protected areas. The composition of the geometry of each thematic feature is listed in Table 1. These illustrations demonstrate that all the geometries of natural protected areas as thematic features are assembled and the scales between thematic data and the base map are synchronised successfully. Also, all the dependency

Table 1. Composition of the geometry of each thematic feature.

	Number of matched components				Number of independent components	Number of components
	Road	River	Lake	Cadastral Unit		
a	1			1	1	3
b	1	1	1	3		6
c				4	4	8
d	1			1	2	4
e	1		1	1	1	4
f				1		1

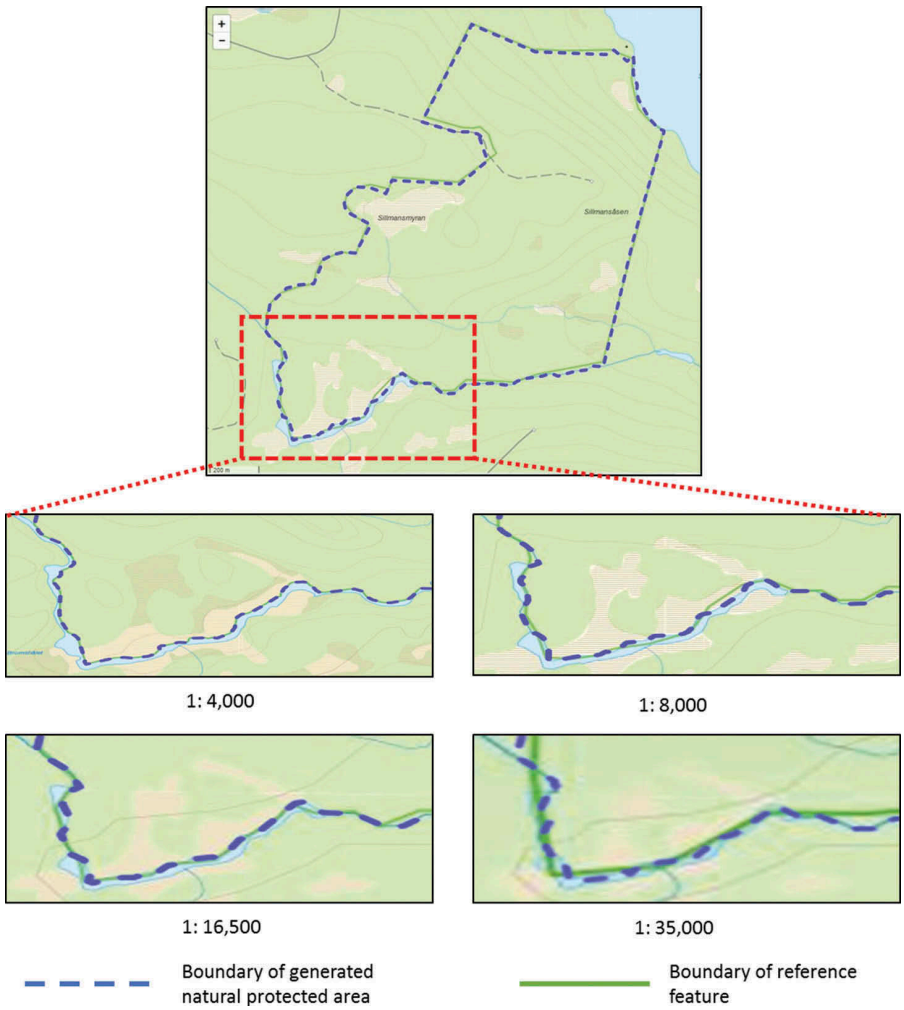


Figure 7. Illustration of the natural protected area *Sillmansåsen* in the scale of 1: 8,000, and enlarged views of the left-bottom part of this feature in scales of 1: 4,000; 1: 8,000; 1: 16,500; and 1: 35,000.

relationships are visually available on the mashup, e.g. in thematic feature *a*, the upper boundary of its geometry exactly fits a road that is a background feature, whereas this matching information is omitted from the geometry of this natural protected area defined by absolute coordinates because of inconsistence in the multi-scale database. Figure 6(e) shows that an interior ring is also successfully assembled.

Figure 7 shows the thematic feature *b* (*Sillmansåsen* in Figure 1) and enlarged views of the left-bottom of this thematic feature in different rendering scales to help to identify the differences between the generated feature and reference feature in difference scales. It demonstrates that although the thematic feature is only created in the most detailed scale, its dependency relations with multi-scale background features enable it to have multiple representations and to be consistent with the background features, e.g. unlike the reference feature, the generated thematic feature's geometry is synchronised with the corresponding part of the river feature in the base map in all scales. To further evaluate the visual improvement of our approach in different scales, we leveraged Hausdorff distance (HD) to estimate the deviations between the geometries of the reference features and their corresponding background features in different scales. The HD values in the four scales are 4.0, 45, 30 and 65 m (from the most detailed to coarser scale levels). The HD values show that the deviations, in general, increase in coarser levels of detail, while the geometries generated by relative positioning approach have no deviation with the base map in any scale because they are generalised in conjunction with the base map.

Discussion

The realisation of the relative positioning approach in this paper utilises three distribution forms of the underlying multi-scale base map data: WMS as a view service to serve the rendered base map; shapefile from a download service to serve as context during the thematic data creation; Linked Data through a SPARQL endpoint to enable the real-time generation of relatively positioned thematic data. Driven by legislation and the open data movement, the multiple distributions of geospatial data are becoming increasingly likely. In Europe, the INSPIRE directive has mandated its member states to provide the view and download services of environmental geospatial data, and it is also investigating the solutions and potential benefits of releasing data as Linked Data in the ARE3NA[15] activity of the European Commission Joint Research Centre, and the draft guidelines for representing INSPIRE data in RDF have been proposed. Currently, several NMAs have released or started releasing authoritative (multi-scale) geospatial data as open data. In the meantime, some NMAs, e.g. OS in the UK, have released the open geospatial data as Linked Open Data (LOD), and some other NMAs are planning or discussing whether the multi-scale geospatial datasets should be released as LOD. In this context, this paper provides a use case of the multi-scale LOD, which can be a strong argument to justify the value of releasing multi-scale base maps as LOD.

In the realm of Volunteered Geographic Information (VGI), OSM data are provided through view services, download services and Linked Data (from the LinkedGeoData project). From this perspective, OSM is suitable to act as a base map to enable the creation of thematic data using our approach, then the created thematic data would have synchronised geometries with the matched features in OSM. This will be useful when creating OSM-based thematic map to have better visualisation performance.

Whereas OSM is not a real multi-scale base map, the geometries of thematic features will generally not be automatically generalised but only be synchronised with the matched features in the base map. On the other hand, VGI can also benefit from the relative positioning through the practice of creating VGI data by linking them to authoritative data, then the highly demanded integration between VGI and authority data can be eased; and given that most VGI data are only produced in one single scale, this is also a way of putting VGI into a multi-scale context.

The focus of this study has been the use of relative positioning and a Linked Data-based technical framework to solve a long-standing visualisation issue in web maps. There are also alternatives to accomplish this goal, e.g. real-time data integration after map matching or real-time generalisation using background features as constraints (cf. Section 2.1). In contrast to these alternatives, our Linked Data-based approach has three advantages: (1) if the geometries in the base map have been updated, the updates can be propagated to the thematic data automatically (persistent URIs are a key to accomplish this, and some NMAs are endeavouring to achieve this goal); (2) if the thematic data need to be visualised in another context (using another base map than the one the thematic data linked to), then the geometries can also be synchronised in a different context (links between the different base map's Linked Data sets or common URIs are crucial for this case, and this is increasingly likely thanks to efforts from the (Geo) Semantic Web community); and (3) our approach reduces the need of computation for feature matching and generalisation in real-time. These three advantages suggest that our approach is promising to enable genuine real-time self-adapting thematic maps. Nonetheless, we could also integrate others' methods to further improve our methodology. For example, when the thematic data have been already produced in absolute coordinates, the semantically enriched map matching (proposed by Toomanian *et al*. 2013) could be used for transforming the thematic data into the data model of our approach.

There are still some issues of relative positioning that should be noted. First, the time efficiency would be one barrier if a large number of features need to be generated simultaneously in real-time. In this case, some caching strategies need to be adopted, and the parallel computation techniques can also be employed to accelerate the feature generation process. Second, this approach requires new routines for the thematic data provides and others that are utilising the base maps for positioning their own data. From a technical point of view, new tools are required, but the main obstacle here is likely to be the change of the traditions of how new geospatial data are produced and positioned. Third, the reliability of the base map needs to be checked before adopting relative positioning. If you start to position your data relative to the base map, you should better be convinced that the organisation providing the base map (most commonly an NMA) has released high quality and up-to-date data, and they would continue to do so. Furthermore, all actors involved in the same mapping project should better agree to use the same base map for thematic data creation. This issue is related to the concept of *trust* in the Semantic Web domain. Carroll *et al*. (2005) argued that Linked Data are trusted depending on: their content, metadata of the Linked Data, and the task the user is performing. To improve the reliability of geospatial Linked Data, Yuan *et al*. (2013) proposed a method to publish geospatial data provenance by analysing how a geospatial metadata catalogue service can be published using Linked Data. This

approach can also be used for Linked Data sets of different base maps to facilitate the thematic data creators' judgement. In short, we believe that the crucial point is that the providers of the base maps should better be trustworthy organisations with a long-term commitment to maintaining their datasets. Fourth, when the thematic data and the base map are not in the same coordinate system, we need coordinate transformation for the starting and ending points of the matched components, and the geometries of the independent components. However, this type of transformation is nothing new, it is also required if the thematic data are simply overlaid on the base maps.

Apart from visualisation, the relative positioning approach, particularly in the Linked Data context, can also be used for many kinds of data querying. The querying capability of Linked Data has gained much attention especially from the geospatial semantics community as the Linked Data paradigm can facilitate the discovery of geospatial data and knowledge. For example, Scheider *et al*. (2014) formally encoded historic map content in Linked Data, and a number of questions about the map metadata and map content were then formulated into SPARQL queries. With our approach, we could also formulate various questions towards the formally encoded spatial relations between thematic and background features in SPARQL and retrieve the answers with low computational cost, for instances:

(1) Which feature type is most involved in the definition of the natural protected areas?
(2) Which national protected areas coincide completely with a single cadastral unit?

Conclusions

This article addresses a long-standing visualisation issue within map mashups, namely the geometric representations between the thematic data and the base maps are usually unsynchronised, particularly in a multi-scale context. In order to solve this problem, this study proposes a methodology that uses relative positioning instead of traditional absolute coordinates to locate geospatial data. Using relative positioning, geospatial features are located relative to background features by e.g. shared geometries. In this study, the relative positioning is implemented using Linked Data technologies, and a use case in a map mashup is designed, implemented and evaluated. The use of relative positioning in map mashups indicates that the relatively positioned geospatial features are naturally integrated and synchronised with the multiple representation background data, i.e. the thematic data automatically obtain synchronised multiple scale representations, which is a prerequisite for proper visualisation. Therefore, the relatively positioned geospatial features avoid substantial visual deficiencies.

Notes

1. https://www.w3.org/TR/owl-features/.
2. http://parliament.semwebcentral.org/.
3. http://www.stardog.com/.
4. http://dbpedia.org.
5. http://www.geonames.org.

6. http://www.opengis.net/ont/geosparql#.
7. http://www.opengis.net/ont/sf#.
8. http://protege.stanford.edu/.
9. http://www.gdal.org/.
10. https://pypi.python.org/pypi/rdflib.
11. http://pro.arcgis.com/en/pro-app/arcpy/main/arcgis-pro-arcpy-reference.htm.
12. https://www.djangoproject.com/.
13. http://geojson.org/.
14. http://leafletjs.com.
15. https://joinup.ec.europa.eu/collection/are3na/about.

Acknowledgments

We thank the editor and the anonymous reviewers for their insightful suggestions and comments that helped to improve the quality of the article. We would also like to thank *Lantmäteriet* for providing the geospatial data used in this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work was funded by China Scholarship Council and Lund University.

ORCID

Weiming Huang http://orcid.org/0000-0002-3208-4208
Ali Mansourian http://orcid.org/0000-0001-6812-4307
Ehsan Abdolmajidi http://orcid.org/0000-0001-7946-2352
Lars Harrie http://orcid.org/0000-0003-3252-1495

References

Abele, A., *et al*., 2017. *Linking open data cloud diagram 2017* [online]. Available from: http://lod-cloud.net/ [Accessed 15 February 2017]

Battle, R. and Kolas, D., 2012. Enabling the geospatial semantic web with parliament and GeoSPARQL. *Semantic Web*, 3 (4), 355–370. doi:10.3233/SW-2012-0065

Bereuter, P. and Weibel, R., 2013. Real-time generalization of point data in mobile and web mapping using quadtrees. *Cartography and Geographic Information Science*, 40 (4), 271–281. doi:10.1080/15230406.2013.779779

Bobzien, M., *et al*., 2008. Multi-representation databases with explicitly modeled horizontal, vertical, and update relations. *Cartography and Geographic Information Science*, 35 (1), 3–16. doi:10.1559/152304008783475698

Carral, D., *et al*., 2013. An ontology design pattern for cartographic map scaling. *In*: P. Cimiano, *et al*., *eds. The semantic web: semantics and big data*. Heidelberg: Springer, 76–93.

Carroll, J.J., *et al*., 2005. Named graphs, provenance and trust. *In*: *Proceedings of the 14th international conference on world wide web*. Chiba, Japan: ACM Press, 613–622. doi:10.1145/1060745.1060835

Couclelis, H., 2010. Ontologies of geographic information. *International Journal of Geographical Information Science*, 24 (12), 1785–1809. doi:10.1080/13658816.2010.484392

Eastman, C., *et al*. 2011. *BIM handbook: A guide to building information modeling for owners, managers, designers, engineers and contractors*. New York: Wiley.

Farnaghi, M. and Mansourian, A., 2013. Disaster planning using automated composition of semantic OGC web services: A case study in sheltering. *Computers, Environment and Urban Systems*, 41, 204–218. doi:10.1016/j.compenvurbsys.2013.06.003

Fichter, D., 2009. What is a mashup. *In*: N.C. Engard, ed. *Library mashups: exploring new ways to deliver library data*. Medford, NJ: Information Today, 3–17.

Fonseca, F., Câmara, G., and Miguel Monteiro, A., 2006. A framework for measuring the interoperability of geo-ontologies. *Spatial Cognition and Computation*, 6 (4), 309–331. doi:10.1207/s15427633scc0604_2

Geonovum, 2018. *URI-strategie linked open data [online]*. Amersfoort: Geonovum. Available from: https://www.geonovum.nl/onderwerpen/linked-data/uri-strategie-linked-open-data [Accessed 20 January 2018].

Goodwin, J., Dolbear, C., and Hart, G., 2008. Geographical linked data: the administrative geography of Great Britain on the semantic web. *Transactions in GIS*, 12 (s1), 19–30. doi:10.1111/j.1467-9671.2008.01133.x

Gruninger, M. and Fox, M.S., 1994. The role of competency questions in enterprise engineering. *In*: *Proceedings of the IFIP WG57 Workshop on Benchmarking - Theory and Practice*. Trondheim, Norway, 1–17.

Harrie, L. and Hellström, A.K., 1999. A prototype system for propagating updates between cartographic data sets. *The Cartographic Journal*, 36 (2), 133–140. doi:10.1179/caj.1999.36.2.133

Hart, G. and Dolbear, C., 2013. *Linked data: a geographic perspective*. Boca Raton, FL: CRC Press.

Hietanen, E., Lehto, L., and Latvala, P., 2016. Providing geographic datasets as linked data in SDI. *In ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B2, 583–586. doi:10.5194/isprsarchives-XLI-B2-583-2016

Hong, J.H. and Kuo, C.L., 2015. A semi-automatic lightweight ontology bridging for the semantic integration of cross-domain geospatial information. *International Journal of Geographical Information Science*, 29 (12), 2223–2247. doi:10.1080/13658816.2015.1072200

Hu, Y., *et al*., 2013. A geo-ontology design pattern for semantic trajectories. *In*: T. Tenbrink, *et al*., *eds. Spatial information theory*. Cham: Springer, 438–456.

INSPIRE, 2014. *D3.4 INSPIRE generic conceptual model*. INSPIRE Drafting Team "Data Specifications".

Jones, C.B., *et al*., 1996. Database design for a multi-scale spatial information system. *International Journal of Geographical Information Systems*, 10 (8), 901–920. doi:10.1080/02693799608902116

Korpi, J. and Ahonen-Rainio, P., 2013. Clutter reduction methods for point symbols in map mashups. *The Cartographic Journal*, 50 (3), 257–265. doi:10.1179/1743277413Y.0000000065

Kuhn, W., Kauppinen, T., and Janowicz, K., 2014. Linked Data – a paradigm shift for geographic information science. *In*: M. Duckham, *et al*., ed. *Geographic information science*. Berlin: Springer, 173–186. doi:10.1007/978-3-319-11593-1_12

Moseme, M.T. and Van Elzakker, C.P.J.M., 2012. Neogeography map users and uses. *In Proceedings of AutoCarto 2012*, Columbus, Ohio, 16-18, 613–622.

Ordnance Survey. 2016. *Ordnance survey linked data* [online]. Available from: http://data.ordnancesurvey.co.uk/datasets/os-linked-data [Accessed 10 February 2016].

Patroumpas, K., *et al*., 2015. Exposing INSPIRE on the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35, 53–62. doi:10.1016/j.websem.2015.09.003

Perry, M. and Herring, J., 2012. *OGC GeoSPARQL – a geographic query language for RDF data* [online]. Technical report, Open Geospatial Consortium. Available from: https://portal.opengeospatial.org/files/?artifact_id=47664 [Accessed 28 August 2016]

Schade, S. and Smits, P., 2012. Why linked data should not lead to next generation SDI. *In*: *Geoscience and remote sensing symposium (IGARSS)*. IEEE, 2894–2897.

Scheider, S., *et al*., 2014. Encoding and querying historic map content. *In*: J. Huerta, S. Schade, and C. Granell, *eds*. *Connecting a Digital Europe Through Location and Place, In: lecture Notes in Geoinformation and Cartography*. Cham: Springer International Publishing. 251–273.

Stadler, C., *et al*., 2012. Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, 3 (4), 333–354. doi:10.3233/SW-2011-0052

Stern, C. and Sester, M., 2013. Deriving constraints for the integration and generalization of detailed environmental spatial data in maps of small scales. *In ICA Workshop on Generalisation and Multiple Representation*, 23–24, August Dresden, Germany.

Toomanian, A., *et al*., 2013. Automatic integration of spatial data in viewing services. *Journal of Spatial Information Science*, 2013 (6), 43–58. doi:10.5311/JOSIS.2013.6.87

Usery, E.L. and Varanka, D., 2012. Design and development of linked data from the national map. *Semantic Web*, 3 (4), 371–384. doi:10.3233/SW-2011-0054

van den Brink, L., *et al*. 2014. Linking spatial data: semi-automated conversion of geo-information models and GML data to RDF. *International Journal of Spatial Data Infrastructures Research*, 9, 59–85.

Xu, H., 2017. *Development of a Digitalization Tool for Linking Thematic Data to a Background Map*, Thesis (MSc), Department of Physical Geography and Ecosystem Science, Lund University, Available from: http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8919462&fileOId=8919470

Yuan, J., *et al*., 2013. A linked data approach for geospatial data provenance. *IEEE Transactions on Geoscience and Remote Sensing*, 51 (11), 5105–5112. doi:10.1109/TGRS.2013.2249523

Paper II

INTERNATIONAL JOURNAL OF DIGITAL EARTH
https://doi.org/10.1080/17538947.2019.1604835

Towards knowledge-based geovisualisation using Semantic Web technologies: a knowledge representation approach coupling ontologies and rules

Weiming Huang and Lars Harrie

Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden

ABSTRACT

Geovisualisation is a knowledge-intensive art in which both providers and users need to possess a wide range of knowledge. Current syntactic approaches to presenting visualisation information lack semantics on the one hand, and on the other hand are too bespoke. Such limitations impede the transfer, interpretation, and reuse of the geovisualisation knowledge. In this paper, we propose a knowledge-based approach to formally represent geovisualisation knowledge in a semantically-enriched and machine-readable manner using Semantic Web technologies. Specifically, we represent knowledge regarding cartographic scale, data portrayal and geometry source, which are three key aspects of geovisualisation in the contemporary web mapping era, coupling ontologies and semantic rules. The knowledge base enables inference for deriving the corresponding geometries and portrayals for visualisation under different conditions. A prototype system is developed in which geospatial linked data are used as underlying data, and some geovisualisation knowledge is formalised into a knowledge base to visualise the data and provide rich semantics to users. The proposed approach can partially form the foundation for the vision of *web of knowledge for geovisualisation*.

ARTICLE HISTORY
Received 28 August 2018
Accepted 3 April 2019

KEYWORDS
Geovisuaisation; Semantic Web; knowledge representation; ontologies; semantic rules

1. Introduction

Geovisualisation is a fundamental, core application of Geographic Information Systems (GIS), and a key enabler for the vision of Digital Earth (Goodchild et al. 2012). It allows users to explore, synthesise, present, and analyse the underlying geospatial data in an interactive manner. Geovisualisation is a knowledge-intensive art where both providers and users are required to possess substantial knowledge about how the geospatial data are visualised, and such knowledge pertains to a wide range of cartographic theories of scaling, portrayal (styles and symbols), etc. For the providers, the knowledge is required to derive sensemaking and cartographically satisfactory applications; for the users, the knowledge is required to interpret the presented data in a meaningful way. Sometimes the users need to reach a high level of cognitive consensus with the providers to better comprehend the information delivered from the geovisualisation applications (MacEachren 2004).

However, geovisualisation knowledge is usually embedded implicitly in complex programs or in the mind of cartographers, which renders the knowledge difficult to be transferred, interpreted, expanded,

CONTACT Weiming Huang weiming.huang@nateko.lu.se GIS Centre, Department of Physical Geography and Ecosystem Science, Lund University, Sölvegatan 12, 223 62 Lund, Sweden

This article has been republished with minor changes. These changes do not impact the academic content of the article.

and reused. To this end, several efforts have been made to standardise the syntaxes used for storing and sharing visualisation information. For example, the Open Geospatial Consortium (OGC) proposed the styled layer descriptor (SLD) standard (Lupp 2007) to represent portrayal information, as used in, for example, the OGC web map service (WMS). However, such syntactic and structural standardisation lacks semantics (meaning of the information), which plays a pivotal role in knowledge representation and information interpretation. The lack of semantics increases the possibility of misinterpretation of the information (Decker et al. 2000; Fellah 2015). The semantic challenge of geovisualisation was also identified by Janowicz et al. (2010), who regarded visualisation as a sink where semantics transferred through all the components of spatial data infrastructures (SDIs) has to be aggregated, interpreted and visualised in a meaningful way. For example, during visualisation, symbols bear abundant semantic information for the delivery of map content to users, and such semantics cannot be fully delivered by the SLD. Furthermore, syntactic approaches like the SLD are too bespoke, and rely on ad-hoc parsers, e.g. to parse the portrayal conditions; such a bespoke nature is also an obstacle in augmenting the transfer of visualisation knowledge to other domains, and its reuse.

Over the last decade, Semantic Web technologies have been increasingly adopted in the geospatial domain to address some long-standing issues, e.g. data integration and reuse (e.g. Schade and Smits 2012), and knowledge formalisation (e.g. Scheider, Ballatore, and Lemmens 2018). Consequently, the amount of geospatial data released as linked data is rapidly growing, a number of geo-ontologies have been designed, and the geospatial linked data query language GeoSPARQL has been standardised (Perry and Herring 2012). This trend fosters the need of representing the visualisation means for geospatial linked data. It also unveils a promising way to mitigate the abovementioned challenges of geovisualisation as the Semantic Web has inherent capacity to formally represent knowledge in a semantically-enriched manner, and such represented knowledge can foster semantic inference to diminish the need for ad-hoc parsers by instead utilising the versatile Semantic Web infrastructure. Moreover, the linked data paradigm provisions a mechanism for interlinking and consolidating distributed information, which produces an opportunity for visualising geospatial data reckoning on semantic and geometric information from diverse sources. These potentials imply the possibility for geovisualisation to move toward a knowledge-based approach.

The research questions that guide this study are: *(1) How can a knowledge base for visualising geospatial linked data be designed?* and *(2) What are the advantages of the knowledge-based approach compared to other means of visualisation, e.g. using the SLD or procedural codes?*

Knowledge-based geovisualisation with Semantic Web technologies implies using geospatial linked data as underlying data. Knowledge concerning how the data are visualised is formalised into a knowledge base consisting of ontologies and rules. Such a knowledge base guides the geovisualisation providers in producing satisfactory applications with formalised knowledge and semantic inference, and also enriches the knowledge represented to the users to ease their perception of map content, e.g. through a semantically-enriched legend with links and relevant resources. The knowledge bases can form part of a *web of knowledge for geovisualisation*, which would facilitate the transfer, reuse, and interpretation of such knowledge within the geospatial domain. This is also a way to augment the usage of such knowledge to other domains in various Digital Earth applications. For example, in the domain of disaster management, knowledge concerning how various geographic objects and events are visualised can be formalised and transferred across several sectors to foster mutual understanding; in the domain of heritage protection, knowledge transfer concerning how heritages are visualised on maps also plays an important role in cross-sector operations.

In this paper, we propose, develop, implement and evaluate a knowledge base in which geovisualisation knowledge is formally represented using ontologies and rules to enable knowledge-based geovisualisation, and to facilitate the transfer, interpretation, and reuse of the knowledge. The background and related work are presented in Section 2. Section 3 elaborates the formalisation of geovisualisation knowledge and the developed knowledge base. The experiments that evaluate the proposed methodology are presented in Section 4. The paper ends with a discussion (Section 5) and conclusions (Section 6).

2. Background and related work

2.1. Geospatial Semantic Web and linked data

The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries (W3C 2013). The Semantic Web is underpinned by a collection of technologies. Linked data refers to a number of recommended best practices for exposing, sharing, and connecting pieces of data, information, and knowledge across the Semantic Web using Uniform Resource Identifiers (URIs) and Resource Description Framework (RDF). The applications of Semantic Web technologies have developed considerably in the geospatial domain in the last decade, and they have fostered a promising approach to connecting SDIs with mainstream IT to augment the application of geospatial data (Schade and Smits 2012). In this context, Vilches-Blázquez et al. (2014) coined the term 'Linked Digital Earth' to represent the scenario where linked data empowers the vision of Digital Earth to facilitate geospatial data integration and retrieval.

With regard to the publication of geospatial linked data, Semantic Web researchers initially showcased the potential of linked data by transforming popular, third-party datasets to RDF, and then more linked data initiatives were run by governmental agencies and large-scale data infrastructures (Regalia et al. 2018). For instance, Ordnance Survey, the national mapping agency (NMA) in the UK, has released several geospatial datasets as linked data (Goodwin, Dolbear, and Hart 2008). In Europe, the e-Government and open data communities are increasingly adopting linked data approaches, and this has motivated the Joint Research Centre (JRC) of the European Commission to investigate the potentials of publishing INSPIRE-compliant data as linked data through the ARE3NA activities (INSPIRE 2017).

The increase in geospatial linked data has stimulated studies concerning techniques for visualising such data. The visualisation of linked data, in general, refers to techniques for visually presenting the links between entities to facilitate the intuitive discovery of underlying information and knowledge (Dadzie and Rowe 2011). For geospatial data, the spatial context is crucial for easing this perception and discovery process. Therefore, the visualisation of geospatial linked data is generally in the form of map mashups, in which the data are spatially presented as thematic data on top of various base maps. To this end, several tools for exploiting such data through visual and graphic interfaces have been developed. For instances, LOD4WFS (Jones et al. 2014) enables geospatial linked data to be queried through the web feature service (WFS) protocol and visualised in GIS programs. SexTant (Nikolaou et al. 2014) allows for visualising and browsing time-evolving geospatial linked data. Map4RDF (Leon et al. 2012) provides the possibility of editing the underlying data and connecting to statistical data. Nevertheless, these tools generally use predefined, hard-coded visualisation settings in their programs, and do not utilise a knowledge-based approach.

2.2. Knowledge representation using Semantic Web technologies

One prominent advantage of harnessing Semantic Web technologies is the inherent knowledge representation capacity equipped with the technology stack. Knowledge representation is a branch of symbolic artificial intelligence, which studies the formalisation of knowledge and its processing within machines (Grimm, Hitzler, and Abecker 2007). Since 1960s, the focus of knowledge representation has evolved through several stages, including general problem solver, expert systems, frame based languages, and rule-based systems, and currently one of the most active areas of knowledge representation research is the Semantic Web. The Semantic Web provisions us with the capacity for representing knowledge, supporting search queries on knowledge and inference. In the Semantic Web, knowledge is represented in different forms, and ontologies (description logics) and rules (horn logic) are the two main paradigms for knowledge representation (Hitzler and Parsia 2009).

