# Using Geospatial Semantic Web for Exploring Geographic Knowledge in Medieval Manuscripts

Nicolò Pratelli[1,2][0000−0003−0364−922X] and Valentina Bartalesi[1][0000−0001−9024−0822]

[1] Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR), Pisa, Italy
{nicolo.pratelli,valentina.bartalesi}@isti.cnr.it
[2] Department of Information Engineering (DII), University of Pisa, Pisa, Italy

**Abstract.** This paper explores the capabilities of the Geospatial Semantic Web to support scholars in studying the geographic knowledge included in medieval and Renaissance works. In the context of the Italian national research project IMAGO, we developed a CRM-based ontology that aligns with the Open Geospatial Consortium (OCG) GeoSPARQL standard. The ontology enables geospatial queries on the IMAGO knowledge graph. The results of these queries, as detailed in this paper, demonstrate the effectiveness of this approach in representing the geospatial data and in inferring new knowledge. For example, using this approach, we are able to identify all the works that mention places in a specific region, or by combining geographic knowledge with knowledge about the literary genre of the works, we can identify authors who travelled to a particular territory, such as the Holy Land. Furthermore, combining temporal and geospatial information enables us to discover places within a particular territory mentioned in manuscripts of a specific century. These examples demonstrate the potential of the Geospatial Semantic Web approach to uncover previously hidden connections and enrich our understanding of historical and geographical data.

**Keywords:** Semantic Web · GeoSPARQL · Medieval Manuscript · Digital Humanities.

## 1 Introduction

The representation of the world, technically known as *imago mundi*, crafted during the Middle Ages and Renaissance, offers a rich and intricate research domain for exploring the history knowledge of the *Ecumene* - i.e., the known, inhabited, or habitable world -, the formation of cultural and national identities, and the evolution of contemporary geographical perspectives. When examined comprehensively, the Middle Ages and Renaissance eras unveil a diverse literary production comprising various works depicting various countries, regions, and urban centres. This literary production encompasses a range of genres, including universal geographic surveys, personal travel diaries, *itineraria*, and detailed geographical essays. Some examples of these different genres are respectively the *Cosmographia* written by Pope Pius II, born Enea Silvio Bartolomeo Piccolomini (1405–1464), the Chronica Major by the English Benedictine monk Mathew Paris (1200–1259), and the *Italia illustrata* written by the Italian Renaissance humanist Blondus Flavius (1392–1463).

The Italian national research project Index Medii Aevi Geographiae Operum (IMAGO)[3] (2020-2024) aims to provide a systematic overview of the Medieval and Renaissance geographic Latin literature exploiting the Semantic Web technologies. Until now, this literature has not been studied using digital methods. The project's final aim is to create a Web application allowing scholars to freely access and visualise the data collected in the IMAGO knowledge base and to publish this knowledge as Linked Open Data (LOD) [5]. In particular, the project wants to provide (i) a collection of the manuscript tradition and printed editions for each work; (ii) a classification of authors, genres and contents; (iii) a collection of critical editions of some more representative works; (iv) a Medieval Latin toponym index.

In this paper, we present ongoing research focused on exploring how we can improve the study of the places mentioned in geographic medieval and Renaissance works using the Semantic Web [6], and particularly the Geospatial Semantic Web [37, 22]. The Geospatial Semantic Web is an extension of the Semantic Web that adds explicit meaning to geospatial data using geospatial ontologies. Geospatial data is unique because it involves specific elements like geometries, a coordinate reference system, and the spatial relationships between geometries, all of which must be carefully considered when encoding them into formats that machines can understand and process. We introduce the IMAGO ontology, a CRM-based vocabulary designed to represent the data related to ancient geographic works gathered by scholars in medieval Latin literature, comprising the information on the mentioned places. We ensure ontology interoperability by adhering to established standards like the ISO standard CIDOC CRM [10] and its extension CRMgeo [14] and the Open Geospatial Consortium (OGC) GeoSPARQL standard[4]. By creating an OWL Knowledge Graph (KG) based on this ontology and applying a semantic reasoner [20] and GeoSPARQL queries on the KG, we can support scholars in inferring new geographical knowledge about the mentioned places, analysing correlations among the places cited in different works and among the places and the other gathered data, and having a complete view of the geospatial knowledge included in this kind of literature.

