Hochschule Mainz – University of Applied Sciences; Université de Bourgogne

INTEGRATION, QUALITY ASSURANCE AND USAGE OF GEOSPATIAL DATA WITH SEMANTIC TOOLS

Timo Homburg, Claire Prudhomme, Frank Boochs, Ana Roxin, Christophe Cruz

Abstract: In this article we want to present an integrational approach of geospatial data into the semantic web in the context of the semantic GIS project. We first highlight the purpose and advantages of the integration and interpretation of data into the semantic web and further on describe the process of data acquisition, data interpretation, quality assurance and provenance and how to access the so integrated data. We continue to highlight the advantages of this integration method by presenting two fields of application of our research project: The evaluation of OpenStreetMap data and the improvement of disaster management. We conclude the article by giving prospects of future work in our project.

Keywords: Geospatial data, linked data, natural language processing, ontology, R2RML, SDI, semantic web, semantification, data quality, provenance

INTEGRATION, BEWERTUNG UND NUTZUNG HETEROGENER DATENQUELLEN MITTELS SEMANTISCHER WERKZEUGE

Zusammenfassung: In diesem Beitrag stellen wir die Integration von Geodaten in einen Semantic-Web-Kontext in Rahmen unseres Projekts Semantic GIS vor. Zunächst möchten wir den Zweck und die Vorteile einer Integration und Interpretation von Daten in das Semantic Web beleuchten und anschließend unseren Integrationsprozess, bestehend aus Datengewinnung, automatischer Interpretation, Qualitätssicherung und Provenienz sowie den Datenzugriff, erklären. Um die Anwendung unserer Forschung zu demonstrieren, gehen wir auf zwei Anwendungsfälle in unserem Projekt ein: Die Bewertung von OpenStreetMap-Daten und die Verbesserung des Katastrophenschutzes mittels semantischem Reasoning. Wir schließen den Beitrag mit einem Fazit sowie einem kurzen Ausblick auf die zukünftige Forschung.

Schlüsselwörter: Geodaten, Linked Data, Sprachverarbeitung natürlicher Sprachen, Ontologie, R2RML, SDI, semantisches Web, Semantifikation, Datenqualität, Provenienz

Authors

M. Sc. Timo Homburg
M. Sc. Claire Prudhomme
Prof. Dr.-Ing. Frank Boochs
Hochschule Mainz – University of Applied Sciences
i3mainz – Institute for Spatial Information and Surveying Technology
Lucy-Hillebrand-Straße 2
D-55128 Mainz
E: timo.homburg@hs-mainz.de
claire.prudhomme@hs-mainz.de
frank.boochs@hs-mainz.de

Dr. Ana Roxin
Dr. Christophe Cruz
Université de Bourgogne
9 avenue Alain Savary
F-21000 Dijon
E: ana-maria.roxin@u-bourgogne.fr
christophe.cruz@u-bourgogne.fr

1 INTRODUCTION

Integration of heterogeneous datasets is a persisting problem in geographical computer science. Many classical GIS approaches exist making use of relational databases to achieve a tailor-made integration of geospatial data according to the needs of the current task. In the SemGIS project we are aiming at integrating heterogenous geodatasets into a semantic web environment to take advantage of the flexibility of semantic data structures and to access a variety of related datasets that are already available in the semantic web. We intend to use the so-formed geospatial knowledge base in the application field of disaster management in order to predict, mitigate or simplify decision making in an event of a disaster. As in our project we are possibly facing a large number of heterogeneous geodatasets of which we often do not know the origin nor intention nor the author and therefore lack an appropriate domain expert to help us understand data fields, we as non-domain experts would be left with a manual integration approach of said data. Dataset descriptions, if available, are often in natural language only which may give us hints but are hard to process in general and contain often hard to resolve ambiguities. However, despite mentioned obstacles we believe that a at least rudimentary classification and interlinking of our given data sets by means of the data values and data descriptions, is feasible. In addition, depending on the data source, data quality metrics as well as provenance information can be added to the to-be-imported data sets and change the way the data is treated not only for the geospatial community but also for the semantic web community. In this article we want to describe our approach to automatically find, process, analyse, interlink and quality-assure geospatial data sets on the web in the context of our project.