Knowledge representation is a good means for precisely conveying long-term intended meaning for expressed information (Lauriault et al. 2007). Specifically, ontologies are controlled vocabularies that describe concepts and relations between concepts using well-understood formal constructs; such constructs formalise the intended meaning of the vocabularies and capture background knowledge about the domain (Horrocks 2008). Ontologies can be connected to further enrich the expressed semantics. Semantic rules are also grounded in formal logic and rich semantics; they can deduce further statements with explanations. Semantic rules are more manageable and understandable than procedural codes to lessen the semantic gap between different parties (Bassiliades 2018). In short, ontologies and rules can provide semantics to disambiguate the meaning of the information concerning how the geospatial data are visualised, and thus foster better transfer, interpretation, and reuse of such knowledge.

Knowledge representation using ontologies and rules has become increasingly prevalent in the geospatial domain, and studies on this subject span several research areas, e.g. visualisation, geoprocessing, semantic geospatial services, and information retrieval. For instance, Hofer et al. (2017) developed a knowledge base to support the composition of geoprocessing workflow, in which ontologies were used to formalise the geooperators, and the Semantic Web rule language (SWRL) was used to formulate the rules associated with geooperator chaining. Scheider, Ballatore, and Lemmens (2018) formalised both geoprocessing tools and the requirements from the users using ontologies and SPARQL CONSTRUCT queries. Falquet et al. (2018) used ontologies and SPARQL CONSTRUCT queries to provide an abstract description for the process of geospatial linked data publication. Keßler, Raubal, and Wosniok (2009) employed ontologies and SWRL rules for context-aware geographic information retrieval, where they used ontologies to organise semantically annotated data and rules for deriving inference for context detection. Gould and Mackaness (2016) formalised knowledge for on-demand map generalisation using ontologies to facilitate the knowledge to be shared, expanded, and reused in the mapping systems.

2.3. Knowledge representation for geovisualisation

The idea of a map as a knowledge base of logical representations is intuitive in view of the implicit concepts and rules inherent in the maps (Kavouras and Kokla 2008). To this end, Varanka and Usery (2018) proposed to semantically represent map content (raw geospatial features) using ontologies to form the knowledge base of maps. we argue that not only can the raw geospatial features be formally represented with this idea, but visualisation knowledge can also be formalised to enrich the knowledge base to clarify how raw data are converted to visualisations, and foster cross-domain and long-term understanding of the visualisations. In addition, several cartographic ontologies have been developed to model common cartographic concepts, e.g. for use in cartographic expert systems (see e.g. Iosifescu-Enescu and Hurni 2007; Smith 2010). However, such ontologies are more like taxonomies; no rule-based inference is enabled, and their models are not in line with the development of contemporary web-based geovisualisation applications.

Geovisualisation has a broad scope and involves multiple aspects of knowledge; from the visual perspective of contemporary web mapping, knowledge concerning *cartographic scale*, *data portrayal* and *geometry source* comprise the core of geovisualisation knowledge. Therefore, in this study, we focus mainly on the knowledge representation of these three aspects, and aim to develop a knowledge base to underpin knowledge-based visualisation.

The concept of *scale* resides at the very core of cartography and is essential to geovisualisation. It also plays a key role in knowledge representation and measurement (Goodchild and Proctor 1997). The modelling of cartographic scale is fundamental to enabling multiple representation of geospatial data, which is a prerequisite to deriving cartographically good visualisations. However, such knowledge is commonly modelled in an implicit way, e.g. by mapping agencies and cartographers. To

address this issue, Carral et al. (2013) designed an ontology for cartographic map scaling, which formalised scale information at the dataset level for representing the scale knowledge associated with geospatial datasets. Huang et al. (2018) formalised the knowledge for both visualisation scales for geospatial features and the relations between thematic data and base maps using ontologies to enable geometrically self-adapting web maps.

The modelling of portrayal information is currently based on established standards, e.g. ISO 19117:2012 (ISO 2012), the SLD, symbol encoding (SE) (Müller 2006) (the SLD is commonly used in WMS, where it is used to link layers to portrayals, while the SE is used to define the portrayal in general; we mainly discuss the SLD in this paper, as they are very similar). The SLD is an XML-based markup language for modelling styles and symbols that are not intrinsically included in geospatial data. In practice, an SLD script is designed for a particular feature type and associated with specific geospatial layers; they must be processed by ad-hoc parsers in order to retrieve the visualisation methods. The SLD has limited capacity for expressing the semantics of the symbols, e.g. assuming a symbol is designed for heritage sites, a computer cannot know that this symbol may apply to a heritage building map. The SLD also has limitations regarding using distributed data, and it still is confined to the niche of layer-based geospatial data management. Moreover, the ad-hoc parsers for parsing the SLD are very few, and are embedded in complex programs, and the portrayal conditions (filters) are often translated to SQL queries so that specific underlying relational databases must be used. The bespoke nature is also revealed in the fact that different parsers handle the same SLD scripts differently (Andersson and Eklöf 2017).

In the context of linked data, it is worth revisiting portrayal information modelling and embracing a knowledge-based approach. This is also advocated by Janowicz et al. (2010) and OGC Testbeds 11, 12, and 13 (Fellah 2015, 2017, 2018). The OGC Testbeds designed ontologies for portrayal information with the initial purpose of semantic mediation of multi-source portrayal data. The ontologies evolved in an SLD-inclined manner through the Testbeds. They mainly modularised the theories into four micro-theories (style, symbol, symboliser and graphic ontologies) to avoid enormous ontology and to underlay better reusability. These studies provide a solid ground for representing portrayal knowledge. In this study, we partially reuse the portrayal ontologies developed during the OGC Testbeds with substantial enrichment and improvement to develop a knowledge base for the representation of portrayal knowledge.

With regard to the geometry source, the linked data paradigm dramatically lowers the barriers to linking distributed data, mainly through the utilisation of URIs; data interlinking is also essential for creating valuable linked data (Berners-Lee 2009). Linked data interlinking has been performed for a number of geospatial linked datasets. For instance, the linked data gazetteer GeoNames[1] has been linked to DBpedia[2]; the LinkedGeoData (a linked data distribution of OpenStreetMap (OSM)) has been linked to GeoNames and United Nations Food and Agriculture Organization geospatial data (Stadler et al. 2012). Such interlinking brings up the opportunity for *integrated visualisation.* For instance, a dataset that has coarse (or even no) geometric information can be visualised with detailed geometries on maps through links with other datasets. This approach would benefit various geovisualisation applications, and thus the modelling of knowledge concerning *which geometry source(s) is used for this visualisation application* would be helpful in enhancing the visualisation knowledge base, particularly in a linked data environment.

3. Knowledge formalisation for geovisualisation

A knowledge-based approach for geovisualisation entails the employment of linked data as the underlying data model, and a knowledge base with knowledge of the means by which the data are visualised. In this section, we elaborate the representation of both geospatial data and visualisation knowledge. This section answers research question *(1)*. The ontologies and example rules are available online.[3]

3.1. Vocabularies for representing geospatial data, metadata, and context

A number of vocabularies for geospatial data have been developed. The commonly used ones are the Basic Geo Vocabulary[4] and GeoSPARQL vocabulary. The GeoSPARQL vocabulary has become increasingly popular, as it allows for embedding spatial predicates in queries. The GeoSPARQL vocabulary is lightweight and represents only some fundamental concepts – essentially the concepts of *feature* and *geometry*; in this paper we use the prefixes *geo* and *sf* to represent the namespaces of GeoSPARQL vocabulary[5] and its simple feature geometry part.[6] Some other geospatial vocabularies with richer semantics and domain knowledge have been developed based on, or with interoperability for the GeoSPARQL vocabulary, e.g. the INSPIRE draft vocabularies.

In Europe, draft guidelines and vocabularies for representing INSPIRE geospatial data in RDF have been proposed, and most of the vocabularies are compatible with the GeoSPARQL query language through the reuse of certain predicates and subclass inheritance. In this study, we adopt the INSPIRE vocabularies that concern 2D buildings to represent geospatial data as linked data, and we showcase our approach through the visualisation of geospatial building data (cf. Section 4). Specifically, we mainly reuse the *bu-base*[7] and *bu-core2D*[8] vocabularies (cf. Figure 1).

The metadata for geospatial datasets is crucial for providing the context for the data. A common practice is to represent geospatial datasets with named graph(s). The named graph is a key concept of Semantic Web architecture, where a collection of RDF statements is organised in a graph with a URI for identification, allowing metadata to be associated with the dataset; the named graphs can also be treated as objects in triples. To this end, the JRC initiated a working group to develop an extension of the DCAT application profile for data portals in Europe (DCAT-AP[9]), and this extension (GeoDCAT-AP[10]) is used for describing geospatial datasets in this context. In this study, we use the developed GeoDCAT-AP as the vocabulary for metadata. Therefore, this work is also an investigation of the benefits obtained from these linked data development endeavours, with a particular focus on visualisation.

The visualisation of geospatial data can benefit from context information to adapt the results to the user's current situation and personal preferences. Semantics plays a pivotal role in modelling context information (Keßler, Raubal, and Wosniok 2009). We aim to develop a knowledge-based approach for visualising geospatial data with context-awareness, i.e. an approach in which a computer is able to deduce different visualisation methods under different client visualisation situations. In this study, we model a lightweight visualisation context ontology with two types of context data: visualisation scale and visualisation phenomenon (theme). We create the class *VisualisationContext*, whose associated values (scale and phenomenon) are updated when the client requests data for visualisation, and these data are stored in a named graph as context information (for detailed usages, see Sections 3.3 and 3.4).

3.2. Formalisation of cartographic scale

A geospatial object can have multiple (geometric) representations with different levels of detail for a real geographic entity. The theory of multiple representation is one of the cornerstones in digital cartography era; representations with different levels of detail are visualised at different visualisation scales (zoom levels). The multiple representation of geospatial data can be organised in multiple representation databases (MRDBs, see e.g. Jones et al. 1996). Hahmann and Burghardt (2010) compared MRDBs with linked data, and they identified several commonalities between these two technologies. They argued that the geospatial objects in both MRDBs and linked data consist of various representations, providing a set of different views of the same object. In this context, linked data eases the linking and reuse of representations of geospatial features.

Most vocabularies for representing geospatial data support the modelling of multiple representations. In the INSPIRE draft building vocabularies, an instance of the *bu-core2d:Building* can be linked to several instances of *bu-base:BuildingGeometry2D* through the object property *bu-core2d:*

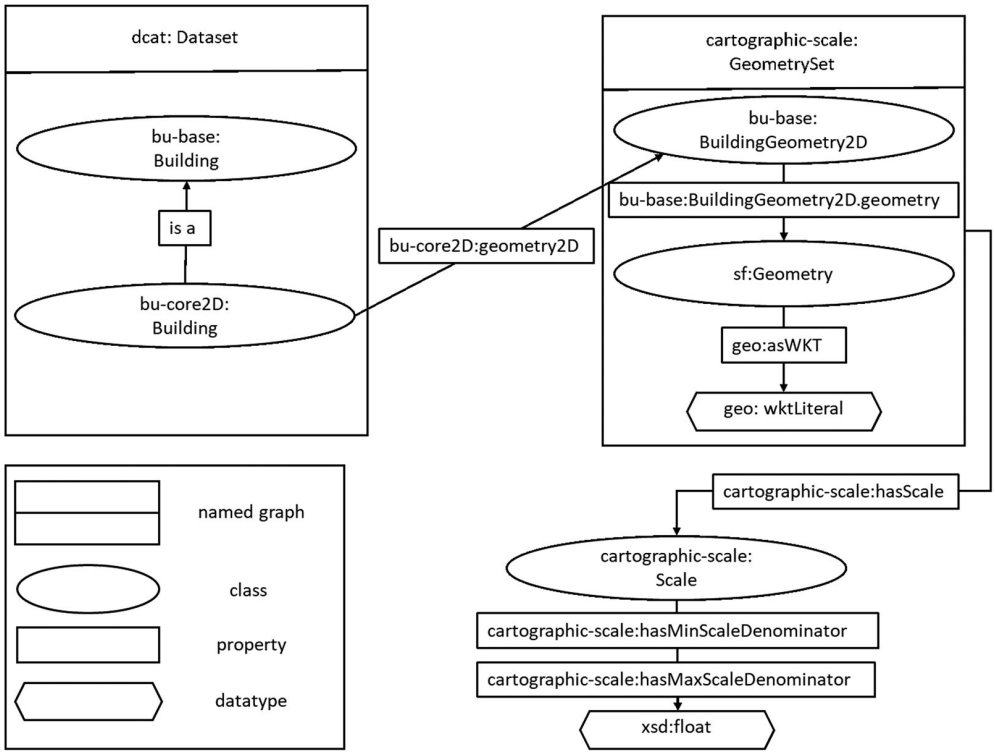


Figure 1. Diagram of the developed cartographic scale vocabulary. The vocabularies used for representing geospatial data are from INSPIRE draft vocabularies for 2D buildings. There can be multiple *GeometrySet* when the geometries are modelled in several levels of detail.

geometry2D to enable multiple representations (cf. Figure 1). However, the information for the cartographic scale in which the representations are rendered is not modelled in these vocabularies, and this is key information for multi-scale visualisation. In this study, we develop a vocabulary for formalising such knowledge (with the prefix *cartographic-scale*).

Unlike previous studies concerning the modelling of cartographic scales (cf. Carral et al. 2013; Huang et al. 2018), we model cartographic scales at the *geometry set* level. Specifically, geometric representations with the same level of detail usually have the same visualisation scales. According to this principle, we develop a cartographic scale vocabulary, where we introduce the concept *GeometrySet*, and encapsulate geometries with the same level of detail in a named graph of the type *GeometrySet*. In the meantime, a class *CartographicScale* is created, and each instance of this class can be linked to the visualisation scale through two datatype properties, *hasMaxScaleDenominator* and *hasMinScaleDenominator*; the object property *hasScale* is created to associate an instance of *GeometrySet* (a named graph) with instance(s) of *CartographicScale*. The cartographic scale information may be different when this knowledge is modelled by different providers and used for different applications. Hence, the metadata, e.g. the application field, is modelled by *hasApplicationField* (an object property whose range is *skos:Concept*), as well as SKOS[11] and Dublin Core[12] vocabularies. Figure 1 illustrates these key concepts and their relations when employing INSPIRE draft building vocabularies for representing geospatial data.

3.3. Formalisation of data portrayal

The visual portrayal of geospatial data transforms raw information into an explanatory or decision-support tool, and plays an indispensable role in map content perception for users to make sense of

the data (Müller 2006). The portrayal bears much semantic information for both information visualisation and retrieval. Janowicz et al. (2010) proposed to semantically annotate the SLD to enrich the semantics and clarify the meaning of styles and symbols presented to users, and to facilitate the recommendation of styles for specific applications. This proposal was accomplished during the OGC Testbed 12, in which ontologies aligned with the SLD standard were developed. However, we argue that, in the linked data environment, a fully SLD-aligned modelling manner, particularly with regard to the modelling of portrayal rules, should be revisited.

Conditional portrayal is prevalent in geovisualisation, i.e. the symbol/symboliser used for visualising a feature depends on the visualisation scale and attribute/geometric data associated with the feature. In the SLD, portrayal rules are modelled by feature filtering using OGC Filter Encodings (ISO/TC211 2009). The ontologies developed by OGC Testbeds also follow this mechanism; SPARQL ASK queries are recommended for modelling such conditions. However, this rule modelling approach has several limitations: (1) although SPARQL can be utilised for expressing rules in the Semantic Web, the queries on their own are not commonly accepted as rule modelling for knowledge representation and inference (W3C 2007), and thus this entails the development of ad-hoc parsers for the conditions; (2) the semantics could potentially be misinterpreted because SPARQL ASK constraints are generally used to check whether certain conditions currently hold in the linked data and thereby facilitating verification and inconsistency checks (Knublauch 2011). To address these limitations, we utilise rule-based inference, and thus augment the use of geospatial rules in other areas in mainstream IT.

Rules are a prominent modelling paradigm for the Semantic Web (Horrocks et al. 2005). They offer a simple model of knowledge representation for both domain experts and programmers. There are several approaches to rule modelling in the Semantic Web, and among them, SWRL rules have been used in several geospatial studies (cf. Section 2.2). The prevalence of SWRL is partly due to its support from the Protégé ontology editor[13], and several rule engines and ontology reasoners. However, SWRL has some limitations in geospatial applications. First, SWRL adopts the *open world assumption*[14], and thereof only supports monotonic semantics. In some geospatial applications, we need to tackle *no data* or *voidable* situations; for example, we cannot use SWRL to represent the rule: *use this specific symboliser to symbolise the feature if the value of a particular attribute does not exist* because this rule entails the handling of non-monotonic semantics. In contrast to SWRL, the object-oriented SPIN (SPARQL Inferencing Notation) rules, which combine concepts from object-oriented languages, SPARQL query language, and rule-based systems to model rules in the Semantic Web, has better expressiveness and several advantages in geospatial applications. For example, SPIN rules can address non-monotonic semantics, and readily allow spatial predicates to be embedded in the conditions within spatially enabled RDF stores (e.g. Stardog[15] and Virtuoso[16]). Therefore, we argue it is time for geospatial Semantic Web researchers to consider a transition from SWRL rules to SPIN rules (before its successor SHACL[17] is better supported by tools). This rule modelling transition is also being advocated by some Semantic Web researchers; see e.g. Bassiliades (2018). In this paper, we use SPIN rules (with the namespaces *spin* and *sp*; for details, see Knublauch (2011)) to model the portrayal rules.

Figure 2 provides an overview of the portrayal knowledge base. We modularise the overall theory into five ontologies, i.e. style, symbol, symboliser, graphic, and legend ontologies; the *symboliser and graphic* ontologies are adopted from OGC Testbed 12; for details, see Fellah (2017). A geospatial linked dataset can be associated with an instance of *style: FeatureTypeStyle* through the property *style:hasStyle* (this relation can be inferred based on semantic relations), the *style: FeatureTypeStyle* is associated with metadata, e.g. using *style:hasApplicationField*. A *style:FeatureTypeStyle* is associated with a portrayal rule base (a named graph whose type is *style:PortrayalRuleBase*), in which all the rules are represented as SPIN rules (*spin:Rule*); the rationale for encapsulating the rules into named graphs is that some RDF stores (e.g. Virtuoso) use the URIs of named graphs to identify the rules that are grouped in which named graphs should be invoked for inference. To facilitate data retrieval, each instance of *spin:Rule* is associated with an instance of *style:PortrayalRule* that is

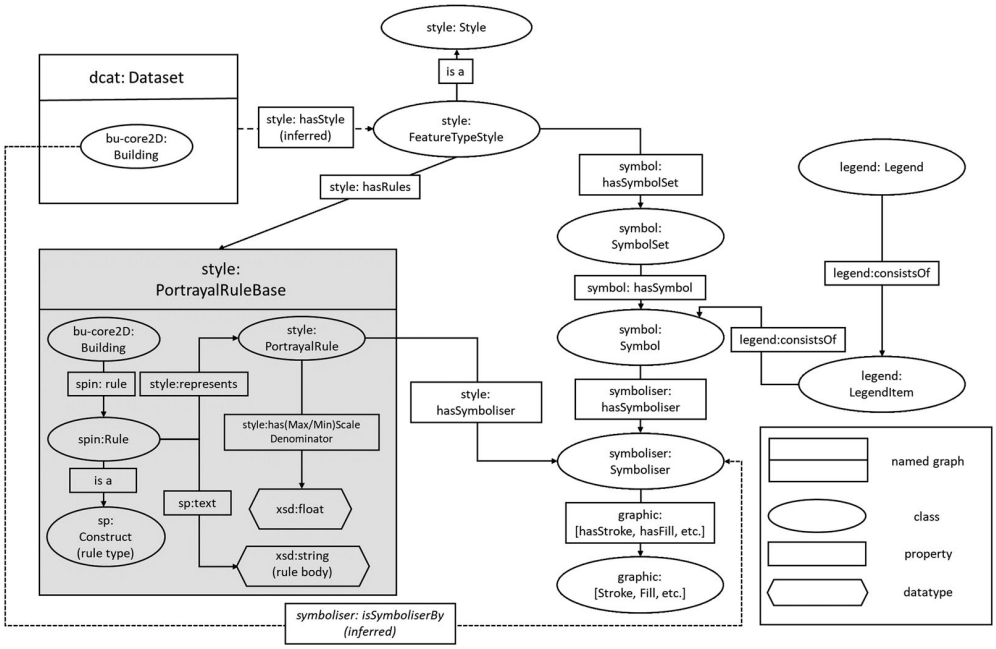


Figure 2. Knowledge base for portrayal information. The grey shaded part is the rule base.

connected to the information about visualisation scale and the associated instance of *symbol:Symbol* (note these relations are modelled as the metadata of the rules and only for the purpose of information retrieval; the inference of SPIN rules relies only on the type and body of the rules). The instances of *symbol:Symbol* are connected to symbolisers and then graphic properties, e.g. the colours of stroke and fill; the graphic information is modelled in an SLD-aligned way. In the meantime, the *symbol:Symbol* instances are used to constitute a legend to enable the knowledge-based generation of map legends and map content retrieval. The labels of the instances in the legend ontology are used to generate the text in legend, and a number of properties, e.g. *legend:represents* (what the legend(item) represents in reality), as well as SKOS and Dublin Core vocabularies are used to model semantic information to facilitate the users interpretation of the map content.

The SPIN rules are encapsulated in a named graph that is processed by the SPIN rule engine. Specifically, we use SPIN CONSTRUCT rules to infer the relation of *symboliser:isSymboliserBy*, and such a rule is attached to the concept describing the geospatial features (*bu_core2D:Building* in this case) due to the object-oriented nature of SPIN. In the main body of a SPIN rule, the inferred relations come first after the CONSTRUCT keyword, and the conditions come afterwards following the WHERE keyword. Listing 1 shows an example of a SPIN rule (in the syntax of Turtle[18]) which infers from the geospatial data and context data (rendering scale transferred from the client) to formulate the rule that *if a building started to be built over 300 years ago, and the rendering scale is larger than 1:10,000, then use the symboliser_0 to symbolise the building*. Furthermore, the *FILTER NOT EXISTS* can be used when it is necessary to deal with the *no data* situation, and spatial predicates (e.g. GeoSPARQL spatial predicates) can be used to develop spatial conditions in spatially enabled RDF stores.

A set of rules can be created and associated with the class *bu_core2D:Building* to enable the usage of different symbols/symbolisers under different conditions. Such rules, when executed, would deduce how to symbolise each feature. The inferred relations can be retrieved through the SPARQL query in Listing 2, which implies that although we define the portrayal rules and symbols only at the dataset level, the relations between features in the dataset and symbolisers are inferred.

```
@[prefix definitions]

bu-core2D:Building a owl:Class;

              spin:rule[a sp:Construct;

              style:representsPortrayalRule portrayal:portrayal_rule_0;

              sp:text"""

              CONSTRUCT {?this symboliser:isSymboliserdBy portrayal:symboliser_0}

              WHERE{

                     ?this bu-base:AbstractConstruction.dateOfConstruction/

                     bu_base:DateOfEvent.beginning ?built_time.

                     BIND(year(now())-year(xsd:dateTime(?built_time)) as ?age)

                     FILTER(?age>300)

                     ?context a context:VisualisationContext;

                     context:hasScaleValue ?rendering_scale.

                     FILTER(?rendering_scale<=10000)}"""].
```

Listing 1. An example of using SPIN rule to represent a portrayal rule in the Turtle syntax.

In addition, we develop a rule for finding appropriate styles according to the visualisation phenomenon; that is, if the visualisation phenomenon and the application field of a style are the same or have a relation among *owl:sameAs*, *rdfs:subClassOf*, *skos:broader*, and *skos:exactMatch*, then the rule-based inference would deduce that the style is applicable to the current visualisation

```
[prefix definitions]

SELECT ?feature ?symboliser

FROM <[URI of the geospatial dataset (named graph)]>

FROM <[URI of the client context named graph]>

WHERE{

       ?feature a bu-core2D:Building;

                symboliser:isSymbolisedBy ?symboliser.

       }
```

Listing 2. The SPARQL query used to retrieve the inferred relations between features and symbolisers.

context, and associate the instance of *dcat:Dataset* with *style:FeatureTypeStyle* through the inferred object property *style:hasStyle*.

3.4. Formalisation of geometry source for visualisation

The geometric representations contained in the geospatial data are one of the most important kinds of information for visualisation purposes. In traditional geovisualisation applications, generally only the geometries contained in the target dataset(s) are used. However, in the linked data environment, geospatial data are increasingly interlinked with each other and data from other domains. Hence, we can adopt an *integrated visualisation* strategy, that is, a strategy in which the visualisation of geospatial linked dataset relies on both the geometries modelled in its own dataset and geometries from other datasets. This is useful when the geometries modelled in the dataset are not sufficient or appropriate (mainly in terms of level of detail) for certain visualisation applications, then the visualisation could (partly) reckon on geometries from other linked datasets to foster better visualisation performance.

Therefore, the representation of knowledge concerning the geometry source(s) used for geovisualisation applications is important in this context. Such knowledge can also be represented by SPIN rules. Specifically, we develop a *geometry source* ontology (its prefix is denoted *gs* in this paper, and the core of it is demonstrated in Figure 3); a geospatial linked dataset can be associated with a *GeometrySourceRuleBase* instance (a named graph) (such an association can also be deduced through the same type of inference for *style:hasStyle* according to contextual information (cf. Section 3.3)), in which the SPIN rule(s) are used to represent the knowledge concerning which geometries should be used for visualisation under different conditions. Metadata are also modelled for instances of *GeometrySourceRuleBase* and *spin:Rule* for information retrieval, e.g. application field of the rule base. Furthermore, the property *isVisualisedBy* is defined to represent the relation between features and the geometries (in well-known text, WKT) used to present them, and this relation is inferred based on the rules.

Listing 3 is an example of a rule that represents the geometry source information. In this example, there are two building datasets that are both modelled using INSPIRE draft building ontologies; one

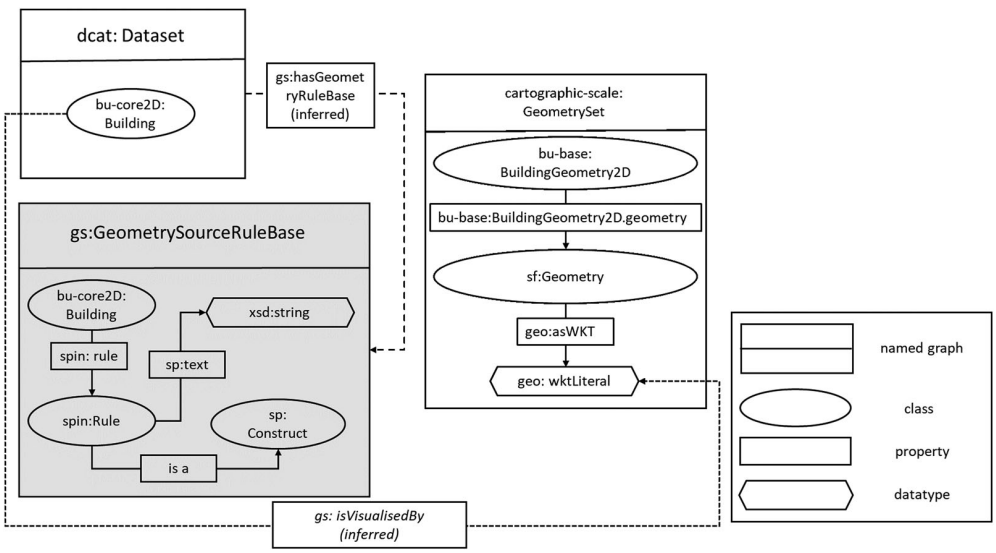


Figure 3. Knowledge base of geometry source(s) used for geovisualisation. The grey shaded part is the rule base.

```
@[prefix definitions]

bu-core2D:Building a owl:Class;

            spin:rule[

                  a sp:Construct;

                  sp:text"""

                  CONSTRUCT {?this gs:isVisualisedBy ?geom}

                  WHERE{

                  ?this skos:closeMatch ?NMABuilding.

                  graph<[NMA dataset URI]>{

                        ?NMABuilding bu_core2D:geometry2D ?2Dgeom.}

                  ?grom_set a cartographic-scale:GeometrySet.

                  graph ?geom_set {

                  ?2Dgeom bu-base:BuildingGeometry2D.geometry/geo:asWKT ?geom.}

                  ?geom_set cartographic-scale:hasScale ?scale.

                  ?scale cartographic-scale:hasMaxScaleDenominator ?maxsd;

                        cartographic-scale:hasMinScaleDenominator ?minsd.

                  ?context a context:VisualisationContext;

                        context:hasScaleValue ?rendering_scale.

                  FILTER(?rendering_scale<=?maxsd && ?rendering_scale>?minsd)

                  }""" ].
```

Listing 3. An example SPIN rule that represents the source of geometries used for visualisation in Turtle syntax. This rule formulates that multi-scale geometries in another dataset are used to render the features in this dataset, and the cartographic scale information modelled in that multi-scale dataset applies.

of them has only coarse geometries, and the other is from an NMA in which multi-scale geometries are stored. The features in these two datasets are linked through *skos:closeMatch*. The visualisation of the first dataset benefits from leveraging the geometries from the second dataset, thus the rule formulates the knowledge that the features are visualised by the multi-scale geometries from the NMA dataset. The rule is also scale-aware, i.e. it infers that different geometries should be rendered for features at different visualisation scales (the visualisation scale transferred from the clients). A set of such rules can be defined to specify that different geometry sources are used under different conditions, and the inferred relations between the features and geometries used for visualisation in

```
[prefix definitions]

SELECT ?feature ?geom

FROM <[URI of the geospatial dataset (named graph)]>

FROM <[URI of the client context named graph]>

WHERE

{

       ?feature a bu-core2D:Building;

                gs:isVisualisedBy ?geom.

}
```

Listing 4. The SPARQL query used to retrieve the inferred relations between features and geometries used for visualisation.