## 2 Related works

The Geospatial Semantic Web provides a framework that uses different vocabularies to standardize geospatial data. As an example, stRDF [16] and its query language stSPARQL [17] have been developed to deal with time-varying geospatial data on top of RDF[27], which is a W3C standard model for expressing data. The Open Geospatial Consortium (OGC) has a counterpart proposal

[3] https://imagoarchive.it/en/index.html

called GeoSPARQL [4], also based on SPARQL[28] for representing and querying geospatial Linked Data in RDF. Although not fully standardized in any single language, qualitative reasoning models like DE-9IM [33] and RCC8 [15], which describe topological relationships between spatial entities, have been partially standardized by the OGC GeoSPARQL.

Various formal models have been proposed for mapping CIDOC CRM to GeoSPARQL. The 2018 paper by Nys et al. [23] suggests a hybrid ontology that allows for concurrent spatial and temporal handling, incorporating GeoSPARQL and OWL-Time into CIDOC CRM. In 2021, Ducatteeuw [12] proposed an urban gazetteer using CIDOC CRM as a top-level ontology, aiming to adhere closely to international gazetteer standards for interoperability with other gazetteer datasets. Lastly, Ranjgar et al. in 2022 [31] developed a POI-based data model for heritage sites in Iran using concepts from CIDOC CRM integrated with GeoSPARQL.

Among these models, CRMgeo [14] has been proposed for approval by CIDOC CRM Special Interest Group (SIG) as an official compatible model that allows the integration of GeoSPARQL and CIDOC CRM. Specifically, CRMgeo aims to integrate geoinformation using the conceptualizations, formal definitions, encoding standards, and topological relations defined by the Open Geospatial Consortium.

## 3 Methodology

As the initial step toward building tools that assist scholars in creating, updating, and consulting a knowledge graph (KG) of medieval and Renaissance geographical works, we constructed the IMAGO ontology[4], a vocabulary [19] that formally represents the knowledge about this literature. The IMAGO ontology [2] is derived from a conceptualisation of the knowledge domain that identifies the categories and the relations that have to be formally represented. This conceptualisation results from a strict collaboration between CNR and the experts in Latin and Italian geographic literature involved in the project. In particular, it is based on authoritative studies of the Medieval and Renaissance Latin geographic works, i.e., [29], [30], [21], [18], [24], [9], [7]. To promote reusability and enhance interoperability, we designed our ontology as an extension of three well-known reference ontologies: the CIDOC CRM vocabulary, FRBRoo [11], including its ongoing reformulation LRMoo [32], and CRMgeo.

Additionally, we designed and developed a web-based tool that allows scholars to populate the ontology with relevant data. In alignment with the Semantic Web paradigm, the tool allows representing every resource stored in the IMAGO KG using a unique identifier, known as an Internationalized Resource Identifier (IRI)[5]. These IRIs are automatically obtained from various open-source datasets available on the web, which are then made available to the scholar through the tool. Specifically, we drew IRIs from the following sources: (i) the

[4] https://imagoarchive.it/en/doc/index-en.html
[5] https://www.w3.org/International/articles/idn-and-iri/

Wikidata knowledge base [36] for authors, works, libraries and locations; (ii) the MIRABILE digital archive[6] for authors and works; and (iii) the Pleiades gazetteer[7] for ancient places; (iv) the Nuovo Soggettario[8] thesaurus for the literary genres. If the IRIs were absent in these datasets, we automatically generated and assigned custom IRIs to the resources.

We developed a Python script to add spatial representations to the Wikidata *location*-type resources as a data augmentation step. The process verified whether each valid *location*-type resource had an associated coordinate pair in its corresponding Wikidata entry. Specifically, it extracted the content of the Wikidata "coordinate location" property (P625) as a reference longitude-latitude coordinate pair. Additionally, the process examined whether a polygon might be linked to the entity. To achieve this goal, our script uses an instance of the open-access QLever endpoint provided by the University of Freiburg [35] to retrieve potential polygon representations from the OpenStreetMap subgraph included in this extensive knowledge graph [3]. Our script returned all geometries found through the QLever service as Well-Known Text (WKT) formatted strings [25].