2 STATE OF THE ART

The geospatial web provides several standards to distribute geospatial data. Since several years it is possible to publish geospatial data with the help of OGC webservices and to categorize said data using OGC catalogue web services (Nogueras-Iso et al. 2005). Despite this fact the access to geospatial data is very limited because it is not possible to search for geospatial data by means of their features and semantics and to make queries over geospatial data in the web on a large scale. This is due to several persisting problems in the publication of geospatial data:

- The scope of data is not semantically accessible using machines.
- Geospatial data is not thematically clustered in the web of data.
- Lack of a dedicated search engine for geospatial data.
- Geospatial data is hard to index because of its heterogeneity.
- Several non-quality annotated data sets depicting the same geometry and/or meaning and varying features might be found in the geospatial web.
- Publications in the form of map APIs like OpenStreetMap are if semantically interpretable only to a certain extent and to a limited amount of knowledge domains.

We can conclude that there is no geospatial search engine nor a unified query interface being able to assess semantically interpreted and quality-assured geospatial webservices on a large scale. In addition many geospatial resources on the web are not even published as webservices or map APIs but in a variety of different formats and/or APIs (e.g. GeoTIFF, KML, GeoJSON, CovJSON) which in many cases are poorly documented and have often neither a quality assessment nor sufficient metadata about its source of origin.

2.1 DATA ACQUISITION

To find thematic data in the geospatial web, traditional approaches are to find an appropriate CSW service which lists appropriate data sources that correspond to the description of the metadata or to its keywords. Recently, approaches to discover and link the geospatial web of data provided by services according to their keyword and service descriptions has been conducted (Pellicer 2011). However, an automated analysis of features and their values was not provided in this work. In our work we would like to overcome this limitation or at least to test how far this limitation can be overcome in the geospatial domain.

2.2 SEMANTIC INTERPRETABILITY AND ACCESSIBILITY

Related work in the interpretation of geospatial data has been conducted for OpenStreetMap by the LinkedGeoData Project (Stadler et al. 2012). Concepts of tags of OpenStreetMap data have been automatically generated and a virtual GeoSPARQL (Perry & Herring 2012) interface accessing OSM data in real time has been created. GeoSPARQL itself as an OGC standard provides us with a standardized method to access semantically interpreted geospatial data, so that in theory, the foundations to create a unified endpoint to search for features and types of geospatial data have been in place since its introduction in 2012.

2.3 PROVENANCE AND DATA QUALITY

Provenance and data quality is of concern to the semantic web community as seen in Mendes et al. (2012) because the semantic web typically lacks such information appended to its knowledge bases. Semantic web data are typically published without a rich provenance hierarchy and from various institutions without a record of trust and a history of how and in which quality the data has been gathered. However, concepts of which parameters to consider and which provenance information to gather can be found in respective standardized ontologies among others by W3C (Lebo et al 2013, Hartig 2009, Fürber, Hepp 2011). In geospatial research data quality is defined in various standards such as INSPIRE and in the respective literature (Shi et al. 2003, Redman 2001), which are useful in their respective applications. By integrating various kinds of heterogeneous data, we intend to use as many quality criteria from those standards as possible to give end-users the possibility to choose among the criteria they deem fitting best.

3 SEMGIS PROJECT

In this section we describe how we implement the aforementioned steps of interpretation in the SemGIS project.

ID	the_geom	Feature1	Feature 2	... FeatureN
Example.1	POINT(..)	123	"ExampleString"	3.4

Table 1: Example file "example" represented as a database table

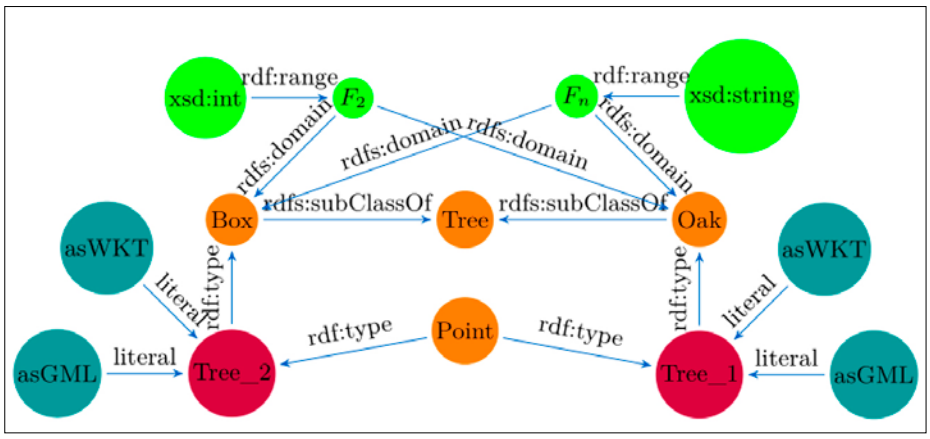

Figure 1: Local ontology example as referenced in Homburg et al. (2016)