WKT can be retrieved through the SPARQL query in Listing 4. If the two datasets are distributed, the keyword *SERVICE* can be used to retrieve geometries from their distributed sources.

4. Experimentation and evaluation

We test our knowledge-based approach for geovisualisation in a case study for visualising heritage building maps. Figure 4 demonstrates the abstract system architecture of the approach. The architecture comprises three components: data (distributed linked data from two endpoints), a knowledge base and a geovisualisation application (for presenting the visualisation). We implemented all three

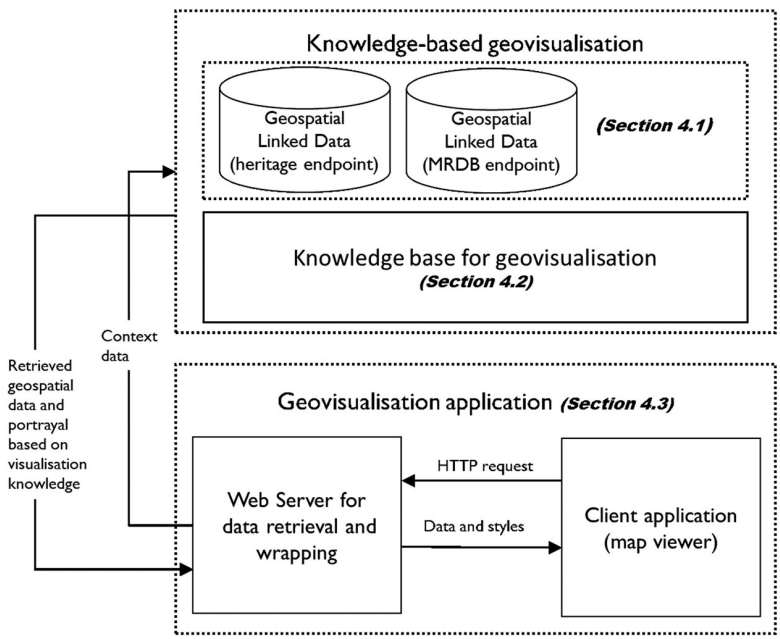


Figure 4. Abstract system architecture of knowledge-based geovisualisation. The corresponding sub-section of each component is annotated in the figure.

components in a distributed architecture in line with the vision of *web of knowledge for geovisualisation*. The implementation details for each part are described in the following sub-sections.

4.1. Data

The study area was central Stockholm, Sweden, and three geospatial datasets (originally all in shapefiles) were used in the experiment (please note that due to licensing reasons, we are not permitted to publish the data):

(*i*) A heritage building thematic dataset from *Riksantikvarieämbetet* (Swedish National Heritage Board). In this dataset, all recorded heritage buildings in Sweden are available as point features, namely a single point is used to represent each building. For most of the heritage buildings, construction time is recorded, while for some buildings, such information is missing.
(*ii*) A building map with detailed geometries from *Lantmäteriet* (Swedish NMA), in which all the buildings are represented by detailed polygons.
(*iii*) Another building map with coarse geometries from *Lantmäteriet*, where large and prominent buildings are represented by coarser geometries than in the previous dataset; other buildings (the small ones) are not present in this dataset. This dataset is used for visualising the buildings on small scale maps.

First, we created an MRDB from datasets *ii* and *iii* using the *spatial join* operation in ArcGIS Pro 2.0.0, and the matched (joined) features were manually checked to ensure each matched pair was semantically correct (we did not consider the aggregated buildings in dataset *iii*; as a proof-of-concept, we employed this semi-automatic matching approach, although more sophisticated methods do exist, see e.g. Zhang et al. (2014); Zhu et al. (2017)). The created MRDB consisted of a number of geospatial features, some of which had two geometries, while others had one geometry. The MRDB was then transformed to RDF according to the ontologies for geospatial data and metadata (see Section 3.1) using R2RML[19] transformations supported by ontop.[20] The geometries with the same level of detail were organised in a *geometry set* (a named graph). The transformed data were exposed through a SPARQL endpoint provided by Stardog (denoted *MRDB endpoint* hereinafter).

Afterward, we matched datasets *i* and *ii* using the same matching method as in the previous step; that is, each heritage building was matched to its corresponding feature in the building dataset from *Lantmäteriet*. The dataset *i* was then transformed to RDF according to INSPIRE draft building vocabularies using R2RML transformations, and the matched features in datasets *i* and *ii* were associated by the property *skos:closeMatch*. The transformed dataset *i* (including the matching relations with dataset *ii*) was exposed through another SPARQL endpoint provided by Stardog (denoted *heritage endpoint* hereinafter).

4.2. Knowledge base for geovisualisation

The cartographic scale information was formalised for the datasets from *Lantmäteriet* (the datasets *ii* and *iii*) as metadata for *GeometrySet*, such information was taken from the recommendations in *Lantmäteriet*, in which dataset *ii* is visualised in large-scale maps, and *iii* is visualised in small-scale maps. The modelled cartographic scale information was exposed along with the metadata of the MRDB through the *MRDB endpoint*.

The knowledge for data portrayal was represented using ontologies and rules (cf. Section 3.3). Specifically, a number of rules were defined to formulate that at different visualisation scales, buildings are symbolised differently according to their ages. Several rules are also defined to form the knowledge about the geometry source and enable inferences concerning the geometries used for visualisation; specifically, in large-scale visualisations, detailed geometries from dataset *ii* were used; at the small scales, coarse polygon geometry was used if it is available for a heritage building

(cf. Listing 3), otherwise point geometry from dataset *i* was used. The portrayal rules were dependent on the geometry source rules because different symbolisers applied according to the types of geometry (point or polygon). The knowledge base for data portrayal and geometry source had the application field *dbpedia:Historic_sites_in_Sweden*.[21] All the semantic rules are stored in Turtle files and can be found online.[22] The knowledge base was implemented using the RDF API Jena[23] and the Topbraid SPIN API[24] (SPIN rule engine). The knowledge base was also implemented as a web service using the Java web framework Spring Boot.[25] This service retrieves data from the two linked data endpoints in a federated manner, which is formulated in the semantic rules.

4.3. Geovisualisation application

We developed a geovisualisation application that was empowered by geospatial linked data and the geovisualisation knowledge base. The application was web-based; its backend is a Python server using the web framework Django[26], and the frontend was developed in HTML and JavaScript employing Leaflet[27] for map visualisation.

In real-time, the frontend sends HTTP requests wrapping the visualisation context in the application to the backend server (the phenomenon to visualise is *dbpedia:Listed_buildings_in_Sweden*[28], which means heritage buildings in Sweden, and has the *skos:broader* relation with *dbpedia:Historic_sites_in_Sweden*), and the backend server updates the context information in a named graph in the knowledge base through the SPARQL UPDATE protocol.[29]The server then retrieves the geospatial data and the symbolisers that are used to symbolise the data (mainly by exposing the SPARQL queries in Listing 2 and Listing 4 to the knowledge base). A lightweight parser is embedded in the backend server to obtain associated e.g. CSS data from the symbolisers. Features and parsed symbolisers are then sent to the frontend, and the features are visualised accordingly. Moreover, the semantically-enriched legend can also be retrieved and visualised in the frontend map to help the users understand the map content.

4.4. Result and evaluation

Knowledge-based geovisualisation can be evaluated with two competency questions:

(1) What geometry should be used to render each feature?
(2) What symboliser should be used to symbolise each feature?

These two questions can be answered by exposing the SPARQL queries in Listings 4 and 2, respectively. The derived answers are context-aware. The queries are simple, because the complex logic lies in the ontologies and rules, e.g. an appropriate style is chosen based on semantic relations; different symbolisers apply depending on the attribute information, visualisation scale, and geometry type; different and distributed geometries are used according to the visualisation scale and the availability of multiple representations for the geospatial objects.

We also evaluated the approach using the visualisation results presented in the frontend application (Figure 5). Figure 5(a) shows the heritage building map at a large scale in the area of *Gamla Stan* (old town, the very centre of Stockholm). In this visualisation, the base map is OSM served through the Mapbox API.[30] The heritage buildings are represented by the detailed geometries from dataset *ii*, and the features are rendered with different colours according to their ages. The ages are derived from the construction time of the heritage buildings recorded in dataset *i*, i.e. this visualisation utilises the semantic information (construction time) from one dataset and detailed geometric information from another distributed dataset. The legend presented in the map is created automatically according to the legend information in the knowledge base. Moreover, the features for which construction time information is missing are also successfully rendered on the map with the corresponding colour, and this indicates that the rule with non-monotonic semantics

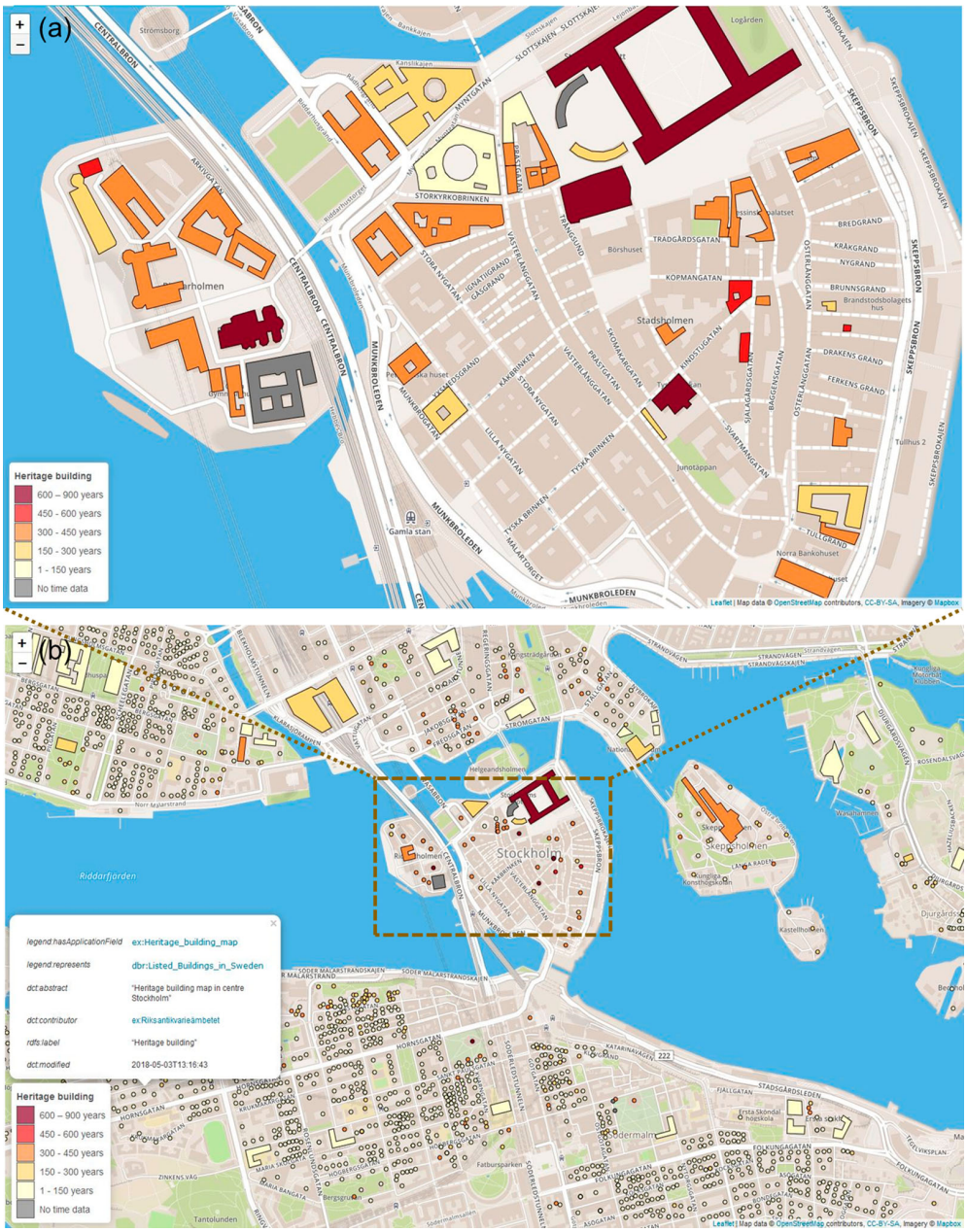


Figure 5. Heritage building maps in central Stockholm underpinned by the knowledge-based geovisualisation. The base map is OSM, and the ages are calculated based on the beginning time of the construction of (part of) each building.

(the rule containing the keyword *FILTER NOT EXISTS* in the conditions) is effectively handled by the rule engine. Figure 5(b) shows the heritage building map at a small scale in a large area of central Stockholm (including *Gamla Stan* in the dashed rectangle). In this visualisation, the large buildings are represented by coarse polygon geometries from dataset *iii*, and the small buildings are represented by point geometries from dataset *i* (with the same colours that indicate building age). Such a combined use of distributed geometry sources is formalised in the SPIN rules, and the

application simply exposes the SPARQL query in Listing 4 to obtain the distributed multi-source geometries for visualisation. The small-scale map gives the map readers a rough sense of how the city expanded in terms of the ages of its heritage buildings. In this way, we observe that the city expanded roughly from *Gamla Stan* to its surroundings, and the buildings by the water are generally older than others.

Furthermore, in Figure 5(b), the enriched semantics of the legend is shown. The knowledge-based legend provides rich semantics that is organised according to the legend ontology (cf. Section 3.3 and Figure 3). In the application, the information shows in a pop-up as the user clicks the legend title. The pop-up provides the meaning of *heritage building in Sweden*, the providers of such information, and so forth; for instance, the legend is associated with the general knowledge base DBpedia's entry *Listed buildings in Sweden* to help the users to explore and understand the map content.

4.5. Comparison with other geovisualisation means

The answer to research question *(2)* was determined with the experiment and evaluation. Although our knowledge-based approach shares some similar design principles with the SLD in terms of the data portrayal, they are fundamentally different. The SLD is confined to the niche of layer-based geospatial data organisation, and is unable visualise data in an integrated and distributed way. For instance, in the small-scale map in Figure 5 (bottom view), the buildings are coloured differently depending on the attribute from one dataset, while the geometries depend on mixed data from two distributed datasets. Such mixed use of semantic and geometric information can be very challenging using the current OGC technology stack, as WMS and WFS do not support federated distributed data retrieval (Zhao, Zhang, and Li 2017) and multiple representations, and the SLD uses a fixed type of symboliser for certain layers (e.g. applying point symbolisers to an entire dataset; the knowledge-based approach is able to assign different types of symbolisers according to the types of the retrieved geometries). Furthermore, the knowledge-based approach substantially reduces the need for ad-hoc parsers for the SLD, as the ontologies and rules are grounded in W3C recommendations (including OWL, RDFS, SPARQL, etc.), which lowers the barrier for the main stream IT community to utilise the geospatial (visualisation) knowledge; this is also in line with the data-centric vision in the IT world. More importantly, the enriched semantics (in contrast to the informal and inexplicit semantics modelled in the SLD) provided by our approach eases the interpretation of information by machines, which is illustrated, e.g. through automated and context-aware style/symbol selection (as shown in this study); it also provides a concrete basis for lifting semantic harmonisation from the data level to the visual level (cf. Karam et al. 2011). Furthermore, users also benefit from enriched semantics, e.g. through the semantically enriched legend, which can be hardly implemented using the SLD.

Using procedural codes (e.g. using a JavaScript mapping library) is an efficient way to develop visualisations, whereas how the data are visualised is not explicitly represented, and is difficult to transfer and interpret, especially by non-developers. Our approach substantially alleviates this issue.

5. Discussion

In this paper, we propose a knowledge-based approach for geovisualisation utilising Semantic Web technologies, in which the knowledge concerning cartographic scale, data portrayal and geometry source is represented using ontologies and rules. The represented knowledge can contribute to the foundation for a *web of knowledge for geovisualisation*, i.e. a distributed knowledge base to provide guidance for geovisualisation, and to facilitate the understanding of the visualisations and thus better reveal the potential for decision-making. The *web of knowledge for geovisualisation* can be used as a *geovisualisation enablement layer* for the linked open data (LOD) cloud to foster sensemaking and cartographically satisfactory visualisations for the increasing geospatial data available in the LOD cloud.

In this study, we discuss and compare our approach with the SLD and procedural codes and illustrate some advantages (see Section 4.5); this does not mean that we believe the latter two methods will

become totally obsolete, and there certainly exist scenarios that they are better adapted to. For example, web developers would be more likely to choose JavaScript libraries. We argue that the knowledge-based approach is more suitable for long-term and cross-domain information transfer and preservation in scenarios such as heritage protection and disaster management, where the visualisations need to be understood by several domains and sectors. Also, for developing visualisations using geospatial linked data, our approach is more appropriate as they stem from the same technology stack.

We believe the presented approach is more friendly for cartographers. Nowadays, the geovisualisation is, in fact, majorly developed by web developers, who sometimes do not possess much cartographic knowledge. Cartographers have to learn rapidly evolving web development techniques to arm the visualisations with their knowledge. With our approach, cartographers could work with domain knowledge modelled in the ontologies and rules, not directly with web development codes, which makes a step forward of moving the visualisation development back to cartographers, who are more (likely to be) competent for mapping.

The knowledge-based approach builds on the premise that the geospatial linked data are interlinked, which is an essential and sometimes expensive work. This is in line with the long-standing research theme of geospatial data matching, which can be difficult especially for complex geometries (e.g. polygon, multi-polygon, etc.). In this regard, the knowledge graph embedding technique (see Wang et al. (2017) for a survey) provides a promising way to facilitate the interlinking of geospatial linked data.

For geospatial data in conventional data models, e.g. PostGIS, our approach can also apply, as the data can be mapped to RDF using e.g. R2RML mapping, the mapped data can be retrieved and queried as virtual RDF graphs. The virtual RDF graph technique is also well-supported by various tools, e.g. Ontop.

A shift in the geospatial domain is underway today: from the creation and maintenance of data, to the creation and maintenance of knowledge as the primary source of value, a.k.a. a transition from SDIs to SKIs (spatial knowledge infrastructures) (Duckham et al. 2017). The SKIs are massively underpinned by Semantic Web technologies. The vision of SKIs is to automatically create, share, curate, deliver, and use knowledge (beyond and not only data or information). To this end, our approach is a way to address the representation of knowledge for visualisation in the SKIs, and such knowledge can also enable context-aware visualisation according to user's specific context, such as previous analysis types and visualisation preference. In order to achieve this goal, we need to incorporate and model more user context data and metadata for the geovisualisation knowledge base to gauge users' purpose and preference, which is a potential future work for this work.

6. Conclusions

This article proposes a knowledge-based approach for geovisualisation in the contemporary web mapping era. We design and implement a knowledge base comprising ontologies and semantic rules to formally represent geovisualisation knowledge in a semantically-enriched and machine-readable manner; the ontologies are mostly dependent on state-of-the-art vocabularies, e.g. GeoSPARQL and INSPIRE draft vocabularies. An architecture for knowledge-based geovisualisation is proposed and a prototype is implemented. A case study using our approach is presented, in which the thematic data for heritage building maps in central Stockholm, Sweden are visualised using the knowledge-based approach.

One incentive of our work is to develop a method for describing the visualisation of geospatial data that are available as linked data. Our approach accomplishes this goal, and such information can be released to the LOD cloud to underlay *a web of knowledge for geovisualisation*, and thus becomes more sharable.

Other advantages of our approach have been unveiled in this study. It is semantically enriched compared with current syntactic methods (e.g. the SLD), and facilitates the clarification of the meaning and selection of the styles and symbols. Furthermore, the enriched semantics is able to foster semantic integration at the visual level for geospatial information. The richer semantics makes the

information easier to interpret and reuse, and eases understanding by visualisation end users. Compared with the state-of-the-art OGC technology stack, our approach enables distributed data retrieval, and thus visualisation, and also supports multiple representations for geospatial features. In the case study, the enriched semantics facilitates the style and symbol selection; the distributed geometric multiple representations foster better visualisation performance; the users benefit from the semantically-enriched legend to better perceive the visualisations.

We believe our approach can also facilitate domain experts (cartographers) to develop geovisualisation, as they will be able to work directly with domain knowledge rather than procedural codes.

Notes

1. http://www.geonames.org.
2. http://dbpedia.org.
3. https://github.com/RightBank/Knowledge-based-geovisualisation.
4. https://www.w3.org/2003/01/geo/.
5. http://www.opengis.net/ont/geosparql#.
6. http://www.opengis.net/ont/sf#.
7. https://raw.githubusercontent.com/inspire-eu-rdf/inspire-rdf-vocabularies/master/bu-base/bu-base.ttl.
8. https://raw.githubusercontent.com/inspire-eu-rdf/inspire-rdf-vocabularies/master/bu-core2d/bu-core2d.ttl.
9. https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe.
10. https://joinup.ec.europa.eu/release/geodcat-ap-v10.
11. https://www.w3.org/2009/08/skos-reference/skos.html#.
12. http://dublincore.org/documents/dcmi-terms/.
13. https://protege.stanford.edu/.
14. http://wiki.opensemanticframework.org/index.php/Overview_of_the_Open_World_Assumption.
15. https://www.stardog.com/.
16. https://virtuoso.openlinksw.com/rdf/.
17. https://www.w3.org/TR/shacl/.
18. https://www.w3.org/TR/turtle/.
19. https://www.w3.org/TR/r2rml/.
20. https://ontop.inf.unibz.it/.
21. http://dbpedia.org/page/Category:Historic_sites_in_Sweden.
22. https://github.com/RightBank/Knowledge-based-geovisualisation/tree/master/rules.
23. https://jena.apache.org/.
24. https://github.com/spinrdf/spinrdf.
25. http://spring.io/projects/spring-boot.
26. https://www.djangoproject.com/.
27. https://leafletjs.com/.
28. http://dbpedia.org/page/Category:Listed_buildings_in_Sweden.
29. https://www.w3.org/TR/sparql11-update/.
30. https://www.mapbox.com/.

Acknowledgments

We thank the editor and the three anonymous reviewers for their comments that helped to improve the quality of the article. We would also like to thank Dr. Carsten Keßler at Aalborg University Copenhagen for his insightful comments and suggestions, and Dr. Oleg Mirzov, Dr. Ali Mansourian and Eiður Eiðsson at Lund University for their advice. We also thank *Lantmäteriet* and *Riksantikvarieämbetet* for providing the geospatial data used in this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by China Scholarship Council and Lund University.

ORCID

Weiming Huang http://orcid.org/0000-0002-3208-4208
Lars Harrie http://orcid.org/0000-0003-3252-1495

References

Andersson, M., and M. Eklöf. 2017. "Stilsättning av geografiska data." Master thesis in Geographic Information Technology at Lund University, Sweden. https://lup.lub.lu.se/student-papers/search/publication/8914527.

Bassiliades, N. 2018. "SWRL2SPIN: Converting SWRL to SPIN." Proceedings of the Doctoral Consortium and Challenge at RuleML+RR 2018 Hosted by 2nd International Joint Conference on Rules and Reasoning, edited by W. Faber, P. Fodor, G. D. Gasperis, A. Giurca, and K. Teymourian, Vol. 2204, CEUR.

Berners-Lee, T. 2009. "Linked Data: Design Issues." http://www.w3.org/DesignIssues/LinkedData.

Carral, D., S. Scheider, K. Janowicz, C. Vardeman, A. A. Krisnadhi, and P. Hitzler. 2013. "An Ontology Design Pattern for Cartographic Map Scaling." In *The Semantic Web: Semantics and Big Data*, edited by P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, 76–93. Berlin, Heidelberg: Springer.

Dadzie, A., and M. Rowe. 2011. "Approaches to Visualising Linked Data: A Survey." *Semantic Web* 2 (2): 89–124. doi:10.3233/SW-2011-0037.

Decker, S., S. Melnik, F. V. Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks. 2000. "The Semantic Web: The Roles of XML and RDF." *IEEE Internet Computing* 4 (5): 63–73.

Duckham M., L. Arnold, K. Armstrong, D. McMeekin, and D. Mottolini. 2017. "Towards a Spatial Knowledge Infrastructure." https://www.crcsi.com.au/assets/Program-3/CRCSI-Towards-Spatial-Knowledge-Whitepaper-web-May2017.pdf.

Falquet, G., C. Metral, S. Ozainne, and G. Giuliani. 2018. "An Abstract Specification Technique for the Publication of Linked Geospatial Data." 21th AGILE conference on Geographic Information Science, Lund, Sweden, June 12–15.

Fellah, S. 2015. "OGC Testbed-11 Symbology Mediation Engineering." Open Geospatial Consortium. https://portal.opengeospatial.org/files/?artifact_id=64385.

Fellah, S. 2017. "Testbed-12 Semantic Portrayal, Registry and Mediation Engineering Report." Open Geospatial Consortium. http://docs.opengeospatial.org/per/16-059.html.

Fellah, S. 2018. "OGC Testbed-13: Portrayal Engineering Report." Open Geospatial Consortium. http://docs.opengeospatial.org/per/17-045.html.

Goodchild, M. F., H. Guo, A. Annoni, L. Bian, K. de Bie, F. Campbell, M. Craglia, et al. 2012. "Next-generation Digital Earth." *Proceedings of the National Academy of Sciences* 109 (28): 11088–11094. doi:10.1073/pnas.1202383109.

Goodchild, M. F., and J. Proctor. 1997. "Scale in a Digital Geographic World." *Geographical and Environmental Modelling* 1: 5–24.

Goodwin, J., C. Dolbear, and G. Hart. 2008. "Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web." *Transactions in GIS* 12 (s1): 19–30. doi:10.1111/j.1467-9671.2008.01133.x.

Gould, N., and W. Mackaness. 2016. "From Taxonomies to Ontologies: Formalizing Generalization Knowledge for On-demand Mapping." *Cartography and Geographic Information Science* 43 (3): 208–222. doi:10.1080/15230406.2015.1072737.

Grimm, S., P. Hitzler, and A. Abecker. 2007. "Knowledge Representation and Ontologies." In *Semantic Web Services: Concepts, Technology and Applications*, edited by R. Studer, S. Grimm, and A. Abecker, 51–106. Heidelberg: Springer.

Hahmann, S., and D. Burghardt. 2010. "Linked Data-a Multiple Representation Database at Web Scale?" Proceedings of the 13th ICA workshop on generalisation and multiple representation, Zürich, Switzerland, September 12–13.

Hitzler, P., and B. Parsia. 2009. "Ontologies and Rules." In *Handbook on Ontologies*, 111–132. Berlin, Heidelberg: Springer.

Hofer, B., S. Mäs, J. Brauner, and L. Bernard. 2017. "Towards a Knowledge Base to Support Geoprocessing Workflow Development." *International Journal of Geographical Information Science* 31 (4): 694–716. doi:10.1080/13658816.2016.1227441.

Horrocks, I. 2008. "Ontologies and the Semantic Web." *Communications of the ACM* 51 (12): 58–67.

Horrocks, I., B. Parsia, P. Patel-Schneider, and J. Hendler. 2005. "Semantic Web Architecture: Tack or Two Towers?" Principles and Practice of Semantic Web Reasoning, Springer.

Huang, W., A. Mansourian, E. Abdolmajidi, H. Xu, and L. Harrie. 2018. "Synchronising Geometric Representations for Map Mashups Using Relative Positioning and Linked Data." *International Journal of Geographical Information Science* 32 (6): 1117–1137. doi:10.1080/13658816.2018.1441416.

INSPIRE. 2017. "Linking INSPIRE Data: Draft Guidelines and Pilots." https://inspire.ec.europa.eu/news/linking-inspire-data-draft-guidelines-and-pilots.

Iosifescu-Enescu, I., and L. Hurni. 2007. "Towards Cartographic Ontologies or" How Computers Learn Cartography." Proceedings 23rd International Cartographic Conference, Moscow, Russia, August 4–10.