Once the data about the works are inserted through the tool and the augmentation step is completed, this knowledge is encoded as an OWL graph and stored in a triple store. The overall process is described in Figure 1.

The data collected by the tool is exported as a JSON object. Indeed, our software uses a JSON schema to represent the data [1], structured according to the IMAGO ontology classes. The JSON object is processed by a Java software, which transforms it into an OWL graph encoded in RDF/XML and Turtle formats. This software performs its task by relying on the Apache Jena library[9]. The graph is finally stored in a GeoSPARQL Fuseki triple store [34], and it can be queried through a SPARQL endpoint[10]. We employed a semantic reasoner to ensure the logical consistency of our entire OWL graph. We used the Openllet open-source semantic reasoner [26] to (i) validate the consistency of our OWL graph, (ii) verify that the class hierarchy aligns with that of the IMAGO Ontology, (iii) check the consistency of geometric data, and (iv) test the execution of complex SPARQL and GeoSPARQL queries. Openllet confirmed the consistency of our knowledge graph and verified that the subclass relationships and the overall class hierarchy conformed to the structure outlined by the IMAGO Ontology. Additionally, all geometric data passed the consistency tests. Finally, GeoSPARQL queries enabled spatial reasoning and supported all algebraic operations among geographic data collected in the KG.

[6] www.mirabileweb.it
[7] https://pleiades.stoa.org/
[8] https://thes.bncf.firenze.sbn.it/index_eng.html
[9] https://jena.apache.org/
[10] https://imagoarchive.it/geosparql/

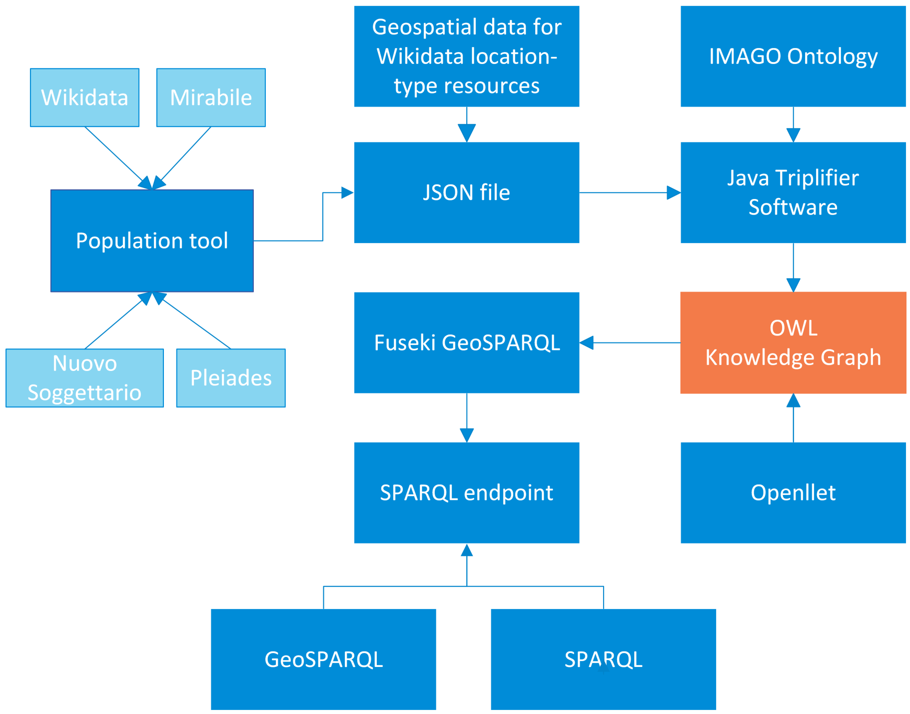


**Fig. 1.** The software architecture to create and query the IMAGO knowledge base. We highlighted the OWL knowledge graph in orange. The knowledge bases we used as external sources are highlighted in light blue.

## 4 Geospatial Reasoning to Infer New Knowledge

Based on discussions with the scholars involved in the project, we identified four key research questions that experts would address. The questions are as follows:

- Which works mention places in a particular area (e.g., in France or along the Black Sea coasts)?
- What are the places located near ancient roads or pilgrimage routes?
- What places in a particular territory are mentioned in manuscripts written in a specific century?
- Who are the authors who visited places within a specific territory?