ISO. 2012. *ISO 19117:2012 – Geographic Information – Portrayal*. Geneva: International Organization for Standardization.
ISO/TC211. 2009. *ISO/DIS 19143: Geographic Information – Filter Encoding*. Geneva: International Standards Organization.
Janowicz, K., S. Schade, A. Bröring, C. Keßler, P. Maué, and C. Stasch. 2010. "Semantic Enablement for Spatial Data Infrastructures." *Transactions in GIS* 14 (2): 111–129. doi:10.1111/j.1467-9671.2010.01186.x.
Jones, C. B., D. B. Kidner, L. Q. Luo, G. L. I. Bundy, and J. M. Ware. 1996. "Database Design for a Multi-Scale Spatial Information System." *International Journal of Geographical Information Systems* 10 (8): 901–920. doi:10.1080/02693799608902116.
Jones, J., W. Kuhn, C. Keßler, and S. Scheider. 2014. "Making the Web of Data Available Via Web Feature Services." Connecting a Digital Europe Through Location and Place, 341–361, Springer.
Karam, R., F. Favetta, R. Kilany, and R. Laurini. 2011. "Location and Cartographic Integration for Multi-providers Location Based Services." In *Advances in Cartography and GIScience*, 1 vol., 365–383. Berlin, Heidelberg: Springer.
Kavouras, M., and M. Kokla. 2008. *Theories of Geographic Concepts: Ontological Approaches to Semantic Integration*. Boca Raton, FL: CRC Press.
Keßler, C., M. Raubal, and C. Wosniok. 2009. "Semantic Rules for Context-aware Geographical Information Retrieval." European Conference on Smart Sensing and Context, 77–92, Springer, Berlin, Heidelberg.
Knublauch, H. 2011. "Spin-modeling vocabulary." W3C Member Submission. https://www.w3.org/Submission/spin-modeling/.
Lauriault, Tracey P., Barbara L. Craig, D. R. Fraser Taylor, and Peter L. Pulsifer. 2007. "Today's Data are Part of Tomorrow's Research: Archival Issues in the Sciences." *Archivaria* 64: 123–179.
Leon, A. D., F. Wisniewki, B. Villazón-Terrazas, and O. Corcho. 2012. "Map4rdf-Faceted Browser for Geospatial Datasets." Proceedings of the First Workshop on Using Open Ddata, W3C, Brussels, Belgium, June 19–20.
Lupp M. 2007. "OGC Implementation Specification 05-078r4: Styled Layer Descriptor Profile of the Web Map Service Implementation Specification." Open Geospatial Consortium. http://portal.opengeospatial.org/files/?artifact_id=22364.
MacEachren, A. M. 2004. *How Maps Work: Representation, Visualization, and Design*. New York: Guilford Press.
Müller, M. 2006. "Symbology Encoding Implementation Specification 05-077r4." Open Geospatial Consortium. http://portal.opengeospatial.org/files/?artifact_id=16700.
Nikolaou, C., K. Kyzirakos, K. Bereta, K. Dogani, S. Giannakopoulou, P. Smeros, G. Schwarz, et al. 2014. "Improving Knowledge Discovery from Synthetic Aperture Radar Images Using the Linked Open Data Cloud and Sextant." Proceedings of ESA-EUSC-JRC 2014-9th Conference on Image Information Mining Conference: The Sentinels Era, 63–66.
Perry, M., and J. Herring. 2012. "OGC GeoSPARQL – A Geographic Query Language for RDF Data." Open Geospatial Consortium. https://portal.opengeospatial.org/files/?artifact_id=47664.
Regalia, B, K. Janowicz, G. Mai, D. Varanka, and E. L. Usery. 2018. "GNIS-LD: Serving and Visualizing the Geographic Names Information System Gazetteer as Linked Data." European Semantic Web Conference, 528–540, Springer, Cham.
Schade, S., and P. Smits. 2012. "Why Linked Data Should not Lead to Next Generation SDI." Geoscience and remote sensing symposium (IGARSS), 2894–2897, IEEE.
Scheider, S., Andrea Ballatore, and R. Lemmens. 2018. "Finding and Sharing GIS Methods Based on the Questions They Answer." *International Journal of Digital Earth*. Advance online publication. doi:10.1080/17538947.2018.1470688.
Smith, R. A. 2010. "Designing a Cartographic Ontology for Use with Expert Systems." Proceedings of A Special Joint Symposium of ISORS Technical Commission IV & AutoCarto in Conjunction with ASPRS/CaGIS 2010 Fall Specialty Conference, Orlando, FL, November 15–19.
Stadler, C., J. Lehmann, K. Höffner, and S. Auer. 2012. "Linkedgeodata: A Core for a Web of Spatial Open Data." *Semantic Web* 3 (4): 333–354. doi:10.3233/SW-2011-0052.
Varanka, Dalia E., and E. Lynn Usery. 2018. "The Map as Knowledge Base." *International Journal of Cartography* 4 (2): 201–223. doi:10.1080/23729333.2017.1421004.
Vilches-Blázquez, Luis M., Boris Villazón-Terrazas, Oscar Corcho, and Asunción Gómez-Pérez. 2014. "Integrating Geographical Information in the Linked Digital Earth." *International Journal of Digital Earth* 7 (7): 554–575. doi:10.1080/17538947.2013.783127.
W3C. 2007. "RDF Data Access WG Charter." World Wide Web Consortium (W3C). https://www.w3.org/2003/12/swa/dawg-charter.
W3C. 2013. "W3C Semantic Web Activity." World Wide Web Consortium (W3C). https://www.w3.org/2001/sw/.
Wang, Q., Z. Mao, B. Wang, and L. Guo. 2017. "Knowledge Graph Embedding: A Survey of Approaches and Applications." *IEEE Transactions on Knowledge and Data Engineering* 29 (12): 2724–2743.
Zhang, X., T. Ai, J. Stoter, and X. Zhao. 2014. "Data Matching of Building Polygons at Multiple Map Scales Improved by Contextual Information and Relaxation." *ISPRS Journal of Photogrammetry and Remote Sensing* 92: 147–163. doi:10.1016/j.isprsjprs.2014.03.010.

Zhao, T., C. Zhang, and W. Li. 2017. "Adaptive and Optimized RDF Query Interface for Distributed WFS Data." *ISPRS International Journal of Geo-Information* 6 (4): 108. doi:10.3390/ijgi6040108.

Zhu, Y., A. X. Zhu, J. Song, J. Yang, M. Feng, K. Sun, J. Zhang, Z. Hou, and H. Zhao. 2017. "Multidimensional and Quantitative Interlinking Approach for Linked Geospatial Data." *International Journal of Digital Earth* 10 (9): 923–943. doi:10.1080/17538947.2016.1266041.

Paper III

Towards Knowledge-Based Geospatial Data Integration and Visualization: A Case of Visualizing Urban Bicycling Suitability

Weiming Huang[1, 3], Khashayar Kazemzadeh[2], Ali Mansourian[1], & Lars Harrie[1]

[1] GIS Centre, Department of Physical Geography and Ecosystem Science, Lund University, Lund 223 62, Sweden

[2] Transport and Roads, Department of Technology and Society, Faculty of Engineering, Lund University, Lund 221 00, Sweden

[3] Center for Spatial Studies, Department of Geography, University of California, Santa Barbara, CA 93106, USA

Corresponding author: Weiming Huang (e-mail: weiming.huang@nateko.lu.se).

Abstract

Geospatial information plays an indispensable role in various interdisciplinary and spatially informed analyses. However, the use of geospatial information often entails many semantic intricacies relating to, among other issues, data integration and visualization. For the integration of data from different domains, merely using ontologies is inadequate for handling subtle and complex semantic relations raised by the multiple representations of geospatial data, as the domains have different conceptual views for modelling the geographic space. In addition, for geospatial data visualization—one of the most predominant ways of utilizing geospatial information—semantic intricacies arise as the visualization knowledge is difficult to interpret and utilize by non-geospatial experts. In this paper, we propose a knowledge-based approach using semantic technology (coupling ontologies, semantic constraints, and semantic rules) to facilitate geospatial data integration and visualization. A traffic spatially informed study is developed as a case study: visualizing urban bicycling suitability. In the case study, we complement ontologies with semantic constraints for cross-domain data integration. In addition, we utilize ontologies and semantic rules to formalize geospatial data analysis and visualization knowledge at different abstraction levels, which enables machines to infer visualization means for geospatial data. The results demonstrate that the proposed framework can effectively handle subtle cross-domain semantic relations for data integration, and empower machines to derive satisfactory visualization results. The approach can facilitate the sharing and outreach of geospatial data and knowledge for various spatially informed studies.

Keywords

Geospatial data integration, data visualization, ontologies, semantic constraints, semantic rules

I. Introduction

Over the last decades, the massive use of geospatial information in various application areas (e.g. traffic analysis and energy simulation) has gradually revealed the indispensable role of geospatial information for interdisciplinary spatially informed research. Geospatial information is a key enabler for solving societal problems across disciplinary boundaries [1], and one of the most powerful information integrators to bridge diverse sources of information [2]. Although increasingly different types of geospatial data (e.g. authoritative and crowd-sourced geospatial data) have been generated and disseminated e.g. through the Internet, readily utilizing such data in a meaningful way still remains a challenge, especially for experts from other domains in which geospatial information is indispensable.

Today's geospatial data analysis heavily relies on data synthesis, as data from a single source usually does not suffice [3]. Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of the data [4]. Integrating geospatial information with information from other domains often entails the challenge of dissolving semantic heterogeneity [5]. Other domains, which are not geospatial *per se*, usually hold different conceptual views of the space that is delineated by geospatial data. Consequently, in different domains, the terminology varies to represent the geographic space. Such a situation induces significant difficulties for both data integration and the consumption of the integrated data.

Accomplishing semantic interoperability for geospatial data integration has been studied intensively; most commonly, ontological approaches are employed to explicitly represent and bridge the semantics in different domains or data sources (see e.g. [6]–[9]). The ontological approaches empower machines to compute the relations between concepts and properties residing in different ontologies, thereby enabling ontology-based data retrieval or transformation to (partially) achieve semantic interoperability. However, ontology-based approaches are inadequate for handling geospatial data with multiple representations, see e.g. [10]. Multiple representations are a special matching problem for geospatial data, as the concepts seem the same, but are not applied in the same way in data, due to differences in the geometric representations [11]. For example, one building object with a point geometry and another building object with a polygon geometry can both be categorized as *Building* in the ontology, but they are fundamentally different in terms of geometric representation and data usage. We have encountered this data integration problem in a case of evaluating urban bicycling suitability, an interdisciplinary study between the geospatial domain and the traffic domain.

Urban planners have been committing to improve urban infrastructure to improve its suitability for bicycling, which is environmentally friendly and beneficial for people's well-being [12]. As a result, traffic researchers have developed several indexes for evaluating transport performance and quality of bicycling experience. To this end, bicycling level of service (LOS) is a framework of quantifying bicycling performance [13]; in this framework, several different indexes have been developed. In this case, we intend to employ a network-based LOS index: the level of traffic stress (LTS) [14]. The rationale for choosing this index is that the network-based nature of this index implies that both links (segments) and nodes (junctions) are quantitatively evaluated to derive a comprehensive understanding of the network's suitability and connectivity. LTS produces four ratings ranging from LTS1 to LTS4 based on the types of network element and three key roadway attributes: (1) number of vehicle lanes; (2) speed limit; and (3) bike lane width (other factors include bike lane lockage, appearance of a centerline on the road, parallel parking, and the presence of traffic signal etc.). Table I demonstrates the means of deriving the LTS value for *mixed traffic* (a type of link in the bicycling network). For the full explanation of the LTS, see [14].

Table I. Criteria for Level of Traffic Stress in Mixed Traffic

Street Width			
Speed Limit	2-3 lanes	4-5 lanes	6+ lanes
Up to 25 mph	LTS 1 or 2	LTS 3	LTS 4
30 mph	LTS 2 or 3	LTS 4	LTS 4
35+ mph	LTS 4	LTS 4	LTS 4

Note: use lower value for streets without marked centerlines or classified as residential and with fewer than 3 lines; use higher value otherwise.

Some of the variables used for the LTS can be found in geospatial (GIS) road databases, while other variables (e.g. the appearance of a centerline on the streets, the appearance and size of median at the junctions, and the type of bicycling link (e.g. mixed traffic)) must be collected in the field. Therefore, data integration between geospatial databases and the field-collected data becomes a prerequisite. However, such data integration is not smooth, partly due to the semantic heterogeneity and different conceptual views held by the two domains. One example is the modelling of links and nodes in the network. In Sweden, the national road database (*Nationell vägdatabas*, NVDB) models the road network in two levels of detail. In the more detailed level, the road links (network element that connects two nodes and represent a homogeneous path in the network) are comprehensively delineated, including multi-direction representations and the features of physically separated bikeways; the nodes in this detailed level are modelled according to the Swedish NVDB mapping rules **[15]**, i.e. the detailed skeleton of the junctions are mapped by the vertices (points) and the links (polylines) between the vertices. In the less detailed (coarse) level, the links are modelled in a more generalized way, i.e. the lanes are aggregated and most of the dedicated bikeways are omitted; the nodes are modelled in a way that each junction is represented by a single node (point). Figure 1 illustrates an example of how a junction is modelled in two different levels of detail in the NVDB.

The multiple representations of geospatial data lead to semantic intricacies for traffic researchers. They need to integrate the records (each record represents either a junction or a road segment) in their field-collected data (spreadsheets) to geospatial features (instances). Traffic researchers need a comprehensive and detailed set of links, which corresponds to the links in the more detailed level; whereas they view the junctions in the same way as they are modelled in the less detailed level (if present). It is unintuitive for them to link a junction record to a set of points and links. That is, in the conceptualization of the road network from the traffic domain, road junctions correspond to the data modelling approach in the coarse level of geospatial data, while links correspond to the more detailed level. Therefore, this becomes a cross-detailed-level data integration task. Such difficulty in geospatial data has led to mainly two types of compromise in previous studies: either the intersections are not

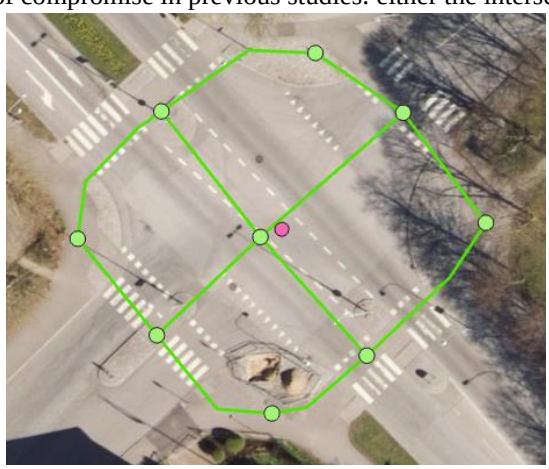


Figure 1. Illustration of different modelling approaches of a road junction in two levels of detail in the NVDB. Green links and nodes comprise the representation of this junction in the more detailed level, while the single red node represents the junction in the less detailed level. The background photograph is from ESRI's world imagery map.

explicitly represented on the map and the indexes of intersections are transferred to links [14], or the less detailed road dataset is used with manual editing, e.g. dedicated bicycling paths [16].

It is desirable to formally represent the subtle and complex semantic relations for sharing this knowledge and facilitate such cross-domain data integration missions, instead of experts from the domains having to discuss for each application of this kind. However, such cross-detailed-level semantic relations are difficult to capture merely using ontology alignment, because in most geospatial (network) ontologies, the level of detail information is modelled at dataset level (see e.g. [17]). For example, in the INSPIRE (infrastructure for spatial information in Europe) network ontology[1], the concept of *node* is defined, while the above semantic relations between the different views of link and node in two different domains can hardly be captured using, for example, Ontology Web Language (OWL)[2] restrictions. For instance, OWL is not able to express the restriction *one junction instance from field-collected data should be linked to one node instance in the less detailed road network*. Therefore, we need a method to formally represent the semantic relations for data integration, and knowledge reuse.

Furthermore, another missing piece for performing this interdisciplinary case resides in the knowledge sharing and formalization of data usage, in which semantic challenges also often arise. This study entails the engagement of multiple analysis from different domains, including how to derive the LTS index and how to appropriately visualize the processed data on the map. In particular, geospatial data visualization (geovisualization) is a knowledge-intensive art and pertains to a wide range of cartographic knowledge, in which there are abundant semantic intricacies [18]–[21]. The knowledge from the two domains is usually embedded implicitly in complex software, or in the mind of domain experts. Traditionally, experts from one side have to either refer to literature or cooperate with the experts from the other side to accomplish such work [22]. Either of these ways is prone to misunderstanding due to the semantic heterogeneity between the domains. Moreover, such an informal way of knowledge sharing impedes the wide sharing, reusing, and expansion of that knowledge. Therefore, we also need methods to formally represent the knowledge for data usage from the two domains to foster better communication and knowledge reuse.

The aim of this paper is to formalize knowledge from different domains for geospatial data integration and visualization for spatially informed studies using semantic technology. Semantic technology has been increasingly adopted in the geospatial domain [23][24], and it possesses several knowledge representation paradigms that empower us to reinforce the bridges between different domains. The approach is showcased in the visualization of urban bicycling suitability with the level-of-traffic-stress (LTS) index, in which semantic heterogeneity is a significant impediment. Specifically, we leverage ontologies, semantic rules, semantic constraints, and linked data for data integration and visualization. The knowledge for data integration, derivation of LTS, and visualization is formally represented to foster better interpretability and reusability. Overall, the contributions of this paper are:

1) A framework for cross-domain and cross-detailed-level geospatial data integration is proposed, in which ontologies and semantic constraints are leveraged to represent complex and subtle semantic relations, in order to ensure the semantic correctness of data integration.

2) A knowledge base consisting of ontologies and semantic rules is developed for formalizing the knowledge of analysis (deriving the bicycling suitability index for a

[1]https://raw.githubusercontent.com/inspire-eu-rdf/inspire-rdf-vocabularies/master/net/net.ttl
[2] https://www.w3.org/TR/owl2-overview/

road network) and visualizing data on maps, which showcases the communication of knowledge from different domains for geospatial applications.

3) The knowledge for data analysis and visualization is represented at different abstraction levels, in order to ease cross-domain knowledge communications.

4) The knowledge base for data visualization is context-aware, i.e. the visualization varies in different contexts.

Following this introduction, Section II presents an overview of the proposed approach, which is showcased in Section III–VII. Section III provides information concerning the multi-source data and the study area of the case study; Section IV and V elaborate our proposed knowledge-based approach for geospatial data integration and visualization in the case study. Section VI presents the implementation details. Section VII evaluates the proposed approach in the case study. The paper ends with a discussion (Section VIII) and conclusions (Section IX).

II. Knowledge-based geospatial data integration and visualization

This section provides an overview of the knowledge-based approach for geospatial data integration and visualization leveraging semantic technology. The approach generally comprises two main parts: data integration and data visualization.

With regard to data integration, a semantic approach is employed. First, ontologies are designed to formally represent the semantics of the data from multiple sources (in this case the semantics of geospatial data with multiple representations and the field-collected data). Ontologies are formal representations of the knowledge within a domain of interest, which are defined by the concepts in the domain and the relationships between the concepts [25]. The ontologies can either be designed from scratch and (partially) reused from state-of-the-art standardized ontologies; the latter is encouraged whenever possible [26]. In the geospatial domain, many ontologies have been designed and standardized for the purposes such as data exchange and query. For example, in Europe, the INSPIRE directive has designed several ontologies for representing geospatial data with different themes, e.g. road network [27]. Yet, for the bicycling LOS evaluation, there is no existing ontology to the best of our knowledge, thus we design the ontologies from scratch. The employed ontologies are then bridged via semantic relations from, for example, OWL, for data integration. However, for the relations that cannot be captured by semantic relations from OWL, we employ semantic constraints [28] to represent such subtle and complex relations. In the study, the complex semantic relations stem from the multiple representations of geospatial road network data. Data from different sources are then transformed to the semantic data model for linked data— Resource Description Framework (RDF) [29]—from their source data models, e.g. ESRI shapefiles for geospatial data. In order to explicitly represent the multiple representation relations of geospatial data, a multiple representation database (MRDB) is constructed before the data transformation to RDF. An MRDB organizes geospatial objects in different levels of detail, and the relations between the representations from different levels of detail are explicitly stored [30]. That is, the geospatial data in RDF have explicit relations between different representations of the geospatial objects. Then, the corresponding data instances, e.g. an intersection from the multi-scale road network and from the field-collected data, are matched. The matching relations are validated against the semantic relations represented by OWL constructs and semantic constraints. Such a knowledge-based geospatial data integration method is detailed in Section IV.

For data visualization, the knowledge is formalized firstly to transform the integrated raw data to the phenomenon that is to be visualized, and the derived phenomenon is visualized

according to the formalized visualization knowledge, i.e. how the geospatial data should be properly visualized on a map in a sense-making and cartographically satisfactory way. The knowledge for phenomenon derivation and data visualization usually stems from different domains. In our case, the knowledge concerning how to derive bicycling suitability indexes comes from traffic experts, and the knowledge for data visualization is from cartographers. The solicited knowledge is then formalized using ontologies and semantic rules [20]. With the formalized knowledge encapsulated in ontologies and semantic rules, reasoners are able to derive phenomenon values and visualization means (e.g. styles and symbols) to develop the final visualization products. This knowledge-based geospatial data visualization approach is detailed in Section V.

All the ontologies, semantic constraints, semantic rules, and source codes used in this study can be found in a GitHub repository at https://github.com/RightBank/Knowledge-based-integration-and-visualization. We are, however, not permitted to distribute the data used in the study.

III. Study area and data

In this study, we showcase our approach in evaluating and visualizing the urban bicycling network in the center-west part of Lund, Sweden. The entire transport network is evaluated using LTS, as according to Swedish traffic regulation *Trafikförordning (1998:1276)*[3], cyclists are legally allowed to ride in motor vehicle infrastructure even if a dedicated cycle path is available, unless bicycling is clearly prohibited. Therefore, we evaluate the dedicated bicycling infrastructure together with the motor vehicle infrastructure that is not prohibited for bicycling.

We utilize two main data sources: geospatial road networks (the NVDB) in two levels of detail with geometries (essential for visualization) and the information of lane numbers and speed limits, as well as the field-collected data containing other necessary information (variables) for LTS derivation. Figure 2 shows the multi-scale road network of this study area, and Figure 3 is a snapshot of field data collected by traffic researchers.

For the geospatial multi-scale road database, we create an MRDB. In the MRDB, the correspondence relations of network elements in two levels of detail are identified and stored. The relations between links in the road network are identified by the tool *Generate Rubbersheet Links*, and the relations between intersections are identified by the tool *spatial join*; both of the tools are from ESRI ArcGIS Pro 2.0.0. The identified relations (links) are thoroughly inspected to ensure their semantic correctness.

IV. Knowledge-based geospatial data integration using ontologies and semantic constraints

This section elaborates the knowledge-based data integration approach in the context of the case study. The approach is based on formal representations of data semantics and the correspondence relations between the means of representing geographic objects in the geospatial domain and the traffic domain. The ontologies and semantic constraints used in this study are all available in the GitHub repository of this paper.

[3] https://open.karnovgroup.se/transport-och-kommunikation/SFS1998-1276

Figure 2. Multi-scale geospatial road networks (NVDB) in the study area (center-west of Lund, Sweden). Left is the more detailed level of NVDB, and the right one is the less detailed level.

	A	B	C	D	F	G	H	I
1								
2		Local ID	Location	Length	Traffic Type	Speed(mph)	N Lane	Centreline
3		S-1	Steglitsvägen	308	Mixed	18	2	N
4		S-2	Steglitsvägen - Örnvägen	85	Mixed	18	2	N
5		S-3	Hökvägen (Steglitsvägen - Örnvägen)	85	Mixed	18	2	N
6		S-4	Starvägen (Steglitsvägen - Örnvägen)	85	Mixed	18	2	N
7		S-5	Örnvägen (Steglitsvägen - Örnvägen)	85	Mixed	18	2	N

Figure 3. Snapshot of the data collected in the field by traffic researchers.

A. Semantic enrichment for geospatial data

It is a common practice to leverage ontologies to formally represent the taxonomy in each domain, and use semantic relations (e.g. relations in OWL or SKOS[4] ontologies) to bridge the domains for e.g. data exchange and integration.

For geospatial data, the ontologies for representing geospatial networks and the information regarding the level of detail (cartographic scale) are necessary. In this study, we utilize the INSPIRE network ontology (*net* as prefix). This ontology defines the key concepts for geospatial networks, such as *Network*, *Link*, and *Node*. The geometric information is defined by incorporating the simple feature part of GeoSPARQL (*sf* as prefix)—a query language for geospatial linked data [31]. A *Network* instance can be associated with a number of *Link* and *Node* instances that are both *NetworkElement*, and *Link* and *Node* instances can also be connected to express the connectivity of the network. We created three subclasses of the class *Link*: *Bikeway*, *Motorway* and *CrossingLink* (the links comprise the junctions) for different types of links in the road network; and also three subclasses of the class *Node*: *Intersection*, *Roundabout* and *DeadEnd*. For the level of detail information, we partly reuse and complement the cartographic scale ontology (*scale* as prefix) from [16]; that is, the two properties of *hasMaxScaleDenominator* and *hasMinScaleDenominator* are defined to associate the datasets (geospatial networks with different levels of detail) with respective visualization scales, and the properties of *isMoreGeneralThan* and *isMoreDetailedThan* (inverse properties) are defined to represent the relations between datasets. The corresponding features identified in two levels of detail when constructing the MRDB (cf. Section III) are associated by properties in the SKOS vocabulary, i.e. using *skos:closeMatch*

[4]https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html

relation to associate the corresponding features at different levels of detail (e.g. a node in the coarse level can be matched to a set of nodes and links in the detailed level). Figure 4 demonstrates the essential parts of the ontologies for semantically organizing the multi-scale NVDB in RDF.

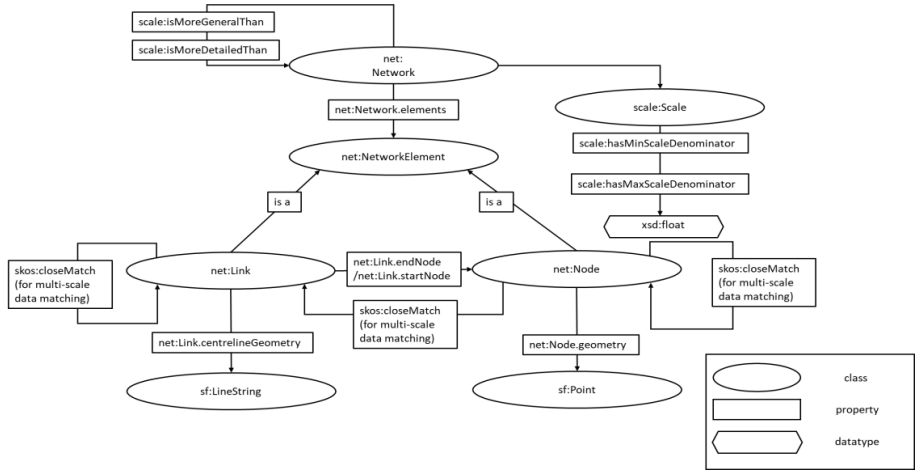


Figure 4. Core parts of the ontologies for semantically representing multi-scale geospatial networks.

B. Semantic enrichment for field-collected road data

The data collected in the field for evaluating the bicycling network based on the LTS are recorded in spreadsheets (tables). Tables are a common way to store and exchange data, e.g. on the Web, whereas most of the tables' information is only understood by humans but not machines. In fact, the tables are sometimes even difficult to understand for humans, particularly in the interdisciplinary studies such as this case where the table data need to be understood by experts from another domain. Therefore, it is important to formally and explicitly represent the semantics in the tables, so that they can be unambiguously understood by both humans and machines. This is in line with the research topic of semantic table interpretation in the semantic web domain [32]. In our study, the meaning of the table data is unclear for the geospatial experts, and this hampers the data integration task. Therefore, we develop ontologies for representing the semantics in traffic domain, especially for the data used for the LTS. Developing ontologies and building relations with geospatial ontologies can not only ease the cross-domain communication, but also facilitate the reuse and sharing of such knowledge. We develop ontologies from scratch, as no previous work has been accomplished in this regard. The ontologies are designed through several comprehensive discussions between traffic researchers and knowledge engineers, and the ontology design approach *METHONTOLOGY* is employed to build glossary, concepts, and relations [33]. The ontologies are developed in two levels: the LOS level and the LTS level (the LTS is an index in the framework of LOS that includes a series of evaluation methods). The rationale for developing the multi-tier ontologies is that we model a part of visualization knowledge at the LOS level, i.e. the cartographic rules apply to all LOS indexes, including the LTS (see Section V.A). In the upper level ontology of LOS, the common concepts and relations for deriving indexes (including the LTS) are defined, including the concepts of *LOSIndex*, *BicyclingNetwork*, *BicyclingNode*, and *BicyclingLink*, and the relations of *hasLOSIndexValue*, and *isMatchedTo* (for associating an instance of bicycling link or node to the instances of geospatial network element). The developed LTS ontology incorporates the concepts and relations used specifically for the LTS index, and is developed on the basis of the LOS ontology. Essentially, the concept *LTS* is defined as a subclass of *LOSIndex* with four instances of this class, i.e. *LTS1*, *LTS2*, *LTS3*, and *LTS4* (including rich semantics about

the four levels); the types of bicycling network elements defined in LTS are incorporated: the concepts of *BikePathWithPhysicalSeparation*, *BikeLaneWithMarking*, *MixedTraffic*, and *PocketBikeLane* are defined as subclasses of *BicyclingLink* in different abstraction levels; the concepts of *Crossing*, *SignalisedCrossing*, *UnsignalisedCrossing*, *CrossingWithMedianRefuge* are defined as subclasses of *BicyclingNode* in different abstraction levels. An object property *hasLTSValue* (this relation is inferred based on semantic rules, see Section V.A) is created as a subproperty of *hasLOSIndexValue*. Other properties needed for LTS derivation are also defined, e.g. *rightTurnLaneType*, *hasCentreline*, and *isConncetedTo* (denoting the connectivity between bicycling nodes and links). Figure 5 illustrates the concepts defined in different abstraction levels in the LOS and LTS ontologies.

C. Data integration with semantic constraints

According to the semantic relations identified and discussed in Section I:

- a bicycling link should be matched to at least one link in the detailed level geospatial road network
- a bicycling node should be matched to exactly one node in the coarse level network data if that node feature is available in the coarse level, otherwise the node should be matched to one node in the detailed level (e.g. small junctions are only present in the detailed level with single points).

Since the level of detail information is defined at dataset level (the scale is associated with *net:Network* instance that presents the entire network in one level of detail), OWL, which is often used for representing data restrictions, is not capable of representing such subtle

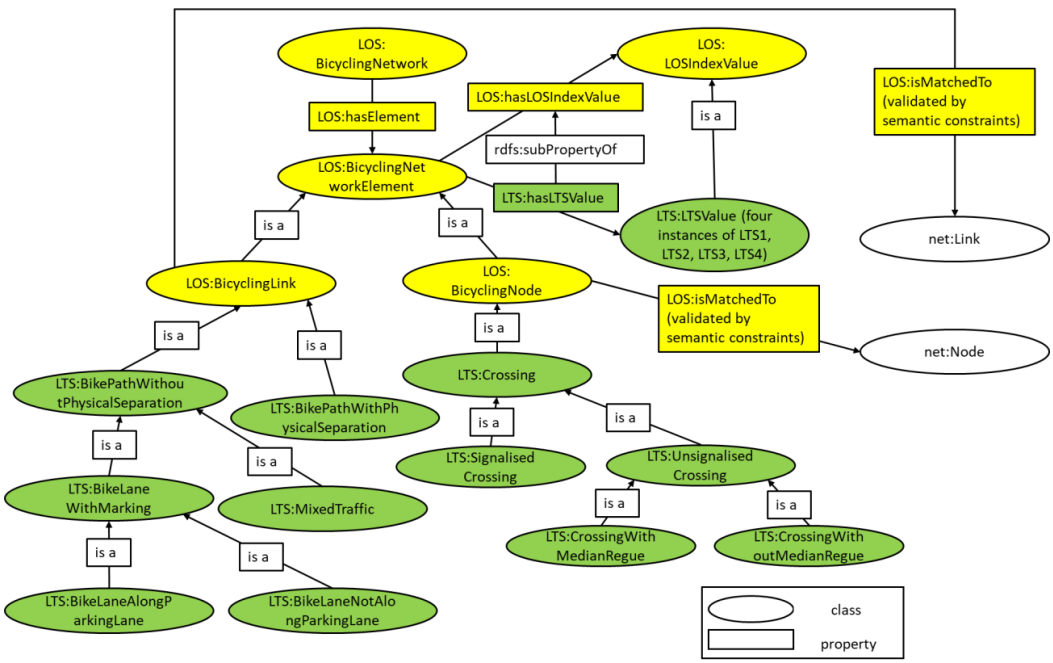


Figure 5. Core concepts and relations in the LOS and LTS ontologies, including their links with the INSPIRE network ontology. The elements in LOS ontology are with yellow color, and green color is for LTS ontology.

semantic relations and complex integrity constraints. Moreover, OWL was designed for reasoning, but not data constraints. OWL restrictions describe the reasoning to be applied based on them [34]. For example, assuming there is an *owl:maxCardinality 1* restriction stating that one bicycling node (*LOS:BicyclingNode*) can only be matched to one geospatial node (*net:Node*) feature at maximum, and there are two *net:Node* instances matched, then an OWL reasoner will assume that the two *net:Node* instances must in fact represent the same real-world entity. Furthermore, OWL adopts the *open world assumption*[5], and thus, assuming an irrelevant instance (e.g. a building instance) is mistakenly matched to a bicycling node, then the reasoner will infer that the building instance is also a *net:Node* instance. Additionally, the *owl:minCardinality* will not report any integrity error of missing values, because more data may appear at any time to satisfy that restriction under the *open world assumption*.

Due to the limitations of OWL, there have been many efforts to develop data constraints for RDF graphs, see e.g. [28] for semantic environmental data validation. In this context, the shapes constraint language (SHACL) became a W3C (World Wide Web Consortium) recommendation in 2017 [35]. The W3C recommendation made SHACL the most promising technique for becoming the *de facto* standard of semantic data constraints. Primarily, SHACL is a language for validating RDF graphs, and can also be used for other purposes including, among others, data integration. SHACL has been increasingly adopted in various domains and applications, e.g. clinical information systems, and software regression testing [36], but has been seldom used in the geospatial domain. We argue that such semantic constraints have unexplored potential for geospatial linked data, which mostly do not adopt the *open world assumption*, and there is a significant need for the integrity assurance and data integration. Such need becomes more prominent in the spatially underpinned interdisciplinary studies, in which subtle and complex semantic relations between geospatial and other domains are often inevitable. This is also in line with the opinions from [10], who identified the problem of missing semantic relations for representing concept relations for multi-source geospatial data. In this context, semantic constraints can be leveraged to handle complex and subtle semantic relations.

We employ SHACL constraints for representing the subtle semantic relations for integrating the field-collected data and multi-scale road network data, i.e. the matching relations between *link* and *node* in the two domains. Listing 1 is a SHACL constraint (*sh* as prefix for namespace of SHACL) that is used for representing subtle semantic relations between *net:Node* and *los:BicyclingNode* for data integration. This constraint assures that an instance of *los:BicyclingNode* will be matched to an instance of *net:Node* in the coarse level of detail network if available, otherwise it must be matched to a *net:Node* in the detailed geospatial network. Once the constraints are violated, SHACL will generate reports to facilitate the identification of semantic mismatches [35]. Moreover, the subtle semantic relation is formally represented thanks to the expressive SHACL semantics and the SPARQL query embedded. With the formalization of such subtle semantic relations, this interdisciplinary study is eased, as such semantic constraints can be readily reused and expanded. Simply put, the bridge between the domains is reinforced than merely using ontologies. Note that the ontologies and semantic constraints provide a semantic framework for data integration to ensure semantic correctness for data integration, and the formally represented knowledge concerning how to incorporate multi-scale geospatial data into analysis can be readily interpreted and reused. By contrast, the matching between individual data objects (e.g. matching a record from field-collected data with a geospatial feature) is not automated by this framework. In this case, object matching (integration) is performed manually depending on the road name information, as the geometric information is not recorded in the field-collected data and distance-based object matching cannot be conducted. The results of the object

[5]http://wiki.opensemanticframework.org/index.php/Overview_of_the_Open_World_Assumption

matching process are validated against the ontologies and semantic constraints to spot the semantically incorrectly matched objects. In addition, the matching is revised according to the hints given in the error reports.

```
@[prefix definitions]

ex:NodeIntegrationShape a sh:NodeShape ;

      sh:targetClass los:BicyclingNode ;

#the shape constraint is for bicycling nodes

      sh:path los:isMatchedTo ;

#the shape constraint is targeted to this relation

      sh:class net:Node;

# the object of the statement in this pattern must be

      #each bicycling node is matched to exactly one geospatial node

sh:minCount 1;    sh:maxCount 1;

      # a SPARQL constraint to represent the conditions incorporating level
      of detail information

      sh:sparql [ a sh:SPARQLConstraint;

      sh:message "a bicycling node should be matched to a node in the
      coarse level data if available." # message in case of error arises
      sh:select """

      SELECT $this (?detailed_road_node as ?value)

      WHERE { $this $PATH ?detailed_road_node .

            ?detailed_network net: Network.elements ?detailed_road_node;

            scale:isMoreDetailedThan ?coarse_network.

            ?detailed_road_node skos:closeMatch ?coarse_road_node.}""" ;] .
```

Listing 1. A SHACL constraint that states a bicycling node must be matched to one node in the geospatial network of coarse level of detail if available, otherwise it must be matched to one node in the detailed geospatial network.

V. Knowledge-based geospatial data visualization with ontologies and semantic rules

Due to the interdisciplinary nature of the case study, evaluating bicycling suitability and then visualizing the evaluation on the map entail the incorporation of knowledge from the two domains. We intend to develop knowledge bases to formally represent the knowledge from the two domains, i.e. derivation of the LTS and the map visualization. Such knowledge bases would foster better communication between the two domains, and they can be readily reused, rather than requiring domain experts to consult literature or cooperate each time.

In this study, we encapsulate the domain data analysis methods (derivation of LTS and visualization rules) using ontologies and semantic rules. Semantic rules (horn logic) are a prominent knowledge representation paradigm in the semantic web. They offer a knowledge representation model for both domain experts and developers; semantic rules are more manageable and understandable than procedural codes as they lessen the semantic gaps

between domains [37]. The developed ontologies and semantic rules are all available in the GitHub repository of this paper.

A. Knowledge base for LTS

Deriving the LTS index values for different types of bicycling links and nodes is complex, as each type of the network element has its own method for its derivation. In this study, we formally represent the LTS derivation using semantic rules to foster rule-based reasoning. Such formalized knowledge can be understood by both humans and machines, and can be reused for the calculation of this index in other use cases.

The LTS derivation is formally represented using the object-oriented SPIN (SPARQL Inferencing Notation) rules, that combine concepts from object-oriented languages, the SPARQL query language, and rule-based systems to model rules in the semantic web [38]. SPIN rules are increasingly widely used as they are expressive and close to SPARQL, and also support non-monotonic reasoning. In fact, SHACL's advanced features include semantic rules, which are an upgrade of SPIN rules, whereas such advanced features are not in the W3C recommendation, and few reasoners currently support SHACL semantic rules. Therefore, we opt to still use SPIN rules, which can be readily migrated to SHACL rules in the future if necessary[6].

As a proof-of-concept, in this study, we develop SPIN rules for a subset of LTS derivation scenarios, i.e. the bicycling network element types of *mixed traffic*, *bike path with physical separation*, *pocket lane*, and *unsignalized crossing with(out) median* that appear in the research area. The index derivation for each type is formalized into a few rules to cover all the logics, and overall 17 SPIN CONSTRUCT rules are developed to formally represent a part of the LTS derivation and enable the reasoner to infer the LTS value, i.e. infer the object property of *lts:hasLTSvalue* and associate each instance of *los:BicyclingNetworkElement* with an instance of *lts:LTSValue*.

B. Knowledge base for visualizing geospatial data

Similarly, the geovisualization (cartographic) knowledge can be formalized with ontologies and semantic rules, in order to facilitate the understanding of such knowledge in interdisciplinary studies where information needs to be visualized on maps. To empower machines to understand cartography, it is commonly acknowledged that the cartographic knowledge framework should be formally represented using ontologies [39]. The authors of [20] designed ontologies and semantic rules used for geovisualization, in which a data portrayal knowledge base comprises a major part. However, that work was for the purpose of web mapping, and did not incorporate high-level cartographic knowledge, i.e. common cartographic principles and rules. In this regard, [40] designed an ontology including many prominent cartographic concepts, e.g. cartographic method, and data types (e.g. according to measurement scale: nominal, interval, and ordinal data). We argue that although the data portrayal ontologies in [20] can explicitly represent the information of how features should be visualized under different conditions, the cartographic theories are unable to be readily utilized, which thus diminishes the automation level of knowledge modelling and representation. Therefore, in order to better leverage the cartographic theories, we complement the data portrayal ontologies in [20] with high-level cartographic knowledge. The visualization knowledge is mostly modelled with high-level cartographic concepts and relations, and then analyzed by semantic reasoning to transfer to lower-level data portrayal knowledge to render information on the maps.

For cartographic ontology, we reuse and extend the work of [39] (*carto* as prefix). We create the relation that *los:LOSIndexValue* is a subclass of *carto:OrdinalData* and

[6] http://spinrdf.org/spin-shacl.html

carto:ThematicData, and we add the concept of *carto:ColorScale* with its subclasses *HSVColorScale*, *CMYKColorScale*, and *RGBColorScale* to represent the color scales in different color systems, as the color distinction is one of the most commonly used visualization practices. The defined concepts enable cartographers to model color scales to different types of thematic data. In this study, we model an HSV (hue, saturation, value), color scale for visualizing bicycling network elements with different LOS index levels, according to cartographic knowledge for ordinal data. We use a traffic signal color scale (green, yellow, and red), as the meaning of the colors in this scale is perceptible in the traffic domain. The defined color scale starts and ends at two certain HSV colors to represent the range (thereby defining the properties *carto:startsAtColor* and *carto:endAtColor*), and the color scale instance is associate with the concept of *los:LOSIndexValue* using the property *carto:hasApplicationField* to denote the application field. Figure 6 illustrates the hierarchy and relations between the cartographic ontology, as well as the LOS and LTS ontologies.

Grounded upon the formalized concepts and relations in the ontologies, we then formalize generic cartographic rules using SPIN rules (with the prefixes of *spin* and *sp*). The color scale is evenly interpolated (one cartographic common rule) and then applied to different values of thematic ordinal data. Different values of line thickness are also applied to different types of links. The interpolation of the color scale is conducted in the three dimensions (hue, saturation, value) respectively. For real-time visualization, a portrayal rule base is created, consisting of four SPIN rules regarding using different symbolizers (basic units of visualization) under different conditions, and thus, how each feature should be portrayed on the map can be deduced using semantic reasoning (cf. [20]). Listing 2 shows the symbolizers used for independent bikeways. In this rule, the color used for the thematic value (LTS value) is from the interpolation of the color scale modelled in the cartographic ontology. The interpolation of color scales for ordinal data (a type of thematic data) is formalized in a SPIN rule and derives the correspondence relations between each thematic value and color (with the property *carto:colorCorrespondsToThematicValue*). The interpolated colors modelled in the cartographic ontology are then transferred to a data portrayal rule (according to the data portrayal ontologies in [20]) in Listing 2 for assigning symbolizers to geospatial objects.

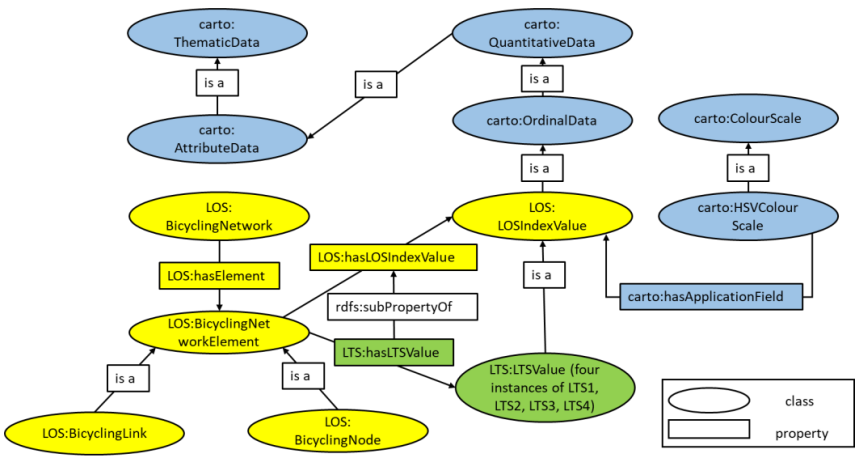


Figure 6. Core concepts and relations in the cartographic ontology and its relations with LOS and LTS ontologies. The cartographic ontology is annotated with blue color, the LOS ontology is with yellow, and

```
@ [prefix definitions]

net: Link a owl:Class;

      spin:rule{

      a sp:Construct;

      style: representsPortrayalRule [portrayal_rule_URI]

      sp:text"""

      CONSTRUCT{ symbol:LOSSymbol symboliser:hasSymboliser ?symboliser.

                        ?symboliser a symboliser:LineSymboliser;

                              graphic:strokeColour ?colour;

                              graphic:strokeWidth 3.

                        ?this symboliser:isSymbolisedBy ?symboliser.}

      WHERE{

      ?bike_path a los:BicyclingLink;
                        los:isMatchedTo ?this.

      ?this a net:Bikeway;

      ?bike_path los:hasLOSIndexValue ?LOSValue.

      ?colour carto:colourCorrespondsToThematicValue ?LOSValue.

# carto:colourCorrespondsToThematicValue was inferred during colour scale
interpolation
      BIND(IRI(CONCAT("urn:symboliser_for_independent_bikeway_",STR(?this))
) as ?symboliser)  }""".
```

Listing 2. Example SPIN rule for assigning symbolisers for the independent bikeways.

C. Abstraction levels of data usage knowledge

As described above, the knowledge concerning data usage, i.e. the derivation of LTS and its visualization is formalized. In this process, we create three knowledge representation abstraction levels, and different types of data usage knowledge are modelled at different levels. The three levels are: (1) cartographic common knowledge; (2) visualization knowledge for the LOS (theoretically it can cover all kinds of LOS indexes); and (3) the particular LTS index level.

At the cartographic common knowledge level, the core concepts and relations of cartographic theories are modelled in the cartographic ontology. In principle many rules can be modelled at this level (e.g. cartographic rules for ordinal data) using semantic rules. In this study we showcase this by developing a rule of color scale interpolation at this level, i.e. the color scale is interpolated evenly according to the number of the ordinal thematic data (LTS1–4 in this case). As the subclass inheritance is formally defined in the ontologies (see Figure 6), all the semantic rules modelled at this level also apply to lower level leveraging ontological reasoning, i.e. the object-oriented SPIN rules modelled in the *carto:OrdinalData* level also apply to lower level concepts of *los:LOSIndexValue*, and thereby, *lts:LTSValue*.

At the second knowledge abstraction level—LOS level—we model all the application-specific visualization knowledge. An instance of *carto:ColorScale* (a color scale that fits the traffic phenomena visualization) is created with the application field of *los:LOSIndexValue*. After the color for each index value is retrieved through color scale interpolation, the semantic rules assign different colors to different LOS index values. For example, in the semantic rule in Listing 2, the property *los:hasLOSIndexValue* is used, which has a subproperty of *lts:hasLTSValue*; each link or node is associated with an LTS value through the property *lts:hasLTSValue*, thereby the upper level property *los:hasLOSIndexValue* can be used to retrieve the LTS values. In addition, all the line thickness rules are formalized at this level.

The derivation of the LTS is formally represented at the lowest level—the LTS abstraction level—to deduce the associations between the bicycling network elements and the LTS values ranging from LTS1 to LTS4 through *lts:hasLTSValue*. No visualization knowledge is defined at this level; they are instead transferred from upper abstraction levels.

The rationale of modelling the knowledge of data usage (analysis and visualization) at different abstraction levels is that, we believe it is unrealistic for cartographers to model how the data should be visualized for every single application, rather the applications can be aggregated to ease such knowledge representation work. The knowledge modelled at the LOS level can be used for every LOS index, as long as the subclass inheritance is explicitly represented. In this case, such knowledge transfer from upper level to lower level is showcased with the LTS—a particular LOS index. Thanks to the semantic reasoning capabilities, every semantic rule modelled at the upper level also apply to lower levels, therefore the number of knowledge abstraction levels can be increased if necessary, e.g. by adding another abstraction level of *traffic thematic data*, of which *los:LOSIndexValue* is a direct subclass.

D. C*ontext-awareness of data usage*

In principle, the data used in this case can be used for different analyses, e.g. the multi-scale geospatial road network data can be also used for traffic congestion analysis, in addition to the bicycling suitability analysis in this study. Therefore, the context information is crucial for spatially informed studies, informing the knowledge bases of the data usage contexts. Semantics plays a pivotal rule for modelling the context information [41]. The knowledge-based approach unlocks the opportunity of context-aware geospatial data visualization, i.e. the analysis or visualization method can vary according to the data usage context.

In this study, the ontologies and semantic rules are context-aware. The visualization context is transferred to the knowledge base from the client side, and the context information thereafter is involved in the semantic reasoning to deduce appropriate analysis and visualization means for the current context. Therefore, we create a light-weight visualization context ontology with the class of *VisualizationContext*; a *VisualizationContext* instance can be associated with a *carto:Phenomenon* instance through the property *visualizesPhenomenon*. In this case, *los:LOSIndexValue* is assigned as a subclass of *carto:Phenomenon*. The context information in the knowledge base can be updated according to the information transferred from the visualization client. In this case, the visualization context (*VisualizationContext*) instance visualizes (*visualizesPhenomenon*) the thematic data of *lts:LTSValue*. Then the rule-based reasoning deduces that the rules of LTS derivation and the visualization knowledge for LTS should be used (the color scale and rules for LOS).

With our knowledge-based approach, it becomes possible that a number of different knowledge bases for different contexts co-exist, and the context data is used for invoking the appropriate knowledge bases (ontologies and semantic rules) for data consumption and visualization.

VI. Implementation

A. Data transformation

An MRDB is created based on the data from the multi-scale NVDB (see Section III). The data (originally in ESRI shapefiles) are transformed to RDF according to the INSPIRE network ontology with the correspondence relations of the features in two levels of detail (*skos: closeMatch*). The scale information is added at the network (dataset) level to denote the visualization scales and the level of detail information of the networks. The field-collected data recorded in spreadsheets are also transformed to RDF according to the LTS ontology (see Section IV.B). The data transformations are performed using R2RML[7] transformation supported by Ontop[8].

B. Cross-domain data matching

This step interlinks the road network element objects in MRDB with the bicycling link or node objects collected in the field using the relation *los:isMatchedTo*. The data matching is empowered by semantic constraints to tackle the subtle semantics of geospatial data raised by multiple representations. Two semantic constraints (in SHACL) are developed, of which one is for nodes, and the other is for links (see Section IV.C). The correspondence relations (*los:isMatchedTo*) are identified manually depending on the road name information and validated against the SHACL constraints using the Jena[9] and SHACL API[10] in a Java environment. After this step, the data from the two domains are matched in a semantically correct way. All the RDF data (including MRDB, field-collected data and the cross-dataset links) are imported to the RDF store RDF4J[11].

C. Enabling rule-based inference for data visualization

In this study, we formally represent the knowledge concerning data usage, i.e. we use ontologies and semantic rules to derive LTS values as the evaluation metric and thereafter derive cartographically satisfactory visualization methods for the bicycling network depending on the LTS values. The semantic rules are developed by writing the domain logic into SPIN rules manually. The rules are also imported into RDF4J, which has the rule-based inference capacity. The LTS values and visualization means for the data objects are inferred over the data with the combination of ontological reasoning and rule-based reasoning.

C. Visualization tool

The results are visualized in a web-based environment and with a client/server architecture. A server is implemented using the Python web framework Django[12] to communicate with the knowledge base (data, ontologies, and semantic rules in RDF4J). The server sends SPARQL queries to the knowledge base, in which it asks the knowledge base to send all the geospatial objects (in the detailed road network) and the visualization means (symbolizer) of each object to the server. The server then parses the retrieved data (e.g. fetches the CSS values associated with the symbolizers) and wraps the data into JSON objects. The JSON objects are then sent to the frontend developed mainly using the web mapping library Leaflet[13]. The frontend (browser) parses the received data and visualizes the bicycling network according to the visualization means (CSS values) encapsulated in the JSON objects. In order to enable the users to interactively understand the visualization, one could click the bicycling network

[7] https://www.w3.org/TR/r2rml/
[8] https://ontop.inf.unibz.it/
[9] https://jena.apache.org/index.html
[10] https://github.com/TopQuadrant/shacl
[11] http://rdf4j.org/
[12] https://www.djangoproject.com/
[13] https://leafletjs.com/

elements and obtain further information (e.g. LTS value, and element type) in the popped-up RDF4J faceted browser. The users could also explore the knowledge base as every data object is dereferenceable in the faceted browser using URIs. The source code of the web-based visualization tool is available in the GitHub repository.

VII. Evaluation

In this paper, we propose a knowledge-based approach with semantic technology (coupling ontologies, semantic constraints, and semantic rules) for geospatial data integration and visualization. The approach is used to solve a real-world geospatial data application—visualizing urban bicycling suitability—where data integration and visualization encounter complex and subtle domain semantics. In this context, we present a workflow for this interdisciplinary spatially informed study, including data integration, analysis, and visualization. We design several knowledge bases to cover all the aspects. Therefore, this approach is evaluated by the visualization results, which is a sink where semantics of the activities of data integration and processing are aggregated, interpreted, and visualized in a meaningful way [42].

Figure 7 is the visualization of the bicycling suitability (LTS) in the study area. The base map is a redistribution of OpenStreetMap (OSM) fed from the Mapbox API14. In the visualization

Figure 7. Visualization of LTS values on the map in the study area. The base map is a redistribution of OSM from

[14] https://www.mapbox.com/

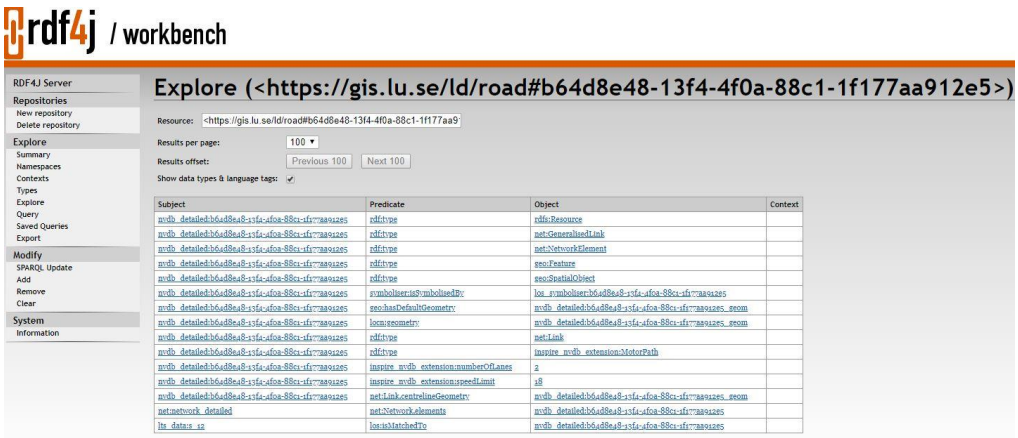


Figure 8. Information of a road link instance in the detailed level of road network in the RDF4J faceted browser.

application, once a user clicks on an object, a pop-up could guide the user to explore further information in the faceted browser of RDF4J. Figure 8 shows the faceted browser with rich semantic information of a road link instance in the detailed level NVDB, and the inferred RDF statements (from ontological and rule-based reasoning) are also included, e.g. the relation of *isSymbolizedBy* that is deduced according to the ontologies and semantic rules.

It can be observed that with our approach, the cross-domain data integration and visualization is accomplished. With the constructed knowledge base for data analysis (LTS derivation) and visualization, essentially the client application (visualization tool) asks a question to the backend knowledge base (RDF4J in this case), "in this visualization context, what are the geospatial objects that should be rendered on the map and their visualization methods?", the knowledge base will then provide the question with answers derived from the represented knowledge for data analysis and visualization. Furthermore, all the objects can be dereferenced, and more comprehensive information can be obtained (cf. Figure 8), which provides users with information beyond the graphic visualization. Such an application is difficult to develop with traditional Web mapping techniques.

VIII. Discussion

In this paper, we propose a knowledge-based approach for geospatial data integration and visualization using semantic technology. We illustrate our approach in an interdisciplinary research application—visualizing urban bicycling suitability. In our study, we had many discussions between traffic experts and geospatial experts. These discussions evidently unveiled that, in spite of massive use of geospatial information for decades in various areas, geospatial information still entails many intricacies for experts from other domains. Multiple representations of geospatial data seemed one of the most puzzling geospatial theories to traffic researchers in this study. We initially planned to perform extensive discussions, so that either geospatial experts would understand the traffic theories and help them integrate the data and develop visualization products, or traffic researchers would (partially) grasp how the tasks should be accomplished. In this scenario, bespoke solutions could be developed, and, most likely, sufficient visualization products could be produced. Nevertheless, such a type of solutions has an intrinsic demerit: the knowledge communication emerging from the discussions and the theories embedded in the developed solutions (procedural codes) can hardly be transferred, interpreted, reused, and potentially expanded.

With our approach, the domain knowledge is formalized in a semantically-enriched and machine-readable manner. In principle, if one agrees with the modelled knowledge, the knowledge base can be readily used in relevant tasks (e.g. deriving LTS values and visualization for another study area, or deriving other types of LOS indexes) instead of domain experts having to sit down together each time or traffic experts having to consult geospatial literature to find appropriate methods. We argue that the knowledge-based approach would benefit the outreach and sharing of geospatial knowledge with a wider audience. This is in line with the research with regard to knowledge sharing using ontologies [41], whereas our approach enriches pure ontological approaches with semantic constraints and rules to cope with the complex semantic landscape in interdisciplinary studies. We believe our approach can be used in many other spatially informed studies in addition to the demonstrated case, as the approach offers a general framework for geospatial data integration and visualization with semantic technology, particularly for handling multiple representations of geospatial data, which is an intricacy for data integration. The approach can also be used in other cross-domain and cross-detailed-level data integration tasks. For example, in spatiotemporal data integration, the events must be linked to the geospatial objects in a certain period of time, i.e. a geospatial object can have multiple (temporal) representations, and each corresponds to a certain period; the events should be linked to the corresponding representations in the time dimension. Another example is that during an emergency, the information of e.g. air pollution caused by fire is produced and should be linked to aggregated levels (e.g. county level), and some information is available at individual level (e.g. heritage building information); in order to analyze the affected heritage buildings during an emergency, cross-detailed-level data integration is necessary. In such cases, semantic constraints can be employed, as ontologies can hardly represent such restrictions.