The answers to these questions are not directly available from the collected data. However, by applying our methodology, we can infer new knowledge from the data formally represented in the KG and thus address these questions.

We defined three geoSPARQL queries that answer some particular cases related to each of the key questions reported above. In particular, we implemented the following queries:

1. The works that mention places located in France (Q1)
2. The places located within a 0.2-degree buffer around the Via Francigena (Q2)
3. The places in Italy that are mentioned in manuscripts written in the fifteenth century (Q3)
4. The authors who have visited the Holy Land (Q4)

An IMAGO expert assisted us in manually verifying the correctness of the retrieved information, ensuring the reliability of the query results.

In the following, we explain the queries that were performed. The SPARQL code of the queries and the corresponding results are available on our GitHub repository[11].

### The works that mention places located in France (Q1)

The query retrieves the work titles and authors along with the toponyms the works mention. The `FROM` clause specifies the graph that is queried. Several ontology prefixes are specified at the beginning of the query to shorten the corresponding IRIs, which are used in the subsequent parts of the query. The `FROM` clause and the prefix declaration are equal for all the queries reported in this paper. The `WHERE` clause contains the conditions that need to be satisfied for each result. It involves several semantic-triple patterns connected by the "." or ";" operators. In this clause, we retrieve the work titles, authors, toponyms, and the polygons of the places identified by the toponyms. A nested `SELECT` statement allows retrieving the WKT geometry of France (Q142) from the QLever SPARQL server of the University of Freiburg. Finally, the `FILTER` clause selects the places included within the France polygon.

### The places located within a 0.2-degree buffer around the Via Francigena (Q2)

The query retrieves the work titles and authors along with the toponyms the works mention. In the `WHERE` clause, we retrieve the work titles and authors and the toponyms as well as the polygons of the places identified by the toponyms. Furthermore, we set the value of the work literary genre equal to *itineraria*. Indeed, we think that a study on the knowledge of places located near the via Francigena is more significant if conducted on works belonging to the genres of travel literature. The Via Francigena is an ancient road and pilgrimage route running from Canterbury in England, through France and Switzerland, to Rome and then to Apulia, Italy, where there were ports of embarkation for the Holy Land. *Itineraria* genre has a unique identifier (100021) that came from a literary genre thesaurus built by the IMAGO scholars based on the subject indexing tool Nuovo Soggettario [8]. Finally, the `FILTER` clause selects the places that are included within a buffer of 0.2 degrees created around the Francigena polygon, specified in the `BIND` operator.

[11] https://l.cnr.it/6eyzp

**The places in Italy that are mentioned in works contained in manuscripts written in the fifteenth century (Q3)**

The query retrieves the toponyms and the corresponding works, and authors, in which these places are mentioned. In the `WHERE` clause, the polygons of the places identified by the toponyms are retrieved. Simultaneously, information regarding manuscripts is retrieved, including the production dates. Subsequently, the time range is established using the `FILTER` operator. A nested `SELECT` statement allows retrieving the WKT geometry of Italy (Q38) from the QLever SPARQL server of the University of Freiburg. Finally, another `FILTER` clause selects the places included within Italy's polygon.

**The authors who have visited the Holy Land (Q4)**

The query retrieves the authors who wrote works in which they tell their journeys in the Holy Land. In the `WHERE` clause the polygons of the places identified by the toponyms are retrieved only for the work belonging to the literary genre "personal travel diaries". As the Q2, this genre has a unique identifier (100026) that came from a literary genre thesaurus built by the IMAGO scholars. A nested `SELECT` statement allows retrieving the coordinates (longitude and latitude) of the Holy Land (Q142) from the Wikidata SPARQL server. Finally, the `FILTER` clause selects the places that are included within a buffer of 0.3 degrees created around the Holy Land coordinates.