Nevertheless, our approach also unveils several challenges. One prominent challenge is the modelling of knowledge, which is a demanding task. Generally, it is easier to train a domain expert with knowledge modelling (representation) than to equip a data scientist with domain knowledge [7]. This is in line with the extensively studied research topic in artificial intelligence, that is, knowledge elicitation (see e.g. [44]), which plays a significant role in expert system development. A recent survey of geospatial expert systems demonstrated that the role of niched and standalone expert systems was downgraded, while the knowledge modelled for making integrated and complex spatial decisions clearly remains imperative [45]. The semantic landscape is increasingly complex, as more diverse sources of data are becoming available for geospatial analysis and visualization. Thus, the knowledge modelled for data integration and usage (visualization) will play a pivotal role to enhance the usability of geospatial information.

One may argue that quite a few semantic technologies are employed in our approach (e.g. ontology, semantic constraint, semantic rule, and linked data), which might be confusing for users. In fact, they are different types of knowledge representation paradigms that facilitate domain experts to formalize knowledge and thus make it explicit. Ontologies are the core in our approach, representing the essential conceptualizations of the domains. Built on the ontologies, semantic constraints and rules can be modelled, to represent more complex semantics, and to derive new facts (e.g. index values and visualization means). Therefore, once the employed ontologies have been decided upon, semantic constraints and rules can be readily developed by domain experts and grounded on the ontologies. In this way, domain experts are able to work with their own domain knowledge, rather than writing programs with procedure codes [20]. However, this approach does not apply to all applications due to the limited expressiveness of the knowledge representation paradigms of semantic technology. There are certainly some analyses or visualization methods that cannot be formalized with ontologies and semantic rules. Nevertheless, it is possible to encapsulate a process that cannot be formalized with semantic technology in, for example, a program, and semantically

annotate its input and output to fit such processes into our knowledge-based approach. This method could be further investigated as a future work.

Another lesson learnt from this study is that the ontologies available for geospatial data have developed considerably, and the trend will most likely remain in the coming years. In this study, we reuse a number of state-of-the-art ontologies, e.g. cartographic scale ontology [16], data portrayal ontologies [20], and INSPIRE network ontology [27]. We acknowledge that such previous works provide solid ground for our work, and we also believe such ontology design works will benefit the outreach of geospatial information in the long run.

The contributions of this work can also be viewed from a semantic web perspective. It has been long discussed and argued that as the open data has proliferated, the data available on the semantic web has increased dramatically in recent years, including geospatial data [46]. However, the representation of knowledge concerning how these data should be used is still sparse. This work advances the modelling of geospatial data and knowledge on the semantic web. The designed knowledge bases for geospatial data integration and visualization can be readily reused, and reached to wide audience.

IX. Conclusions

This article proposes a knowledge-based approach for geospatial data integration and visualization with semantic technology. Compared to other ontology-based approaches for (partially) accomplishing semantic interoperability for data integration, we reinforce the semantic bridge between the data from different domains using semantic constraints (SHACL constraints) to cope with complex semantic relations raised by multiple representations of geospatial data. In addition, we leverage semantic rules for modelling domain knowledge (analysis and visualization means) at different abstraction levels to enable machines to deduce the desired analysis results and visualization methods. The proposed framework is showcased and evaluated in a case study of visualizing urban bicycling suitability with the LTS index, in a study area in Lund, Sweden. The case study illustrates that the knowledge-based approach successfully overcomes semantic heterogeneity for cross-domain data integration with subtle and complex semantic relations. In addition, the knowledge modelled for data analysis as well as visualization effectively empowers machines to derive desired outcomes. This work provides a methodological framework for the sharing and outreach of geospatial data and knowledge to a wider audience for interdisciplinary spatially informed studies.

References

[1]. W. Kuhn, "Core concepts of spatial information for transdisciplinary research," *International Journal of Geographical Information Science,* vol. 26, pp. 2267-2276, 2012.

[2]. K. Janowicz, "The role of space and time for knowledge organization on the semantic web," *Semantic Web,* vol. 1, pp. 25-32, 2010.

[3]. K. Janowicz, F. Van Harmelen, J. A. Hendler, and P. Hitzler, "Why the data train needs semantic rails," *AI Magazine,* vol. 36, pp. 5-14, 2014.

[4]. M. Lenzerini, "Data integration: A theoretical perspective," in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002, pp. 233-246.

[5]. K. Janowicz, S. Scheider, T. Pehle, and G. Hart, "Geospatial semantics and linked spatiotemporal data–Past, present, and future," *Semantic Web,* vol. 3, pp. 321-332, 2012.

[6]. H.T. Uitermark, P. J. van Oosterom, N. J. Mars, and M. Molenaar, "Ontology-based geographic data set integration," in *International Workshop on Spatio-Temporal Database Management*, 1999, pp. 60-78.

[7]. Y. Chen, S. Sabri, A. Rajabifard, and M. E. Agunbiade, "An ontology-based spatial data harmonisation for urban analytics," *Computers, Environment and Urban Systems*, 2018.
[8]. Y. Shu, D. Ratcliffe, M. Compton, G. Squire, and K. Taylor, "A semantic approach to data translation: A case study of environmental observations data," *Knowledge-Based Systems*, vol. 75, pp. 104-123, 2015.
[9]. M. Lutz, J. Sprado, E. Klien, C. Schubert, and I. Christ, "Overcoming semantic heterogeneity in spatial data infrastructures," *Computers & Geosciences*, vol. 35, pp. 739-752, 2009.
[10]. L. van den Brink, P. Janssen, W. Quak, and J. Stoter, "Towards a high level of semantic harmonisation in the geospatial domain," *Computers, Environment and Urban Systems*, vol. 62, pp. 233-242, 2017.
[11]. S. Volz, "Data-driven matching of geospatial schemas," in *International Conference on Spatial Information Theory*, 2005, pp. 115-132.
[12]. J. R. Pucher and R. Buehler, *City cycling* vol. 11: MIT Press Cambridge, MA, 2012.
[13]. H. C. Manual, "Highway capacity manual," *Washington, DC*, vol. 2, 2000. ISBN 0-309-06681-6.
[14]. M. C. Mekuria, P. G. Furth, and H. Nixon, "Low-stress bicycling and network connectivity," Mineta Transportation Institute, San José, USA, May 2012.
[15]. L. Petterson, “Regler för insamling och leverans av vägdata,” Trafikverket, Sweden, 2012. [Online] Avaiable at http://www.nvdb.se/globalassets/upload/styrande-och-vagledande-dokument/tdok-2013-0381-regler-for-insamling-och-leverans-ver-9.pdf
[16]. R. Pritchard, Y. Frøyen, and B. Snizek, "Bicycle Level of Service for Route Choice—A GIS Evaluation of Four Existing Indicators with Empirical Data," *ISPRS International Journal of Geo-Information*, vol. 8, p. 214, 2019.
[17]. D. Carral, S. Scheider, K. Janowicz, C. Vardeman, A. A. Krisnadhi, and P. Hitzler, "An ontology design pattern for cartographic map scaling," in *Extended Semantic Web Conference*, 2013, pp. 76-93.
[18]. L. Harrie and R. Weibel, "Modelling the overall process of generalisation," *Generalisation of geographic information: cartographic modelling and applications*, pp. 67-87, 2007.
[19]. S. Scheider and M. D. Huisjes, "Distinguishing extensive and intensive properties for meaningful geocomputation and mapping," *International Journal of Geographical Information Science*, vol. 33, pp. 28-54, 2019.
[20]. W. Huang and L. Harrie, "Towards knowledge-based geovisualization using Semantic Web technologies: a knowledge representation approach coupling ontologies and rules," *International Journal of Digital Earth*, Advance online publication, 2019.
[21]. W. Huang, A. Mansourian, E. Abdolmajidi, H. Xu, and L. Harrie, "Synchronising geometric representations for map mashups using relative positioning and Linked Data," *International Journal of Geographical Information Science*, vol. 32, pp. 1117-1137, 2018.
[22]. D. Callister and M. Lowry, "Tools and strategies for wide-scale bicycle level-of-service analysis," *Journal of Urban Planning and Development*, vol. 139, pp. 250-257, 2013.
[23]. S. Wiemann and L. Bernard, "Spatial data fusion in spatial data infrastructures using linked data," *International Journal of Geographical Information Science*, vol. 30, pp. 613-636, 2016.
[24]. W. Huang, S. A. Raza, O. Mirzov, and L. Harrie, "Assessment and Benchmarking of Spatially Enabled RDF Stores for the Next Generation of Spatial

Data Infrastructure," *ISPRS International Journal of Geo-Information*, vol. 8, p. 310, 2019.
[25]. F. Baader, D. Calvanese, D. McGuinness, P. Patel-Schneider, and D. Nardi, *The description logic handbook: Theory, implementation and applications*: Cambridge university press, 2003.
[26]. N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," ed: Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, Stanford, CA, 2001.
[27]. European Commission. Guidelines for the RDF encoding of spatial data, Technical specification, 2017. Available at: http://inspire-eu-rdf.github.io/inspire-rdf-guidelines.
[28]. Y. Shu, "A practical approach to modelling and validating integrity constraints in the Semantic Web," *Knowledge-Based Systems*, vol. 153, pp. 29-39, 2018.
[29]. F. Manola, E. Miller, and B. McBride. "RDF primer". W3C recommendation, 2004. [Online] Available at: https://www.w3.org/TR/rdf-primer/
[30]. C. B. Jones, D. B. Kidner, L. Luo, G. L. Bundy, and J. M. Ware, "Database design for a multi-scale spatial information system," *International Journal of Geographical Information Systems*, vol. 10, pp. 901-920, 1996.
[31]. M. Perry and J. Herring, "OGC GeoSPARQL-A geographic query language for RDF data," *OGC Implementation Standard. Sept*, 2012. [Online] Available at: https://portal.opengeospatial.org/files/?artifact_id=47664
[32]. Z. Zhang, "Towards efficient and effective semantic table interpretation," in *International Semantic Web Conference*, 2014, pp. 487-502.
[33]. M. Fernández-López, A. Gómez-Pérez, and N. Juristo, "Methontology: from ontological art towards ontological engineering," 1997.
[34]. H Knublauch. "SHACL and OWL Compared". Spinrdf.org. Available at: http://spinrdf.org/shacl-and-owl.html [Accessed 24 Feb. 2019].
[35]. H Knublauch, and A Ryman. "Shapes constraint language (SHACL)". W3C recommendation, 2017. [Online]. Available at: https://www.w3.org/TR/shacl/
[36]. S Steyskal, and K Coyle. "SHACL Use Cases and Requirements". W3c working group note, 2017. [Online]. Available at: https://www.w3.org/TR/shacl-ucr/
[37]. N. Bassiliades, "SWRL2SPIN: A tool for transforming SWRL rule bases in OWL ontologies to object-oriented SPIN rules," *arXiv preprint arXiv:1801.09061*, 2018.
[38]. H Knublauch. "Spin-modeling vocabulary". W3C Member Submission, 2011. [Online]. Available at: https://www.w3.org/Submission/spin-modeling/
[39]. R. A. Smith, "Designing a cartographic ontology for use with expert systems," in A special joint symposium of ISPRS Technical Commission IV & AutoCarto in conjuction with ASPRS/CaGIS, 2010
[40]. J. Brus, D. Zdena, J. Kanok, and V. Pechanec, "Design of intelligent system in cartography," in Roedunet International Conference (RoEduNet), 2010 9th, 2010, pp. 112-117. IEEE.
[41]. C. Keßler, M. Raubal, and C. Wosniok, "Semantic rules for context-aware geographical information retrieval," in European Conference on Smart Sensing and Context, 2009, pp. 77-92.
[42]. K. Janowicz, S. Schade, A. Bröring, C. Keßler, P. Maué, and C. Stasch, "Semantic enablement for spatial data infrastructures," *Transactions in GIS*, vol. 14, pp. 111-129, 2010.
[43]. M. Jelokhani-Niaraki, "Knowledge sharing in Web-based collaborative multicriteria spatial decision analysis: An ontology-based multi-agent approach," *Computers, Environment and Urban Systems*, 2018.

[44]. N. J. Cooke, "Varieties of knowledge elicitation techniques," *International Journal of Human-Computer Studies,* vol. 41, pp. 801-849, 1994.
[45]. D. Demetriou, "A review of spatial expert systems: Do they still have a role to play?," in *Sixth International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2018)*, 2018, p. 107730H.
[46]. L. van den Brink, P. Barnaghi, J. Tandy, G. Atemezing, R. Atkinson, B. Cochrane, *et al.*, "Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web," *Semantic Web,* vol. 10, pp. 95-114, 2019.

Paper IV

International Journal of Geo-Information

MDPI

Article

Assessment and Benchmarking of Spatially Enabled RDF Stores for the Next Generation of Spatial Data Infrastructure

Weiming Huang [1,*], Syed Amir Raza [1], Oleg Mirzov [1,2] and Lars Harrie [1,2]

1 Department of Physical Geography and Ecosystem Science, Lund University, 223 62 Lund, Sweden
2 ICOS Carbon Portal, Lund University, 223 62 Lund, Sweden
* Correspondence: weiming.huang@nateko.lu.se

Received: 27 May 2019; Accepted: 12 July 2019; Published: 19 July 2019

Abstract: Geospatial information is indispensable for various real-world applications and is thus a prominent part of today's data science landscape. Geospatial data is primarily maintained and disseminated through spatial data infrastructures (SDIs). However, current SDIs are facing challenges in terms of data integration and semantic heterogeneity because of their partially siloed data organization. In this context, linked data provides a promising means to unravel these challenges, and it is seen as one of the key factors moving SDIs toward the next generation. In this study, we investigate the technical environment of the support for geospatial linked data by assessing and benchmarking some popular and well-known spatially enabled RDF stores (RDF4J, GeoSPARQL-Jena, Virtuoso, Stardog, and GraphDB), with a focus on GeoSPARQL compliance and query performance. The tests were performed in two different scenarios. In the first scenario, geospatial data forms a part of a large-scale data infrastructure and is integrated with other types of data. In this scenario, we used ICOS Carbon Portal's metadata—a real-world Earth Science linked data infrastructure. In the second scenario, we benchmarked the RDF stores in a dedicated SDI environment that contains purely geospatial data, and we used geospatial datasets with both crowd-sourced and authoritative data (the same test data used in a previous benchmark study, the Geographica benchmark). The assessment and benchmarking results demonstrate that the GeoSPARQL compliance of the RDF stores has encouragingly advanced in the last several years. The query performances are generally acceptable, and spatial indexing is imperative when handling a large number of geospatial objects. Nevertheless, query correctness remains a challenge for cross-database interoperability. In conclusion, the results indicate that the spatial capacity of the RDF stores has become increasingly mature, which could benefit the development of future SDIs.

Keywords: linked data benchmark; RDF stores; geospatial data; GeoSPARQL; spatial data infrastructure

1. Introduction

Geospatial information is indispensable for spatially informed decision-making and analyses and is thereby a prominent part of today's data science landscape. Significant progress in geospatial data availability and sharing has been achieved as a result of the development of spatial data infrastructures (SDIs) that aim to make geospatial data available for the benefit of the economy and the society [1]. In Europe, the INSPIRE directive—a legal framework and standardization body for SDI development—sets the data specifications, and it mandates its member states to provide data mainly using Open Geospatial Consortium (OGC) web services [2].

Despite the significant progress, SDIs still face a number of limitations, especially in terms of discovery, reuse, and integration of the data. SDIs have partially achieved dissolving environmental

and geospatial data held in silos, but the data is still largely isolated from other information domains [3]. For example, the OGC web features service (WFS) can make geospatial data available through its data query protocol, yet such data cannot be discovered by search engines or, more importantly, linked by other data resources. This makes the data lying in the so-called deep web [4].

Today's geospatial data is available and used not only in dedicated SDIs but also in various general data infrastructures/projects that are not dedicated to geospatial data. One open data example is the general-purpose knowledge graph DBpedia (https://wiki.dbpedia.org/), which has a large number of geospatial objects. In other words, geospatial data has become a part of today's big data landscape; thus, siloed data management and delivery should be revisited [5]. This is also in line with the development and vision of open SDIs, which highlight the integration and harmonization with other data [6].

Another significant issue in SDIs is semantic heterogeneity, which is an impediment to integrating multi-source geospatial data and fusing geospatial data with other types of data, as the semantics of metadata, schemas, and data content are not usually harmonized for multi-source geospatial data or with other types of data [7].

Semantic Web technologies, particularly the parts relevant to linked data, provide a promising way to resolve the aforementioned limitations. Linked data is built around a set of data publishing best practices and facilitates data access, interlinking, and integration on the web. A recent survey conducted in 2018 by EuroSDR demonstrated that linked data is seen as one of the most important research issues and key factors moving SDIs toward the next generation [8]. Linked data was also voted one of the most important SDI research topics during the AGILE 2018 workshop 'SDI research and strategies towards 2030' [9]. An increasing amount of geospatial data has been delivered as linked data on the web and has become part of the linked open data (LOD) cloud (https://lod-cloud.net/).

Linked data is organized in the data model Resource Description Framework (RDF) [10], which is a generic graph-based data model that describes entities and relations. Linked data is also built upon formally defined ontologies, providing the means to define the concepts and relations in data, in order to make explicit any underlying assumptions regarding the data, and make it easier to understand and reuse the data. In practice, linked data needs to be managed, stored, and delivered by utilizing RDF stores (also known as triplestores), which are databases for storing and retrieving RDF data (linked data) through semantic queries (SPARQL queries [11]). The OGC extended SPARQL to develop the query language for geospatial linked data—GeoSPARQL, which comprises a lightweight vocabulary to represent and query geospatial data [12]. The number of spatially enabled RDF stores (RDF stores that handle geospatial queries) is currently growing, and their compliance with GeoSPARQL has progressed. Therefore, there is a need to survey the status of spatially enabled RDF stores in terms of both geospatial query performance and GeoSPARQL compliance.

The aim of this study is to assess and benchmark several well-known and popular spatially enabled RDF stores for potential use in future SDIs and the geospatial linked data community at large (see supplementary files). In this context, we performed benchmarking in two different scenarios in future SDIs. The first scenario is one in which geospatial data plays an important role in and constitutes a part of a large data infrastructure; here, the focus is on the integration of geospatial data with other data. Two issues must be resolved here: the ontology of the geospatial components of the data should conform to the GeoSPARQL standard, and the RDF stores should be able to efficiently perform geospatial queries on a large volume of data that is a mixture of geospatial and other data. To evaluate the first scenario, we used data from the Integrated Carbon Observation System (ICOS) carbon portal (ICOS CP) [13]—a large-scale Earth Science scientific data infrastructure. The second scenario illustrates a dedicated SDI with purely spatial data; for this case, we used test datasets from Geographica, a previous geospatial benchmark for RDF stores [14]. These datasets include crowd-sourced (e.g., GeoNames, DBpedia, and LinkedGeoData) and authoritative geospatial data.

Following this introduction, the background and related work are presented in Section 2. The data used in this study is illustrated in Section 3, including the ICOS CP's ontology design. Section 4 describes the assessment and benchmarking methodology, and the results are presented in Section 5 (for

qualitative evaluation) and Section 6 (for quantitative evaluation). The paper ends with a discussion (Section 7) and conclusions (Section 8).

2. Background and Related Work

2.1. Geospatial Semantic Web and Linked Data

The Semantic Web is a common framework that allows data to be shared and reused across application, enterprise, and community boundaries [15]. In order to make the Semantic Web a reality, it is important to make a huge amount of data on the web available with recommended best practices for exposing, sharing, and connecting pieces of data, information, and knowledge. These best practices, as well as the delivered data, are also referred to as linked data. At the core of the linked data principles are the ideas of globally unique identifiers, i.e., Uniform Resource Identifiers (URIs) for data elements and a universal graph data model Resource Description Framework (RDF). By reusing the addressing system used for web pages, one can uniquely identify and link to data elements and datasets anywhere on the web [16]. The appreciation of Semantic Web technologies and linked data has increased considerably in the geospatial domain in the last decade, and they have fostered a promising approach to connecting SDIs with mainstream IT to augment the application of geospatial data [3]. Semantic Web technologies, especially linked data, provide a promising means to address some long-standing challenges in the geospatial domain, e.g., data integration (e.g., [3]) and knowledge formalization (e.g., [17]).

Pilot studies have been performed releasing INSPIRE-compliant data as linked data, and draft guidelines and vocabularies have been developed [18]. The development of INSPIRE linked data's URIs leveraged previous work on the standardization of unique identifiers for geospatial objects [19]. In the meantime, an increasing amount of geospatial data has been delivered as linked data, mainly by governmental agencies and large-scale data infrastructures [20]. The UK is a pioneer to this end; Ordnance Survey, Great Britain's national mapping agency (NMA), released several geospatial datasets as linked data nearly a decade ago [21]. However, the data relied on unstandardized methods to represent data semantics and thus lacked usability. In the Netherlands, Kadaster delivered several key geospatial datasets, e.g., building data and address data, as linked data on the web, together with other governmental open data, e.g., statistical data [22]. In Finland, the National Land Survey piloted the delivery of geographic name data, authoritative data, and building data as linked data [23]. In Norway, Kartverket also released some geospatial datasets as linked data [24]. A recent report summarized and reflected on the development of geospatial linked data in the Netherlands, Finland, Norway, and Spain. The fact that different projects use different RDF stores also renders the aim of this study necessary [25]. In the US, several geospatial linked data projects have been conducted: a pilot of design and development of linked data from The National Map was performed [26]; the Geographic Names Information System was served as linked data, and its geospatial visualization was enabled [20]; the GeoLink knowledge graph was published following linked data principles and served through a SPARQL endpoint, including Earth Science information captured by oceanographic cruises, physical sample metadata, etc. [27]. Along with these linked data, development endeavors from authorities, crowd-sourcing projects have also produced several geospatial linked datasets, and some of them are serving as central hubs of the LOD cloud, e.g., GeoNames (https://www.geonames.org/) and LinkedGeoData (a linked data distribution of OpenStreetMap [28]). Moreover, van den Brink et al. [29] proposed the best practice of delivering geospatial linked data, and they bridged the OGC web services and the Semantic Web. In the Earth Science domain, there have also been several discussions about how to utilize linked data for data integration and discovery (e.g., [30]).

Semantic Web technologies and linked data have also been utilized in a number of studies in the geospatial domain. The studies on this subject span several research areas, e.g., geoprocessing, information retrieval, and visualization. For example, Hofer et al. [31] developed a knowledge base to support the composition of geoprocessing workflows with ontologies and Semantic Web rule language (SWRL). Keßler et al. [32] leveraged linked data, ontologies, and SWRL rules for geospatial information

retrieval with context awareness. Wiemann and Bernard [33] used linked data for data integration in the environment of SDIs. Huang et al. [34] leveraged linked data and ontologies to realize the relative positioning of geospatial data, thus enabling geometrically self-adapting web maps. Huang and Harrie [17] used linked data, ontologies, and semantic rules to realize knowledge-based visualization of geospatial data, thereby formalizing some visualization knowledge on the aspects of cartographic scale, data portrayal, and geometry source. To realize the potentials revealed by the above studies (e.g., the use of ontological reasoning, rule-based reasoning, and spatial operations), we need RDF stores with capabilities such as semantic query, semantic reasoning, and geospatial query. Therefore, we used these capabilities in this study as part of the RDF store selection criteria (cf. Section 4.1).

2.2. Assessment and Benchmarking of Spatially Enabled RDF Stores

As the Semantic Web evolved into the mainstream of the web and has been adopted in many scientific domains (e.g., life sciences, geosciences), assessments and benchmarks of RDF stores have been abundant, mainly on synthetic and artificial test datasets. Popular benchmarks include, in chronological order, the Lehigh University Benchmark (LUBM) [35], the SPARQL performance benchmark (SP^{2}Bench) [36], and the Berlin SPARQL Benchmark (BSBM) [37]. The DBpedia SPARQL benchmark (DBSB) [38] is a popular benchmark used for real-world linked data and queries (the queries are extracted from actual server logs). However, these benchmarks are mainly for common-use data and data from other domains, not geospatial data and queries. In addition, benchmarks based on synthetic data have been criticized because they have very little in common with the needs of real application domains [39].

For the assessment of spatially enabled RDF stores, in which an even higher level of complexity arises [40,41], Kolas [42] proposed and performed a benchmark for the geospatial query capacity of RDF stores; however, since it was proposed before the standardization of GeoSPARQL, not much from that work can be applied to today's developments. Battle and Kolas [43] demonstrated the geospatial capacity of Parliament and successfully ran a number of GeoSPARQL-compliant queries. Garbis et al. [14] presented the benchmark Geographica to assess several spatially enabled RDF stores in which spatial queries were written in both GeoSPARQL and stSPARQL (the spatiotemporal query language in the RDF store Strabon). In that benchmark, three RDF stores were evaluated, i.e., Strabon, uSeekM, and Parliament, in a micro-benchmark and a macro-benchmark. The micro-benchmark aims to test the efficiency of primitive spatial functions in spatially enabled RDF stores; the macro-benchmark aims to test the performance of the stores in some certain application scenarios, e.g., reverse geocoding, map search, etc. This benchmark's datasets and queries have been published online (http://geographica.di.uoa.gr/), and the benchmark was based on both real-world geospatial data (e.g., LinkedGeoData) and synthetic data. The GeoKnow project, which dealt with geospatial Semantic Web and linked data, released a thorough survey and evaluation of spatially enabled RDF stores, with a partial focus on GeoSPARQL compliance [44]. The stores evaluated in GeoKnow include Virtuoso, Parliament, OWLIM, uSeekM, and Strabon, as well as spatially enabled relational databases, i.e., Oracle Spatial and PostgreSQL with PostGIS extension. Bellini and Nesi [45] assessed several well-known RDF stores, including Virtuoso, GraphDB, Oracle, and Stardog, for semantically enabled smart city services. The geospatial capacity of these RDF stores was one of the focuses of this study, as smart city services also have the need for capabilities such as temporal data query. The benchmark was based on the Florence Smart City model; the used datasets and tools are available online. These benchmarks clearly demonstrated the sparse support for spatial operations in RDF stores, and the RDF stores supporting GeoSPARQL were very few. Specifically, many RDF stores, e.g., Virtuoso, used their own syntaxes for geospatial queries rather than GeoSPARQL, and most RDF stores supporting GeoSPARQL queries were developed in academic environments, e.g., Parliament. Furthermore, the query performance was generally unsatisfactory, which also undermined the usability of these very few spatially enabled RDF stores.

The abovementioned previous works provide useful grounds for this study to evaluate the geospatial query capacity of RDF stores for future SDIs and for the geospatial linked data community at large. However, these previous studies have some limitations. First, the results are now mostly outdated, as the status of the tested RDF stores have changed considerably: some of them have developed with more advanced support for geospatial queries and increased GeoSPARQL compliance, and some of them have become obsolete and are rarely used. Second, the assessments and benchmarks targeting geospatial query (i.e., Geographica and GeoKnow benchmarks) depended on either synthetic data or purely geospatial data (in which nearly all the data objects have geometric information and are involved in spatial indexing/search). Our first test scenario, which uses data from ICOS CP, is, however, an Earth Science data infrastructure with a portion of geospatial data, which is more in line with the current role of geospatial data in large data infrastructures (open SDI). In addition, we provide a reproducible benchmark with deliverables that others can use to assess the RDF stores on their own datasets. Additionally, one shortcoming of previous spatially enabled RDF stores' benchmarking works is that they fully focused on evaluating the query performance (response time), but they did not assess the correctness of the returned results. In this paper, we assess query correctness in the first scenario.

3. Benchmarking Datasets

3.1. ICOS Carbon Portal Metadata

In the first scenario, we used data from ICOS CP (see supplementary files). ICOS is a Pan-European research infrastructure that currently has 12 member countries and a legal status of European Research Infrastructure Consortium (ERIC) (https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/eric_en). It is a European measurement system for high-quality and precision greenhouse gas observations and environmental monitoring. Currently, there are 135 measurement stations (including co-located ones), with 33 atmosphere stations, 81 ecosystem stations, and 21 ocean stations (Figure 1 shows the geographic locations of the stations).

ICOS CP is the data portal that provides free and open access to all ICOS datasets. ICOS data products include quality-controlled observational data, elaborated (model) products, and synthesis reports, which is material for policymakers. The users of ICOS CP span various domains, e.g., (Earth Science) researchers, education users, policymakers, and stakeholders in the negotiation of carbon reduction policies. ICOS produces around 25–30 TB of sensor data per year, together with about 1 GB of processed data products and 5–20 TB of elaborated data products. Additionally, as ICOS CP has become a well-recognized data sharing and distribution platform, some other data initiatives and producers, e.g., SOCAT (https://www.socat.info/), have also contributed by publishing their data through ICOS CP. The observation data at ICOS CP is linked to georeferenced locations. The atmospheric and ecosystem observations are connected to the coordinates of the measurement stations. For the ocean data, ship trajectories are stored as lists of *XY* coordinate pairs. The huge amount of data delivered and the complex organizational structure and responsibility raise the importance of data cataloging and discovery.

ICOS CP is an active practitioner of the FAIR principles, which aim to make data **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable [46,47]. In this context, ICOS CP has adopted linked data for delivering and publishing all its metadata (including metadata for ICOS data and other data harvested by ICOS CP, e.g., SOCAT data) to make such data more discoverable. The metadata is available through, among others, a SPARQL endpoint (https://meta.icos-cp.eu/sparqlclient). Geospatial data forms a part of the ICOS CP metadata. As the size of ICOS CP metadata is constantly growing because observational data is continually ingested, query performance will become a notable issue. To accelerate the spatial search of ocean data, each trajectory is simplified into a line string or a polygon (concave hull of the trajectory) containing a maximum of 20 coordinate pairs (by an in-house developed

European research infrastructure that currently has 12 member countries and a legal status of European Research Infrastructure Consortium (ERIC) (https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/eric_en). It is a European measurement system for high-quality and precision greenhouse gas observations and environmental monitoring. Currently, there are 135 measurement stations (including 7 associated ones), with 33 atmosphere stations, 53 ecosystem stations, and 21 ocean stations (Figure 1 shows the geographic locations of the stations).

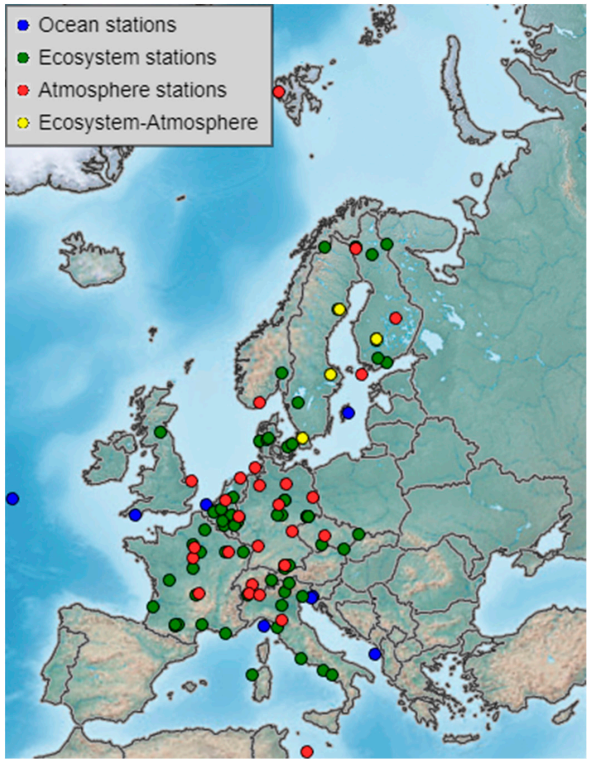


Figure 1. Geographic locations of Integrated Carbon Observation System (ICOS) measurement stations.

The linked data implementation is built upon a set of ontologies for different scopes of the data portal responsibility. Among them, the most important ontology is the ICOS CP metadata ontology (with the prefix *cpmeta* (https://meta.icos-cp.eu/ontologies/cpmeta/)). The ICOS CP metadata ontology relies on and has strong interoperability with some W3C standard ontologies, e.g., W3C PROV ontology [49] and W3C organization ontology [50]. For the details of ICOS CP ontologies, please refer to its GitHub repository (https://github.com/ICOS-Carbon-Portal/meta/tree/master/src/main/resources/owl) or the online description (http://static.icos-cp.eu/share/slides/dataServiceWorkshop/#/).

In the ICOS CP metadata ontology, the instances of the class *DataObject* can be associated with the instances of the class *SpatialCoverage*, and the instances of *SpatialCoverage* can be associated with the serialization of the corresponding geometries (Figure 2 demonstrates a part of the ICOS metadata ontology that is relevant to spatial information.). Currently, the ICOS metadata ontology is not GeoSPARQL-compliant (the GeoSPARQL classes are not introduced into ICOS metadata ontology, and the geometries are serialized in GeoJSON, which is not supported by GeoSPARQL). To support geospatial (GeoSPARQL) queries, we redesigned the ontology to accomplish GeoSPARQL compliance, as illustrated in Figure 2 (we use *geo* for the prefix of GeoSPARQL). That is, we built an inheritance relation in which *SpatialCoverage* is a subclass of *geo:Geometry*, and the instances can thereby be associated with the geometries in Well-Known Text (WKT) to enable GeoSPARQL-compliant geospatial queries. Afterward, we transformed all the geometries from GeoJSON to WKT using several SPARQL CONSTRUCT queries (the queries are available online at https://github.com/RightBank/Benchmarking-spatially-enabled-RDF-stores/tree/master/TransformationSPARQLQueries.

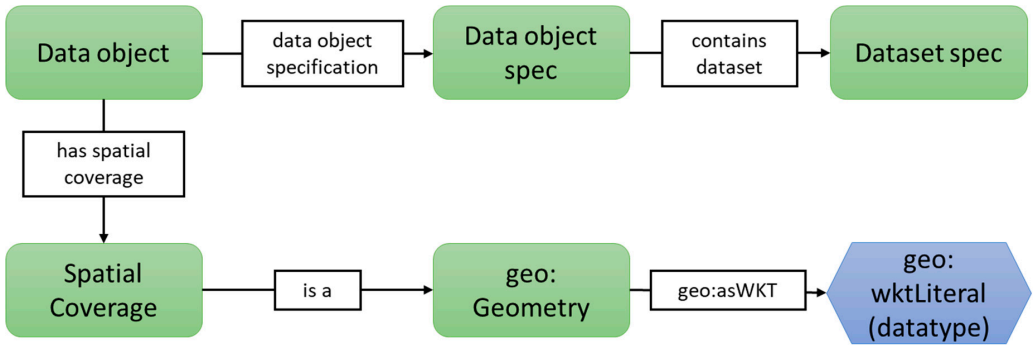


Figure 2. Geospatial part of the ICOS metadata ontology. The concepts and relations without prefix annotation are from ICOS metadata ontology.

The test data for RDF store assessment and benchmarking is the entire set of metadata of ICOS CP, which has 2,194,299 RDF statements as of 18 March 2019. The dataset has been published online [51]. Among the data, there are 1068 spatial objects (88 polygons, 853 polylines, and 127 points). We believe that this situation mirrors the current development of geospatial data that it forms a part of a large-scale information infrastructure. Therefore, the results of this study can also be used as a reference for other linked data implementations with similar situations. Technically, extracting and querying on relevant geospatial data from mass data, including relevant and irrelevant data, is costlier for query planners in the RDF stores than merely operating without query-irrelevant data.

The most important geospatial query requirement for ICOS CP is to enable users to directly spatially select different types of data objects (e.g., measurement trajectories) in user-defined geometric ranges, which could be a simple rectangle or an arbitrary complex polygon that is drawn by the users. In this context, the topological relations within, intersects, and overlaps are useful, but we also would like to support other geospatial functions available in GeoSPARQL, such as buffer, disjoint, and crosses, for specific user needs and requirements. Therefore, we tested the available spatial functions in some RDF stores that are not restricted to the functions for spatial selections (cf. Section 4.2).

3.2. Geographica Benchmarking Datasets

For the second scenario, in which the benchmarking is performed on a large amount of purely geospatial data, we used real-world datasets from the Geographica benchmark. Six real-world geospatial datasets in RDF were used: DBpedia, GeoNames, road networks and rivers from Greece, the Greek Administrative Geography dataset, the CORINE Land Use/Land Cover dataset, and wildfire hotspots from the National Observatory of Athens. The geographic coverages of the six datasets are in Greece. The six datasets contain more than 30,000 points, 12,000 polylines, and 104,000 polygons. Details of the datasets are provided in [14] and its online repository (http://geographica.di.uoa.gr/).

4. Evaluation Methodology

The evaluation of spatially enabled RDF stores was carried out in two stages. In the first stage, we selected the RDF stores using a set of criteria and deeply analyzed the geospatial features provided by the selected stores (e.g., GeoSPARQL compliance, licensing, spatial indexing, etc.). The successive second stage applied a benchmark to the RDF stores in the above-discussed two scenarios. It is based on a set of SPARQL queries that are capable of testing the geospatial query performance of the stores.

4.1. RDF Store Selection and Analysis

The selection of the tested RDF stores is based on the needs both of large-scale information infrastructures (ICOS CP in this case) and dedicated SDIs. First, general selection criteria were applied:

- The RDF store should be popular, well-known, and actively supported by a community or backed by a commercial vendor.
- The RDF store should support W3C standards, e.g., SPARQL 1.1.
- The RDF store should support semantic reasoning, which can be either triple materialization at load time or at query time (query rewriting), and the widely used reasoning types should be supported (e.g., RDFS, OWL, OWL2, OWL2-DL, etc.). Additionally, rule-based reasoning should be supported.
- The RDF store should have geospatial query capacity, preferably with GeoSPARQL support and compliance.

On the basis of these criteria, a pre-selection was made. The final selection was then based on a qualitative analysis of the pre-selected RDF stores by reading the documentation (we contacted the vendor for Stardog, as we could not find information about its spatial index technique in its documentation). The key aspects of this analysis include the following:

- Software components, architecture, deployment, and licensing;
- The means of data loading, query, and management;
- Utilization of software components from other solutions (e.g., if it is based on open-source frameworks);
- Supported semantic reasoning types;
- Geospatial query capacity and GeoSPARQL compliance;
- The employment of spatial indexing for geospatial data and the types of indexing;
- The popularity of the RDF stores is partially consulted from DB-Engines ranking (https://db-engines.com/en/ranking/rdf+store).

Through the qualitative analysis, not only can we choose the evaluated RDF stores in our work, but we can also obtain an up-to-date view of the popular RDF stores, especially to gain insight concerning the recent development of spatially enabled RDF stores and their GeoSPARQL compliance.

4.2. Performance Benchmark of Geospatial Query in RDF Stores

In this study, we reused and tailored the micro-benchmark from the Geographica benchmark [14] to evaluate the RDF stores. The micro-benchmark from Geographica aims to test the efficiency of primitive spatial functions in spatially enabled RDF stores. Simple SPARQL queries that consist of one or two triple patterns and a spatial function were used as benchmark queries. This benchmark includes non-topological geometric construction, simple spatial selections, and more complex operations (e.g., spatial join). In the first scenario, we tailored the benchmark queries for ICOS CP metadata; a brief description of the tailored queries can be found in Table 1. For the second scenario, we adopted the original query set from Geographica [14]. In addition, in both scenarios, *Q6* (area calculation), *Q28* (extension constructing), and *Q29* (union constructing) were removed because these functions are not supported by GeoSPARQL and seldom supported by RDF stores. *Q14* (spatial within function to real-time constructed buffers) was also removed, as this query is semantically equivalent to *Q15* but more computationally expensive than *Q15* [14], and this type of nested spatial function is not always supported by RDF stores.

In our benchmark, we first warmed up the RDF stores with warm-up SPARQL queries in order to get the benchmark systems under normal working conditions, as the query performance in a cold state is often unstable and unpredictably low in the beginning because of factors such as the initial interpretation and compilation of codes. The warm-up queries are disjoint from the actual benchmark queries (cf. Table 1), and they are taken from the pre-defined queries at ICOS CP's SPARQL endpoint.

Table 1. Benchmark queries for spatially enabled Resource Description Framework (RDF) stores in the first scenario with Integrated Carbon Observation System carbon portal (ICOS CP) metadata. Q1–Q5 are non-topological construct functions, Q7–Q17 (excluding Q14) are spatial selection queries, and Q18–Q27 are spatial join queries.

	Operation	Query Description
Q1	Boundary	Construct boundary for each polygon
Q2	Envelope	Construct envelope for each polygon
Q3	Convex Hull	Construct convex hull for each polygon
Q4	Buffer	Construct buffer for each line string (polyline)
Q5	Buffer	Construct buffer for each polygon
Q7	Equals	Find all line strings that are spatially equal to a given line string
Q8	Equals	Find all polygons that are spatially equal to a given polygon
Q9	Intersect	Find all line strings that intersect with a given Polygon
Q10	Intersect	Find all polygons that intersect with a given polygon
Q11	Overlaps	Find all polygons that overlap a given polygon
Q12	Crosses	Find all line strings that cross a given line string
Q13	Within Polygon	Find all points that are spatially within a given polygon
Q15	Near a Point	Find all points that are within a fixed distance to a given point
Q16	Disjoint	Find all points that are disjoint from a given polygon
Q17	Disjoint	Find all line strings that are disjoint from a given polygon
Q18	Equals	Find point-to-point equality among all the points
Q19	Intersects	Find all points and lines that intersect with each other
Q20	Intersects	Find all points and polygons that intersect with each other
Q21	Intersects	Find all line strings and polygons that intersect with each other
Q22	Within	Find all points and polygons where the point lies inside the polygon
Q23	Within	Find all line strings, polygons where the line string lies inside the polygon
Q24	Within	Find all pairs of polygons where one polygon is within the other
Q25	Crosses	Find all line strings, polygons where the line string crosses the polygon
Q26	Touches	Find all pairs of polygons where the polygons touch each other
Q27	Overlaps	Find all pairs of polygons where the polygons overlap each other

4.3. Implementation—Reusable Benchmark Deliverables

The benchmarking of the RDF stores was implemented in Java. We encapsulated the SPARQL queries and the codes interoperating with the underlying RDF stores in executable Jar (Java archive) packages that can be directly run with Java Runtime Environment (JRE). The delivered Jar packages request the location of data source, warm-up query iteration times, and benchmark query iteration times. The deliverable programs and source codes (including the benchmark queries) are available online at https://github.com/RightBank/Benchmarking-spatially-enabled-RDF-stores.

After benchmarking, text files were generated with comprehensive information regarding data loading time, the execution time of each query in each iteration, and the query results (including resulted object numbers and the resulted features—mainly their geometries). The query execution time refers to the time elapsed between the point a query is sent to the RDF store and the point the query results are completely returned to the benchmark systems. The benchmark systems use the RDF stores in an embedded mode whenever possible.

5. Results of RDF Store Selection and Analysis

Using the selection criteria for testing RDF stores for this work, we thoroughly investigated a number of RDF stores, and we ultimately selected the following RDF stores for evaluation.

1. RDF4J 2.4.2: an open-source Java RDF framework under the license of Eclipse Distribution License, v1.0, formerly known as Sesame. It supports parsing, storing, inferencing, and querying RDF data. It supports SPARQL 1.1 and both ontological and rule-based reasoning. Inferred statements are materialized. It supports geospatial query in GeoSPARQL, and its spatial queries can be performed without spatial indexing or with Lucene Spatial (currently, Lucene Spatial in

RDF4J results in errors). RDF4J can be used as an RDF store or a library that communicates and operates with many third-party storage solutions (RDF stores).

2. Jena 3.9.0 + GeoSPARQL-Jena 1.0.3: an open-source Java framework for building Semantic Web and linked data applications. It supports SPARQL 1.1 and both ontological and rule-based reasoning. It provides both RDF API, which manipulates RDF data, and TDB, an RDF store solution. Jena is one of the most widely adopted RDF frameworks in various research and production projects. Jena itself has very limited spatial query capacity and does not support GeoSPARQL. The recently developed open-source plugin GeoSPARQL-Jena (https://github.com/galbiston/geosparql-jena) provides fully GeoSPARQL-compliant spatial query capacity with a custom spatial indexing technique. Both Jena and GeoSAPRQL-Jena are under Apache License 2.0.
3. Virtuoso Enterprise 8.2: one of the most well-known RDF stores because of its adoption by DBpedia. It supports SPARQL 1.1 and ontological and rule-based reasoning. The reasoning is performed by query rewriting, so inferred statements are not materialized. It has had geospatial query support for a few years, and it started to support GeoSPARQL in its commercial version in 2018 (it also claimed to support GeoSPARQL in its open-source edition, but, to date, no release has appeared, so we chose to use the commercial version). It uses R-tree as its spatial indexing technique. A proprietary license for the commercial edition and a GPL 2 license for the open-source version are used.
4. Stardog 6.0.1: a commercial knowledge graph product that supports parsing, storing, inferencing, and querying RDF data. It supports SPARQL 1.1 and both ontological and rule-based reasoning with a query rewriting strategy. It supports a few GeoSPARQL query functions with Lucene Spatial for spatial indexing. It is actively supported by a commercial company and uses proprietary licenses.
5. GraphDB 8.8.0: a linked data platform built upon RDF4J. It is a commercial solution that provides support for SPARQL 1.1 and ontological and rule-based reasoning. It supports GeoSPARQL with spatial indexing of Lucene Spatial (specifically, quad-prefix-tree and geohash-prefix-tree). It utilizes different strategies for handling queries with and without using a spatial index. GraphDB is under proprietary licenses.

The rationale for not selecting the formerly assessed and benchmarked spatially enabled RDF stores Parliament, Strabon, and uSeekM is that they are currently not actively supported by the community, and some of them have limited capacity for reasoning, particularly rule-based reasoning. That is, we only evaluated fully fledged and popular RDF stores with spatial query support.

The qualitative analysis of the selected stores resulted in a cross-store qualitative comparison. Table 2 compiles the results of qualitative analysis with a focus on spatial query capacity and GeoSPARQL compliance. The storage solutions adopted by the RDF stores are mainly divisible into two types: native (designed from scratch) and RDBMS-based (based on an existing relational database management system). Four of the five tested stores utilize native solutions for storage; only Virtuoso relies on an underlying RDBMS. All tested RDF stores support spatial operations for geometries serialized in WKT; only GraphDB and GeoSPARQL-Jena support GML as well. RDF4J, GeoSPARQL-Jena, Virtuoso, and GraphDB currently provide full support for GeoSPARQL functions (the queries with spatial relations in the simple features relation family), including non-topological construct functions (*Q1*–*Q5* in Table 1), spatial selection functions (*Q7*–*Q17* in Table 1), and spatial join functions (*Q18*–*Q27* in Table 1). Stardog only supports the functions that find the relations within, nearby, intersect, contains, disjoint, and equal, and it uses its own spatial query syntax. With regard to the spatial index technique, Lucene Spatial is commonly used because of its fast development and active support from the community. GeoSPARQL-Jena indexes and caches intermediate spatial query results to accelerate queries with similar graph patterns thereafter, and it supports dataset-custom spatial index constructing, which cannot be migrated to other datasets. Virtuoso uses R-tree for spatial indexing. In Virtuoso and Stardog, there is no way to switch off spatial queries with a spatial index, while the others support switching off spatial indexing. GeoSPARQL-Jena has been very recently

developed, and it supports transformation between different spatial reference systems (SRSs), whereas the other stores only support WGS84. This usually entails SRS transformation before importing into the stores.

Table 2. Qualitative analysis results of geospatial query support of the selected RDF stores.

	RDF4J	GeoSPARQL-Jena	Virtuoso	Stardog	GraphDB
Storage	Native	Native	RDBMS	Native	Native
Geometry serialization	WKT	WKT, GML	WKT	WKT	WKT, GML
GeoSPARQL-compliance [1]	Full	Full	Full	Partly	Full
Use of spatial index	Optional [2]	Optional	Must	Must	Optional
Spatial index technique	Lucene Spatial	Custom	R-tree	Lucene Spatial	Lucene Spatial
Supported SRS	WGS84	Geographic and project SRSs	WGS84	WGS84	WGS84

[1] It refers to the compliance with spatial functions in the simple features relation family; e.g., it does not include support for SRS and GML. [2] The support of Lucene Spatial in RDF4J currently has problems.

6. Results of the Spatially Enabled RDF Store Benchmark

6.1. Experimental Setup

We ran the benchmark in a machine with the processor Intel Core i7-6700 (8M Cache, up to 4.00 GHz), 24 GB of RAM, and the operating system Ubuntu 18.04.1 LTS.

In the first scenario, the ICOS CP metadata was exported from its current RDF4J-based store into an RDF dump file with the 2.2 M triples. In the second scenario, the Geographica data was downloaded from its online repository as dump files. The benchmark programs first loaded the dump files into each store and recorded the loading time (including the spatial index construction time).

Each query in the benchmark (Table 1) was run three times after a number of warm-up queries were finished. In order to test the difference between using and not using a spatial index, we tested GraphDB in both modes (the queries *Q1–Q5* and *Q15* do not differ in either manner, as spatial indexing cannot be used in these queries in GraphDB). To determine the influence of the means of communication with the stores, we tested different communication interfaces with Virtuoso and Stardog. We tested Virtuoso's native interface Java Database Connectivity (JDBC) and RDF4J for operation and communication (as RDF4J is also commonly used as a library to manipulate other stores). We also tested Stardog's native interface SNARL and RDF4J for communication. We set a 1-h timeout for all queries.

6.2. Benchmark Results with ICOS CP Metadata

6.2.1. Query Performance

Table 3 summarizes the loading time for the ICOS CP metadata of each store. All the stores import, and possibly construct, the spatial index for the 2.2 M triple dataset in a reasonable time. Notice that the loading time is for the entire ICOS CP dataset, which contains around 1000 spatial objects and many other object types.

Table 3. Loading time of each store for ICOS CP metadata.

	RDF4J	GeoSPARQL-Jena	Virtuoso	Stardog	GraphDB
Loading time	62.4 s	88.0 s	94.5 s	134.1 s	154.1 s

Table 4 summarizes the results for the average query execution time regarding RDF4J, GeoSPARQL-Jena, Virtuoso (connected through JDBC and RDF4J), Stardog (connected through

SNARL and RDF4J), and GraphDB (with and without using a spatial index). For non-topological functions (*Q1*–*Q5*), GraphDB generally triumphs over the other stores. The performance of RDF4J is comparable to that of GraphDB. Compared with the other stores, GeoSPARQL-Jena and Virtuoso take much more time to calculate buffers of polylines and polygons, which might be the result of their more complex custom implementations. Stardog does not support any of the non-topological functions. For spatial selection queries (*Q7*–*Q17*), RDF4J provides generally good performance in terms of query response time. GraphDB also has comparable performance records, and it is much faster than the other stores for *Q7* (equal polyline finding). Virtuoso has the best performance for *Q13* (i.e., find all points in a given polygon, which is a very useful query for ICOS CP and many other linked data-based projects). Stardog has a reasonable performance but is much slower for *Q7* using its native SNARL interface. For spatial join queries (*Q18*–*Q27*), RDF4J provides the best performance for four queries (*Q20*, *Q21*, *Q22*, *Q27*), and it is generally fast at intersection queries. GeoSPARQL-Jena is fastest at *Q23*, *Q24*, and *Q25* and is generally superior at within functions. GraphDB is the best at *Q18* (without using a spatial index), *Q19* (with a spatial index), and *Q26* (with an index), and it generally provides reasonable performance for all queries. Virtuoso and Stardog are relatively slow for *Q19*, *Q20*, *Q23*, and *Q24*, which are mainly within and intersection queries; for these queries, the query performance differs by nearly three orders of magnitude, which indicates that some stores (Stardog and Virtuoso) may not be suited to the tasks of conducting spatial join queries.

Table 4. Average query response time of selected stores of benchmark queries with ICOS CP metadata (shortest response times in bold). Time unit is millisecond. The results that are different from the results produced from JTS (ArcGIS for Q15) are shaded (see Section 6.2.2).

Query Time (ms)	RDF4J	GeoSPARQL-Jena	Virtuoso		Stardog		GraphDB	
			JDBC	RDF4J	SNARL	RDF4J	Indexed	Non-Indexed
Q1	1.70	4.04	3.76	7.37			**1.51**	
Q2	1.27	2.62	2.14	2.14			**1.15**	
Q3	1.44	6.85	4.29	5.02			**1.19**	
Q4	1.45	100.93	944.95	979.93			**1.12**	
Q5	**1.29**	3.68	64.98	70.41			2.51	
Q7	21.48	12.84	53.11	56.62	142.72	33.58	**3.57**	5.83
Q8	7.13	4.34	8.97	10.20	11.93	4.23	13.13	**2.58**
Q9	**1.93**	4.83	21.02	22.80	10.19	5.20	5.57	19.32
Q10	**1.10**	3.68	10.13	11.71	11.90	4.02	4.27	2.53
Q11	**1.20**	3.39	9.39	12.54			9.73	2.80
Q12	**1.19**	3.64	55.17	47.79			2.83	2.95
Q13	2.54	5.04	**1.85**	4.05	10.05	4.49	4.03	7.17
Q15	2.35	20.20	2.10	4.80	24.57	3.30		**1.78**
Q16	**1.47**	2.51	1.78	4.63	8.96	3.39	2.83	5.31
Q17	**1.37**	1.87	28.15	26.82	10.80	3.73	2.34	2.24
Q18	136.81	71.19	39.92	45.84	280.10	196.99	31.01	**9.33**
Q19	2569.42	454.32	776.08	666.98	5786.66	5363.58	**4.55**	29.82
Q20	**1.75**	7.40	1536.15	1541.63	621.66	583.33	19.10	3.61
Q21	**1.51**	2.20	47.86	45.06	10.95	4.17	14.20	3.57
Q22	**1.25**	10.13	759.84	783.62	11.27	6.21	14.97	11.86
Q23	422.94	**3.57**	277.79	279.90	1605.72	1499.08	3.58	3.81
Q24	76.92	**2.80**	111.08	90.40	211.97	226.99	18.05	5.24
Q25	2.09	**1.73**	42.27	32.47			6.93	5.29
Q26	719.31	165.46	619.50	629.43			**19.11**	23.49
Q27	**2.19**	2.62	58.81	52.85			5.34	11.43

We also observe that the performance with Virtuoso's native JDBC interface is similar to that with the RDF4J interface. With Stardog, using RDF4J as the interface generally leads to better performance than using its native interface SNARL, as RDF4J caches some intermediate query results. From the results, we observe that GeoSPARQL-Jena and RDF4J demonstrate a significant caching effect, i.e., the query time of the second and third times substantially drops compared with that of the first time. This

is in line with their means of implementation: they cache a lot of intermediate query results. Other stores do not show a clear caching effect.

6.2.2. Query Correctness

Evaluating query correctness for spatial queries is complex, particularly when the queries deal with a large amount of data. However, query correctness is an important aspect in the assessment of the selected stores, especially because it is common for different stores to implement the spatial query functions differently. In this paper, we partially evaluate and discuss the query correctness by observing the results from the above-described benchmarking.

For topological queries, GeoSPARQL follows the definitions of topological relations in the dimensionally extended nine-intersection model DE-9IM [52]. A well-known and reliable implementation of DE-9IM is the Java library JTS Topology Suite, JTS (https://github.com/locationtech/jts). In this study, we performed all the benchmark queries using the JTS library, and we treat the returned results as reference results for the evaluation of the RDF stores. Queries whose number of returned results from the RDF stores differs from the number returned from JTS are shaded in Table 4. One exception is *Q15*, which is not supported by JTS (as JTS does not support distance calculation is geographic SRSs). Thus, we calculated it in ArcGIS 10.3.1 as reference results.

For *Q1–Q9*, all the evaluated stores provide the same number of returned results as JTS. For *Q10*, we find that Stardog handles the spatial relation intersect (for polygons) in a manner that differs from the other stores; it returns the same results as the other stores return for *Q11*, which queries all the polygons that overlap a given polygon. That is, the intersect function for polygons in Stardog is actually equivalent to the overlap function in other stores, and Stardog does not have the function overlap. For *Q15*, only GraphDB provides the same results as ArcGIS (10 results); RDF4J, Virtuoso, and Stardog return 11 results (probably linked to precision settings); and GeoSPARQL-Jena fails to give any result in spite of the relatively long time it takes on this query. For *Q18*, RDF4J fails to return any result, and this problem is potentially linked to the precision setting in RDF4J when finding equal points. For *Q21* and *Q25*, RDF4J, Stardog (only for *Q21*), and GraphDB (using spatial indexing) return 563 results; Virtuoso returns 567 results; and GeoSPARQL-Jena, and GraphDB (without using spatial indexing) return 565 results. This divergence may be linked to Lucene Spatial filtering out some results because of factors such as precision settings in different stores. JTS returns 565 results for these queries.

6.3. Benchmark Results with Geographica Datasets

In the second scenario, we tested the selected RDF stores with large geospatial datasets. This scenario is more in line with conventional SDIs, in which geospatial data dominates. Therefore, benchmarking the RDF stores with such large datasets to test their scalabilities will potentially benefit the SDI and geospatial linked data communities, as it is common for a project (especially dedicated SDIs) to have a vast number of geospatial objects.

The loading time of the six datasets in the five selected stores is presented in Table 5, and the query performance is demonstrated in Table 6.

From Table 5, we can observe that a large number of geospatial objects do not lengthen the loading time for RDF4J, GeoSPARQL-Jena, and GraphDB. For RDF4J and GeoSPARQL-Jena, this is because they do not build a spatial index while data loading; for GraphDB, the spatial index construction is completed in a short time. Virtuoso takes longer (more than 10 min) to load and construct a spatial index for the data. For Stardog, the spatial indexing process is slow, as the whole loading and index construction process takes nearly five hours.

Table 5. Loading time of each store for Geographica datasets.

	RDF4J	GeoSPARQL-Jena	Virtuoso	Stardog	GraphDB
Loading time	48.6 s	89.6 s	620.0 s	4.6 h	89.7 s

Table 6. Average query response time of selected stores of benchmark queries with Geographica datasets (shortest response time in bold). Time unit is second unless specified as hour.