## 5 Discussion

The Geospatial Semantic Web brings several functionalities, including geographic analysis, spatio-temporal queries, and the ability to interlink with other data sources to add geospatial context to a given subject. These capabilities distinguish the Geospatial Semantic Web from the standard Semantic Web, which is not typically equipped to handle complex geospatial data [38, 13]. The ontology we created within the IMAGO project aligns with the OGC GeoSPARQL standard, enabling geospatial queries on the IMAGO knowledge graph. The results of the queries we presented in this paper demonstrated their efficacy in representing all the gathered data and in inferring new knowledge, which is particularly relevant for an in-depth analysis of the geospatial data represented in medieval and Renaissance works. Applying the semantic reasoner Openllet on GeoSPARQL data representation, we are able to infer new geospatial knowledge, which is not explicitly represented in the gathered data. For example, we are able to identify all the works that mentioned places located in a specific territory. Furthermore, since our KG incorporates geospatial data, general concepts, and temporal information, we can link these different types of data to provide geospatial and spatio-temporal context to the gathered knowledge. For example, combining geographic knowledge with information about the literary genre allows us to identify the authors who visited a particular territory, e.g. the Holy Land. We can also combine temporal and geospatial knowledge, for example, to find the places in a specific area mentioned in the manuscripts written in a particular century. These results show that the Geospatial Semantic Web can significantly enhance the exploration of historical and geographical data, providing a valuable resource for studying ancient literature.

## 6 Conclusions

This paper presents the first results of ongoing research focused on using the Geospatial Semantic Web technologies to represent and query a collection of data related to medieval and Renaissance geographic Latin works created within the Italian national research project IMAGO. The use of Semantic Web technologies to formally represent geographic data, combined with GeoSPARQL and a semantic reasoner to query and process these data, enables the inference of new knowledge of significant interest to scholars in medieval and Renaissance geographical literature. According to the scholars involved in the project, we identified four key research questions that experts would address, and we implemented the corresponding GeoSPARQL queries that are able to answer these questions, inferring new knowledge. This approach allows researchers to explore connections and spatio-temporal relationships within the data, providing valuable insights into historical contexts and enhancing our understanding of geographical concepts from those periods.

## References

1. Bartalesi, V., Pratelli, N., Lenzi, E.: Json schema of the internal imago tool data representation (2022), https://imagoarchive.it/imago_json.html
2. Bartalesi, V., Pratelli, N., Lenzi, E., Pontari, P.: Using semantic web to create and explore an index of toponyms cited in medieval geographical works. ACM Journal on Computing and Cultural Heritage **16**(2), 1–18 (2023)
3. Bast, H., Buchhold, B.: Qlever: A query engine for efficient sparql+ text search. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 647–656 (2017)
4. Battle, R., Kolas, D.: Geosparql: enabling a geospatial semantic web. Semantic Web Journal **3**(4), 355–370 (2011)
5. Bauer, F., Kaltenböck, M.: Linked open data: The essentials. Edition mono/monochrom, Vienna **710** (2011)
6. Bizer, C., Heath, T., Berners-Lee, T.: Linked data: The story so far. In: Semantic services, interoperability and web applications: emerging concepts, pp. 205–227. IGI global (2011)
7. Bouloux, N.: Culture et savoirs géographiques dans l'Italie du XIVe siècle. Ph.D. thesis, Paris 1 (1999)
8. Cheti, A., Viti, E.: Functionality and merits of a faceted thesaurus: The case of the nuovo soggettario. Cataloging & Classification Quarterly **61**(5-6), 708–733 (2023)