Query Time (s)	RDF4J	GeoSPARQL -Jena	Virtuoso		Stardog		GraphDB	
			JDBC	RDF4J	SNARL	RDF4J	Indexed	Non-Indexed
Q1	0.015	0.011	0.020	0.088			**0.009**	
Q2	0.011	0.003	0.023	0.016			**0.002**	
Q3	0.011	**0.005**	0.059	0.074			0.009	
Q4	0.006	0.079	0.043	0.061			**0.005**	
Q5	**0.003**	**0.003**	0.203	0.250			**0.003**	
Q7	0.527	**0.055**	0.120	0.130	0.515	0.533	0.079	2.515
Q8	0.482	0.139	0.148	0.156	0.178	0.140	**0.139**	5.536
Q9	0.013	**0.005**	0.022	0.035	0.021	0.017	17.442	0.046
Q10	0.776	**0.012**	0.120	0.181	0.095	0.083	0.125	0.867
Q11	0.685	**0.014**	0.077	0.125			0.404	0.034
Q12	**0.009**	**0.005**	0.094	0.116			1.880	0.076
Q13	0.093	**0.003**	0.052	0.054	0.786	0.776	13.163	0.026
Q15	0.222	4.529	**0.119**	0.147	0.921	0.760	1.645	
Q16	**0.003**	0.384	0.006	**0.009**	0.013	0.009	124.135	**0.003**
Q17	0.003	**0.002**	0.027	0.030	0.008	0.007	148.334	**0.002**
Q18	0.027	0.060	0.060	**0.009**	0.014	0.008	0.010	1.491
Q19	>1 h	**544.082**	938.553	932.699	>1 h	>1 h	1026.021	>1 h
Q20	9.031	**0.021**	2.677	2.679	2416.730	2439.824	1.013	9.887
Q21	3.985	**0.005**	1.715	1.673	4.174	4.471	2.969	3.573
Q22	8.569	**0.003**	2.071	2.130	0.380	0.386	0.441	9.104
Q23	5.940	**0.004**	2.370	2.463	4.681	4.857	1.677	3.382
Q24	7.875	**0.007**	2.358	2.529	0.129	0.113	0.099	3.304
Q25	3.940	**0.017**	6.531	6.596			0.612	62.865
Q26	0.040	**0.033**	3.460	3.758			0.531	7.431
Q27	18.274	0.111	1.059	0.644			**0.077**	17.337

The query performance of GraphDB is generally better than that of the others for the non-topological construct queries *Q1*–*Q5*, and RDF4J, GeoSPARQL-Jena, and Virtuoso have comparable performances. For the spatial selection queries *Q7*–*Q17*, all the RDF stores respond in a reasonable time, and GeoSPARQL-Jena performs better than the others in most of the queries. The spatial join query *Q19* is the most computationally expensive query in the benchmark: RDF4J, Stardog, and GraphDB without spatial indexing all time out for this query, while GeoSPARQL-Jena provides the shortest time for this query (less than 10 min). For other spatial join queries, *Q20*–*Q27*, GeoSPARQL-Jena generally performs better than the others, and all stores have reasonable response times. It is observed that different query interfaces do not have much effect on the query response time. For GraphDB, the indexed mode generally returns the results much quicker than the non-indexed mode. The exceptions are *Q16* and *Q17*, for which GraphDB has a very similar performance to that of RDF4J with quick responses; this might be the result of the simplistic implementation of the disjoint function in RDF4J (GraphDB is dependent on RDF4J in the mode that does not use spatial indexing).

7. Discussion

In this paper, we comprehensively assess and benchmark five popular and well-known spatially enabled RDF stores, i.e., RDF4J, GeoSPARQL-Jena, Virtuoso, Stardog, and GraphDB. It is encouraging to see the increasing maturity of the technical environment for the support of geospatial linked data, as well as the increasing compliance with GeoSPARQL compared with previous benchmarks. That is, progressively more mainstream and well-known RDF stores are (partially) supporting GeoSPARQL. Another positive observation is that the syntaxes used for geospatial queries with GeoSPARQL are the same in RDF4J, GeoSPARQL-Jena, Virtuoso, and GraphDB in this benchmark, which implies that the geospatial queries are cross-database interoperable in terms of query syntax (Stardog does not have the same geospatial query syntax as the others). Listing 1 is an example query of *Q23* in the first scenario in RDF4J, GeoSPARQL-Jena, Virtuoso, and GraphDB (without using spatial indexing, as the

filter should be replaced with a triple relation in the query when using spatial indexing in GraphDB, i.e., `?geom1 geo:sfWithin ?geom2.`). Listing 2 is the corresponding query used in Stardog.

Listing 1. Query syntax of Q23 in the first scenario in RDF4J, GeoSPARQL-Jena, Virtuoso, and GraphDB (without indexing).

```
PREFIX geo:  <http://www.opengis.net/ont/geosparql#>
PREFIX geof:  <http://www.opengis.net/def/function/geosparql/>
PREFIX sf:  <http://www.opengis.net/ont/sf#>
SELECT ?geom1 ?geom2
WHERE {
?geom1 a sf:LineString; geo:asWKT ?wkt1.
?geom2 a sf:Polygon; geo:asWKT ?wkt2.
FILTER(geof:sfWithin(?wkt1,?wkt2)).}
```

Listing 2. Query syntax of Q23 in the first scenario in Stardog.

```
PREFIX geo:  <http://www.opengis.net/ont/geosparql#>
PREFIX geof:  <http://www.opengis.net/def/function/geosparql/>
PREFIX sf:  <http://www.opengis.net/ont/sf#>
SELECT ?geom1 ?geom2
WHERE {
?geom1 a sf:LineString.
?geom2 a sf:Polygon.
FILTER(geof:relate(?geom1,?geom2,geo:within)).}
```

The query performance is generally acceptable, and it is much better than previous benchmarking results because RDF stores have developed and computer hardware has advanced. GeoSPARQL was supported in all the stores except for Stardog after 2018, which also makes this paper timely in its contribution to the comprehensive understanding of this subject. We believe the increasingly mature technical environment will benefit the development of the next generation of SDIs, in which linked data will expectedly play an important role.

From the query performance of the evaluated stores in the two scenarios, we observe that GraphDB is generally better than the others at non-topological queries, which are useful in many real-world spatial analyses: e.g., buffering is important for location selection analysis. GeoSPARQL-Jena and RDF4J are generally better than the other RDF stores at spatial selection queries, which are useful for many real-world use cases: e.g., for ICOS CP, the overlap and within functions are the most useful queries for enabling a user-defined spatial search. GeoSPARQL-Jena is superior at spatial join queries—operations used for functions such as establishing relations between the cadaster registries (points) and building objects (polygons).

A prerequisite of (partially) achieving cross-database interoperability is that the GeoSPARQL standard should be used when possible. The lightweight nature of the GeoSPARQL vocabulary means that accomplishing interoperability with GeoSPARQL for other spatial-relevant ontologies does not entail much work since, in most cases, it can be accomplished with subclass/subproperty inheritance. Nevertheless, we believe that GeoSPARQL should support more serializations to realize its wider adoption. It is especially desirable to have support for GeoJSON, which is widely accepted by the web development community.

One lesson learned from the experimental results is that, for a moderate amount of geospatial data (scenario 1 with about 1000 spatial objects), spatial indexing could be an overhead both for data loading and querying, whereas spatial indexing is certainly necessary when querying a large number of geospatial objects (scenario 2 with about 150,000 spatial objects). Most selected RDF stores provide reasonable data loading and spatial index construction times, except for Stardog, which takes nearly five hours to load and index the Geographica datasets. That is, we believe that enabling spatial indexing for querying large geospatial datasets is imperative, and constant change and injection of data

are also feasible as long as the data loading and indexing times are reasonable. In this context, further assessment of the RDF stores with an even larger amount of data is desirable, which is interesting for large-scale geospatial linked data deployment.

From this assessment, we observe that most of the selected RDF stores with spatial indexing use Lucene Spatial for its easy deployment and wide support from the community. We argue that no spatial indexing technique can best fit all applications. In fact, it would be better to also enable developers and geospatial experts to configure specific and optimized spatial indexes tailored for certain datasets. This functionality is already provided by some RDF stores, e.g., RDF4J and GeoSPARQL-Jena.

Despite the promising results and advancements, there are still some challenges. One of the most significant challenges is query correctness. Although the queries are interoperable in terms of query syntax across most of the selected RDF stores, the returned results are sometimes not the same because of different implementations and interpretations of, for example, spatial topological relations. This issue renders the cross-database interoperability problematic for geospatial queries, which is rarely the case for other types of queries following the W3C recommendations. We think further development of the RDF stores might mitigate this issue, but to overcome this problem, we may need a community-backed and commonly used compliance testing suite regarding the OGC Implementation Standard for Geographic Information [52] for the implementation and interpretation of spatial functions. For the query correctness issue, we propose that a major cause is the different strategies for handling precision in the stores. Furthermore, as only four of the five stores support the SRS of WGS84, conducting spatial operations in a geographic SRS and converting data from other SRSs to WGS84 can lead to precision loss and thus incorrect or inaccurate results. Therefore, further investigation of the effect of precision settings in RDF stores is deemed necessary.

Another important topic that deserves investigation is the performance comparison between spatially enabled RDF stores and state-of-the-art OGC services (e.g., WFS). We speculate that current OGC services are superior to RDF stores at spatial queries. This raises the question of how much faster OGC services are than RDF stores. The answer to that question will potentially unveil the answers to two other questions: (1) Should we (partly) leave the spatial operations to RDBMS-backed OGC services or other GIS tools, especially since spatial join queries do not perform favorably in the evaluated RDF stores, until their spatial capacities are significantly advanced? (2) Should data publishers or third parties pre-compute important and relevant spatial relations and publish them along with the data, which will greatly diminish the need for real-time spatial operations at the cost of pre-computation and increase in data volume? Our initial opinion is that it will be beneficial to pre-compute some important spatial relations and release the relations together with geospatial linked data.

8. Conclusions

Linked data is a promising means to resolve the limitations concerning data integration and semantic heterogeneity of the current SDI solutions; thus, linked data has been seen as one of the key factors moving SDIs toward the next generation. The technical environment and support are important for deploying geospatial linked data. In this paper, we present an assessment and benchmarking concerning the spatial query capacities of five RDF stores, i.e., RDF4J, GeoSPARQL-Jena, Virtuoso, Stardog, and GraphDB. We tested the selected stores in two scenarios. One scenario involves benchmarking the RDF stores with ICOS CP metadata, a large-scale Earth Science data infrastructure in which geospatial data is integrated with other types of data. The other scenario is in a dedicated SDI environment with a large amount of purely geospatial data, which is a mixture of crowd-sourced and authoritative geospatial data. The queries used in this study are mainly from the Geographica benchmark. The results demonstrate that GeoSPARQL compliance has advanced dramatically in the last several years for the RDF stores, and query performances are generally acceptable. Furthermore, spatial indexing is important when querying a large number of geospatial objects. However, query correctness remains a challenge for cross-database interoperability.

Supplementary Materials: The benchmarking programs used in this study are available at https://github.com/RightBank/Benchmarking-spatially-enabled-RDF-stores. The test data from ICOS CP has been published at https://doi.org/10.18160/9D9W-WT2P.

Author Contributions: W.H., O.M., and L.H. conceived and designed this study; S.A.R. and W.H. implemented the benchmarking for the RDF stores; W.H. wrote the paper, with revisions from S.A.R., O.M., and L.H.; O.M. is the system architect at ICOS CP and partially proposed the need for this study; all authors read and approved this manuscript.

Funding: The work was supported by Lund University and China Scholarship Council.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Van den Brink, L.; Janssen, P.; Quak, W.; Stoter, J. Towards a high level of semantic harmonisation in the geospatial domain. *Comput. Environ. Urban Syst.* **2017**, *62*, 233–242. [CrossRef]
2. INSPIRE. Available online: https://inspire.ec.europa.eu/ (accessed on 2 December 2018).
3. Schade, S.; Smits, P. Why linked data should not lead to next generation SDI. In Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Munich, Germany, 22–27 July 2012; pp. 2894–2897.
4. Parsons, E. If You Can't Link to it … Does it Exist? Available online: https://www.edparsons.com/2017/09/cant-link-exist/ (accessed on 24 April 2019).
5. Janowicz, K.; Scheider, S.; Pehle, T.; Hart, G. Geospatial semantics and linked spatiotemporal data–Past, present, and future. *Semant. Web* **2012**, *3*, 321–332.
6. Vancauwenberghe, G.; Valeckaite, K.; Van Loenen, B.; Donker, F.W. Assessing the Openness of Spatial Data Infrastructures (SDI): Towards a Map of Open SDI. *IJSDIR* **2018**, *13*, 88–100.
7. Lutz, M.; Sprado, J.; Klien, E.; Schubert, C.; Christ, I. Overcoming semantic heterogeneity in spatial data infrastructures. *Comput. Geosci.* **2009**, *35*, 739–752. [CrossRef]
8. EuroSDR. EuroSDR Annual Report 2018. Available online: http://www.eurosdr.net/sites/default/files/images/inline/eurosdr_annual_report_2018.pdf (accessed on 12 June 2019).
9. AGILE 2018 Workshop 'SDI Research and Strategies towards 2030'. Available online: https://kcopendata.eu/sdi2030/ (accessed on 25 July 2018).
10. W3C. Resource Description Framework (RDF). Available online: https://www.w3.org/RDF/ (accessed on 6 January 2018).
11. W3C. SPARQL Query Language for RDF. Available online: https://www.w3.org/TR/rdf-sparql-query/ (accessed on 20 March 2019).
12. Perry, M.; Herring, J. OGC GeoSPARQL-A Geographic Query Language for RDF Data. Technical report, Open Geospatial Consortium, 2012. Available online: http://www.opengeospatial.org/standards/geosparql (accessed on 1 May 2019).
13. ICOS Carbon Portal. Available online: https://www.ICOSCP.eu/ (accessed on 7 January 2019).
14. Garbis, G.; Kyzirakos, K.; Koubarakis, M. Geographica: A benchmark for geospatial RDF stores (long version). In Proceedings of the International Semantic Web Conference, Sydney, NSW, Australia, 21–25 October 2013; pp. 343–359.
15. World Wide Web Consortium (W3C). W3C Semantic Web Activity. Available online: https://www.w3.org/2001/sw/ (accessed on 25 July 2017).
16. Kuhn, W.; Kauppinen, T.; Janowicz, K. Linked data-A paradigm shift for geographic information science. In Proceedings of the International Conference on Geographic Information Science, Vienna, Austria, 24–26 September 2014; pp. 173–186.
17. Huang, W.; Harrie, L. Towards knowledge-based geovisualisation using Semantic Web technologies: A knowledge representation approach coupling ontologies and rules. *Int. J. Digit. Earth* **2019**. [CrossRef]
18. INSPIRE. Linking INSPIRE Data: Draft Guidelines and Pilots. Available online: https://inspire.ec.europa.eu/news/linking-inspire-data-draft-guidelines-and-pilots (accessed on 20 December 2018).
19. INSPIRE. Guidelines for the Encoding of Spatial Data. Available online: https://inspire.ec.europa.eu/documents/Data_Specifications/D2.7_v3.3rc3.pdf (accessed on 28 April 2019).

20. Regalia, B.; Janowicz, K.; Mai, G.; Varanka, D.; Usery, E.L. GNIS-LD: Serving and Visualizing the Geographic Names Information System Gazetteer as Linked Data. In Proceedings of the European Semantic Web Conference, Heraklion, Crete, Greece, 3–7 June 2018; pp. 528–540.
21. Goodwin, J.; Dolbear, C.; Hart, G. Geographical linked data: The administrative geography of great britain on the semantic web. *Trans. Gis* **2008**, *12*, 19–30. [CrossRef]
22. Folmer, E.; Beek, W.; Rietveld, L. Linked Data Viewing as part of the Spatial Data Platform of the Future. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*, 49–52. [CrossRef]
23. Hietanen, E.; Lehto, L.; Latvala, P. Providing Geographic Datasets as Linked Data in SDI. *Isprs-Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *41*, 583–586. [CrossRef]
24. Shi, L.; Sukhobok, D.; Nikolov, N.; Roman, D. Norwegian State of Estate Report as Linked Open Data. In Proceedings of the OTM Confederated International Conferences on the Move to Meaningful Internet Systems, Rhodes, Greece, 23–27 October 2017; pp. 445–462.
25. Ronzhin, S.; Folmer, E.; Mellum, R.; von Brasch, T.E.; Martin, E.; Romero, E.L.; Kytö, S.; Hietanen, E.; Latvala, P. Next Generation of Spatial Data Infrastructure: Lessons from Linked Data implementations across Europe. Report of Open ELS Project. Available online: https://openels.eu/wp-content/uploads/2019/04/V2_Next_Generation_SDI_Lessons-from-LD-implementations-across-Europe_1.pdf (accessed on 20 May 2019).
26. Usery, E.L.; Varanka, D. Design and development of linked data from the national map. *Semant. Web* **2012**, *3*, 371–384.
27. Cheatham, M.; Krisnadhi, A.; Amini, R.; Hitzler, P.; Janowicz, K.; Shepherd, A.; Narock, T.; Jones, M.; Ji, P. The GeoLink knowledge graph. *Big Earth Data* **2018**, *2*, 131–143. [CrossRef]
28. Stadler, C.; Lehmann, J.; Höffner, K.; Auer, S. Linkedgeodata: A core for a web of spatial open data. *Semant. Web* **2012**, *3*, 333–354.
29. Van den Brink, L.; Barnaghi, P.; Tandy, J.; Atemezing, G.; Atkinson, R.; Cochrane, B.; Fathy, Y.; Castro, R.G.; Haller, A.; Harth, A. Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web. *Semant. Web* **2019**, *10*, 95–114. [CrossRef]
30. Narock, T.; Shepherd, A. Semantics all the way down: The Semantic Web and open science in big earth data. *Big Earth Data* **2017**, *1*, 159–172. [CrossRef]
31. Hofer, B.; Mäs, S.; Brauner, J.; Bernard, L. Towards a knowledge base to support geoprocessing workflow development. *Int. J. Geogr. Inf. Sci.* **2016**, *31*, 694–716. [CrossRef]
32. Keßler, C.; Raubal, M.; Wosniok, C. Semantic rules for context-aware geographical information retrieval. In Proceedings of the European Conference on Smart Sensing and Context, Guildford, UK, 16–18 September 2009; pp. 77–92.
33. Wiemann, S.; Bernard, L. Spatial data fusion in spatial data infrastructures using linked data. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 613–636. [CrossRef]
34. Huang, W.; Mansourian, A.; Abdolmajidi, E.; Xu, H.; Harrie, L. Synchronising geometric representations for map mashups using relative positioning and Linked Data. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 1117–1137. [CrossRef]
35. Guo, Y.; Pan, Z.; Heflin, J. LUBM: A benchmark for OWL knowledge base systems. *Web Semant. Sci. Serv. Agents World Wide Web* **2005**, *3*, 158–182. [CrossRef]
36. Schmidt, M.; Hornung, T.; Lausen, G.; Pinkel, C. SP^ 2Bench: A SPARQL performance benchmark. In Proceedings of the 2009 IEEE 25th International Conference on Data Engineering, Shanghai, China, 29 March–2 April 2009; pp. 222–233.
37. Bizer, C.; Schultz, A. The berlin sparql benchmark. *Int. J. Semant. Web Inf. Syst.* **2009**, *5*, 1–24. [CrossRef]
38. Morsey, M.; Lehmann, J.; Auer, S.; Ngomo, A.-C.N. DBpedia SPARQL benchmark–performance assessment with real queries on real data. In Proceedings of the International Semantic Web Conference, Bonn, Germany, 23–27 October 2011; pp. 454–469.
39. Duan, S.; Kementsietsidis, A.; Srinivas, K.; Udrea, O. Apples and oranges: A comparison of RDF benchmarks and real RDF datasets. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, Athens, Greece, 12–16 June 2011; pp. 145–156.
40. Perry, M.; Sheth, A.P.; Hakimpour, F.; Jain, P. Supporting complex thematic, spatial and temporal queries over semantic web data. In Proceedings of the International Conference on GeoSpatial Sematics, Mexico City, Mexico, 29–30 November 2007; pp. 228–246.
41. Papadimitriou, F. The algorithmic complexity of landscapes. *Landscape Res.* **2012**, *37*, 591–611. [CrossRef]

42. Kolas, D. A Benchmark for Spatial Semantic Web Systems. In Proceedings of the International Workshop on Scalable Semantic Web Knowledge Base Systems, Karlsruhe, Germany, 26–30 October 2008.
43. Battle, R.; Kolas, D. Enabling the geospatial semantic web with parliament and geosparql. *Semant. Web* **2012**, *3*, 355–370.
44. Athanasiou, S.; Bezati, L.; Giannopoulos, G.; Patroumpas, K.; Skoutas, D. GeoKnow– Making the Web an Exploratory for Geospatial Knowledge: Deliverable 2.1.1 Market and Research Overview. Available online: http://svn.aksw.org/projects/GeoKnow/Public/D2.1.1_Market_and_Research_Overview.pdf (accessed on 30 September 2018).
45. Bellini, P.; Nesi, P. Performance assessment of RDF graph databases for smart city services. *J. Vis. Lang. Comput.* **2018**, *45*, 24–38. [CrossRef]
46. FORCE 11. Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.0. Available online: https://www.force11.org/fairprinciples (accessed on 5 March 2019).
47. Bechhofer, S.; Buchan, I.; De Roure, D.; Missier, P.; Ainsworth, J.; Bhagat, J.; Couch, P.; Cruickshank, D.; Delderfield, M.; Dunlop, I. Why linked data is not enough for scientists. *Future Gener. Comput. Syst.* **2013**, *29*, 599–611. [CrossRef]
48. Abam, M.A.; De Berg, M.; Hachenberger, P.; Zarei, A. Streaming algorithms for line simplification. *Discret. Comput. Geom.* **2010**, *43*, 497–515. [CrossRef]
49. W3C. PROV-O: The PROV Ontology. W3C Recommendation. Available online: https://www.w3.org/TR/prov-o/ (accessed on 28 April 2019).
50. W3C. The Organization Ontology. W3C Recommendation. Available online: https://www.w3.org/TR/vocab-org/ (accessed on 28 April 2019).
51. Mirzov, O.; Huang, W.; Raza, S.A. ICOS CP metadata used for RDF store benchmarking. *Res. Data.* **2019**. [CrossRef]
52. Herring, J. OpenGIS Implementation Standard for Geographic Information-Simple feature access-Part 1: Common architecture. *OGC Doc.* **2011**, *4*, 122–127.

Avhandlingar från Institutionen för naturgeografi och ekosystemanalys (INES), Lunds universitet

---

Dissertations from Department of Physical Geography and Ecosystem Science, University of Lund

---

Martin Sjöström, 2012: Satellite remote sensing of primary production in semi-arid Africa.

Zhenlin Yang, 2012: Small-scale climate variability and its ecosystem impacts in the sub-Arctic.

Ara Toomanian, 2012: Methods to improve and evaluate spatial data infrastructures.

Michal Heliasz, 2012: Spatial and temporal dynamics of subarctic birch forest carbon exchange.

Abdulghani Hasan, 2012: Spatially distributed hydrological modelling: wetness derived from digital elevation models to estimate peatland carbon.

Julia Bosiö, 2013: A green future with thawing permafrost mires?: a study of climate-vegetation interactions in European subarctic peatlands. (Lic.)

Anders Ahlström, 2013: Terrestrial ecosystem interactions with global climate and socio-economics.

Kerstin Baumanns, 2013: Drivers of global land use change: are increasing demands for food and bioenergy offset by technological change and yield increase? (Lic.)

Yengoh Genesis Tambang, 2013: Explaining agricultural yield gaps in Cameroon.

Jörgen Olofsson, 2013: The Earth: climate and anthropogenic interactions in a long time perspective.

David Wårlind, 2013: The role of carbon-nitrogen interactions for terrestrial ecosystem dynamics under global change: a modelling perspective.

Elin Sundqvist, 2014: Methane exchange in a boreal forest: the role of soils, vegetation and forest management.

Julie Mari Falk, 2014: Plant-soil-herbivore interactions in a high Arctic wetland: feedbacks to the carbon cycle.

Finn Hedefalk, 2014: Life histories across space and time: methods for including geographic factors on the micro-level in longitudinal demographic research. (Lic.)

Sadegh Jamali, 2014: Analyzing vegetation trends with sensor data from earth observation satellites.

Cecilia Olsson, 2014: Tree phenology modelling in the boreal and temperate climate zones : timing of spring and autumn events.

Jing Tang, 2014: Linking distributed hydrological processes with ecosystem vegetation dynamics and carbon cycling: modelling studies in a subarctic catchment of northern Sweden.

Wenxin Zhang, 2015: The role of biogeophysical feedbacks and their impacts in the arctic and boreal climate system.

Lina Eklund, 2015: "No Friends but the Mountains": understanding population mobility and land dynamics in Iraqi Kurdistan.

Stefan Olin, 2015: Ecosystems in the Anthropocene: the role of cropland management for carbon and nitrogen cycle processes.

Thomas Möckel, 2015: Hyperspectral and multispectral remote sensing for mapping grassland vegetation.

Hongxiao Jin, 2015: Remote sensing phenology at European northern latitudes: from ground spectral towers to satellites.

Bakhtiyor Pulatov, 2015: Potential impact of climate change on European agriculture: a case study of potato and Colorado potato beetle.

Christian Stiegler, 2016: Surface energy exchange and land-atmosphere interactions of Arctic and subarctic tundra ecosystems under climate change.

Per-Ola Olsson, 2016: Monitoring insect defoliation in forests with time-series of satellite based remote sensing data: near real-time methods and impact on the carbon balance.

Jonas Dalmayne, 2016: Monitoring biodiversity in cultural landscapes: development of remote sensing- and GIS-based methods.

Balathandayuthabani Panneer Selvam, 2016: Reactive dissolved organic carbon dynamics in a changing environment: experimental evidence from soil and water.

Kerstin Engström, 2016: Pathways to future cropland: assessing uncertainties in socio-economic processes by applying a global land-use model.

Finn Hedefalk, 2016: Life paths through space and time: adding the micro-level geographic context to longitudinal historical demographic research.

Ehsan Abdolmajidi, 2016: Modeling and improving Spatial Data Infrastructure (SDI).

Giuliana Zanchi, 2016: Modelling nutrient transport from forest ecosystems to surface waters.

Florian Sallaba, 2016: Biophysical and human controls of land productivity under global change: development and demonstration of parsimonious modelling techniques.

Norbert Pirk, 2017: Tundra meets atmosphere: seasonal dynamics of trace gas exchange in the High Arctic.

Minchao Wu, 2017: Land-atmosphere interactions and regional Earth system dynamics due to natural and anthropogenic vegetation changes.

Niklas Boke-Olén, 2017: Global savannah phenology: integrating earth observation, ecosystem modelling, and PhenoCams.

Abdulhakim M. Abdi, 2017: Primary production in African drylands: quantifying supply and demand using earth observation and socio-ecological data.

Nitin Chaudhary, 2017: Peatland dynamics in response to past and potential future climate change.

Ylva van Meeningen, 2017: Is genetic diversity more important for terpene emissions than latitudinal adaptation?: using genetically identical trees to better understand emission fluctuations across a European latitudinal gradient.

Patrik Vestin, 2017: Effects of forest management on greenhouse gas fluxes in a boreal forest.

Mohammadreza Rajabi, 2017: Spatial modeling and simulation for disease surveillance.

Jan Blanke, 2018: European ecosystems on a changing planet: integrating climate change and land-use intensity data.

Min Wang, 2018: Characteristics of BVOC emissions from a Swedish boreal forest: using chambers to capture biogenic volatile organic compounds (BVOCs) from trees and forest floor.

Wilhelm Dubber, 2018: Natural and social dimensions of forest carbon accounting.

Emma Johansson, 2018: Large-Scale Land Acquisitions as a Driver of Socio-Environmental Change: From the Pixel to the Globe.

Helen Eriksson, 2018: Harmonisation of geographic data: between geographic levels, hierarcic structures and over time. (Lic.)

Zhendong Wu, 2018: Modelling the terrestrial carbon cycle: drivers, benchmarks, and model-data fusion.

Zhanzhang Cai, 2019: Vegetation observation in the big data era: Sentinel-2 data for mapping the seasonality of land vegetation.

Fabien Rizinjirabake, 2019: Dissolved organic carbon in tropical watersheds: linking field observation and ecohydrological modelling.

Jeppe Agård Kristensen, 2019: Biogeochemistry in Subarctic birch forests: perspectives on insect herbivory.

Yanzi Yan, 2019: The role of hydrological cycle in forest ecosystems: flow path, nutrient cycling and water-carbon interaction.

George Oriangi, 2019: Urban resilience to climate change shocks and stresses in Mbale municipality in Eastern Uganda.

Alex Paulo Lubida, 2019: Investigating spatial data infrastructure planning in Tanzania using system modelling and social concepts.

Weiming Huang, 2020: Geospatial data and knowledge on the Web: knowledge-based geospatial data integration and visualisation with Semantic Web technologies.