9. Defilippis, D.: Da flavio biondo a leandro alberti. corografia e antiquaria tra quattro e cinquecento, atti del convegno di studi (foggia, 2 febbraio 2006), a cura di d. defilippis (2009)
10. Doerr, M.: The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine **24**(3), 75–75 (2003). https://doi.org/https://doi.org/10.1609/aimag.v24i3.1720
11. Doerr, M., Bekiari, C., LeBoeuf, P., nationale de France, B.: Frbroo, a conceptual model for performing arts. In: Proceedings of the 2008 Annual Conference of CIDOC, Athens. pp. 15–18 (2008)
12. Ducatteeuw, V.: Developing an Urban Gazetteer: A Semantic Web Database for Humanities Data. In: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities. pp. 36–39. GeoHumanities '21, Association for Computing Machinery, New York, NY, USA (Nov 2021). https://doi.org/10.1145/3486187.3490204
13. Egenhofer, M.J.: Toward the semantic geospatial web. In: Proceedings of the 10th ACM international symposium on Advances in geographic information systems. pp. 1–4 (2002)
14. Hiebel, G., Doerr, M., Eide, Ø.: Crmgeo: A spatiotemporal extension of cidoc-crm. International Journal on Digital Libraries **18**, 271–279 (2017)
15. Kontchakov, R., Pratt-Hartmann, I., Zakharyaschev, M.: Spatial reasoning with rcc8 and connectedness constraints in euclidean spaces. Artificial Intelligence **217**, 43–75 (2014)
16. Koubarakis, M., Kyzirakos, K.: Modeling and querying metadata in the semantic sensor web: The model strdf and the query language stsparql. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) The Semantic Web: Research and Applications. pp. 425–439. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
17. Koubarakis, M., Kyzirakos, K.: Modeling and Querying Metadata in the Semantic Sensor Web: The Model stRDF and the Query Language stSPARQL. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) The Semantic Web: Research and Applications. pp. 425–439. Springer, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13486-9_29
18. Kristeller, P.O.: Iter Italicum: Alia itinera 1. Australia to Germany, vol. 3. Warburg Institute (1963)
19. Krötzsch, M.: Owl 2 profiles: An introduction to lightweight ontology languages. In: Reasoning Web International Summer School. pp. 112–183. Springer (2012). https://doi.org/https://doi.org/10.1007/978-3-642-33158-9_4
20. Maarala, A.I., Su, X., Riekki, J.: Semantic reasoning for context-aware internet of things applications. IEEE Internet of Things Journal **4**(2), 461–473 (2016)
21. Menestò, E.: Relazioni di viaggi e di ambasciatori. Cavallo, Guglielmo; Leonardi, Claudio; Menestò, Enrico (dir. da), Lo spazio letterario del Medioevo **1**, 535–600 (1993)
22. Nishanbaev, I., Champion, E., McMeekin, D.A.: A survey of geospatial semantic web for cultural heritage. Heritage **2**(2), 1471–1498 (2019)
23. Nys, G.A., Van Ruymbeke, M., Billen, R.: Spatio-temporal reasoning in CIDOC CRM: An hybrid ontology with GeoSPARQL and OWL-Time. In: CEUR Workshop Proceedings. vol. 2230. RWTH Aachen University, Aachen, Germany (Oct 2018)
24. Odoric, F.: The travels of friar odoric. Edited by Henry Yule and Paolo Chiesa. Translated (2002)

25. Open Geospatial Consortium: Well-known text representation of coordinate reference systems. https://www.ogc.org/standard/wkt-crs/ (2024)
26. Openllet: Openllet: An Open Source OWL DL reasoner for Java. https://github.com/Galigator/openllet (2023)
27. Pan, J.Z.: Resource description framework. In: Handbook on ontologies, pp. 71–90. Springer (2009)
28. Pérez, J., Arenas, M., Gutierrez, C.: Semantics and complexity of sparql. ACM Transactions on Database Systems (TODS) **34**(3), 1–45 (2009)
29. Pontari, P.: «nedum mille qui effluxerunt annorum gesta sciamus». l'italia di biondo e l'invenzione del medioevo (2016)
30. Potthast, A.: Repertorium fontium historiae medii aevi (1970)
31. Ranjgar, B., Sadeghi-Niaraki, A., Shakeri, M., Choi, S.M.: An ontological data model for points of interest (POI) in a cultural heritage site. Heritage Science **10**(1), 13 (Mar 2022). https://doi.org/10.1186/s40494-021-00635-9
32. Riva, P., Žumer, M.: Frbroo, the ifla library reference model, and now lrmoo: a circle of development (2017)
33. Strobl, C.: Dimensionally extended nine-intersection model (de-9im) (2008)
34. The Apache Software Foundation: GeoSPARQL Fuseki. https://jena.apache.org/documentation/geosparql/geosparql-fuseki.html (2024)
35. University of Freiburg: QLever instance. https://qlever.cs.uni-freiburg.de/ (2024)
36. Vrandecic, D.: The rise of wikidata. IEEE Intelligent Systems **28**(4), 90–95 (2013). https://doi.org/10.1109/MIS.2013.119
37. Zhang, C., Zhao, T., Li, W.: Geospatial semantic web. Springer (2015)
38. Zhang, C., Zhao, T., Li, W., Zhang, C., Zhao, T., Li, W.: Conceptual frameworks of geospatial semantic web. Geospatial Semantic Web pp. 35–56 (2015)