3.1 DATA AQUISITION AND DISCOVERY

To discover potential geospatial data in the SemGIS project we plan to develop a webcrawler similar to Pellicer (2011) which uses search engines and given addresses of geoportals to crawl for the data we would like to integrate. Along with the results we also allow lists of resources provided by users. We are prepared to integrate the following resources:

- Web services (WFS, WCS, WMS, CSW, SOS);
- Spatial Databases (PostGIS, Oracle Spatial);
- OGC data formats (SHP, KML, GML and dialects,GeoJSON,GeoTIFF);
- Using Geoparsing to make sense of geotagged web pages and/or web APIs.

Once an appropriate resource has been found, the metadata found along with the resource has to be gathered as well if it is existent. For OGC webservices we can usually rely on given metadata standards of the web service definition itself. Metadata for files might be stored along with them or on the surrounding homepage as for example in CKAN-based geoportals.

3.2 DATA INTERPRETATION

Once a list of suitable geospatial resources has been gathered from the web or has been provided by the end-user, the datasets need to be interpreted to link them to semantic web concepts. Every aforementioned data source can be seen as one or many relational database tables, which we depict generically as shown in table 1.

When interpreting data from a spatial database, foreign keys aka. relations between existing database tables can be considered and extracted using established methods like R2RML (Das et al. 2012). Geodatasets in the form of files represent single relational database tables of which it is typically unknown which relations to other data sets exist. It is on this premise that we employ interpretation algorithms to create such relations to semantic web concepts. The goal of such efforts is to produce a so-called local ontology (figure 1) of each resource with induced links to other semantic web resources and concepts.

3.2.1 CONCEPT MATCHING PROCESS

The information we would like to extract from a single database table data set includes

- At least one concept for the whole data set;
- At best one or more concepts per column of the database table;
- At best several additional attributes using additional knowledge sources e.g. geocoding information.

We can extract at least one concept for the whole data set by analyzing its filename/database table name or by using a reference data set of geometries e.g. OpenStreetMap/Linked Geodata to find classifications of geometries in the vicinity of the geometries provided in the data set. Fitting concepts for columns can be found by either analyzing the columns' title and/or the values of the column using Natural Language Processing algorithms. To do that we rely on BabelNet (Navigli & Ponzetto 2010), a multilingual network, partially connected to the semantic web and on the labels of ontologies we would like to link to e.g. DBPedia (Lehmann et al. 2015) or Wikidata (Vrandecič & Krötzsch 2014). Using a rudimentary detection of the language the dataset is written in we can analyze column titles and values for existing terms that we can subsequently match in the mentioned ontologies. Having sufficiently many values of a similar classification allows a generalization of a concept description for the respective column. We are therefore able to detect at best one concept per column and if modeled the instance of each value present in the corresponding dataset. By doing further linguistic analysis we are able to detect the role of each column, which might be:

- A foreign key corresponding to an ObjectProperty in the semantic web.
- A DataTypeProperty corresponding to a simple value.
- An AnnotationProperty corresponding to metadata annotated to an instance or class in our knowledge base.

Using this additional information we further interpret and distinguish columns by the following categories:

- Address columns: Columns that represent components of an address matchable with traditional geocoding.
- SubClass columns: Columns including nouns that represent a sub-categorization of the database tables content.
- Object Property Columns: Columns including adjectives that represent a categorization of a relation or an attribute of the data set.
- Common regular expression columns: Columns that can be associated by executing a common regular expression on its values (e.g. email addresses, UUIDs).
- Label and comment columns: Columns that represent a description of one row (= instance of the data set).
- Unit columns: Columns containing numbers which have an identifiable unit and/or concept when analyzing the column description.

We are currently not able to analyze remaining column types so that they remain in the system as associated values in its primitive types (e.g. double, integer). The end-user is still able to access them, but the semantic meaning could not be determined automatically and is therefore not accessible if not corrected by a human being. For a more detailed description of the match-

ing process and its results, we refer to Prudhomme et al. (2017). Our goal is to improve the automated concept detection in further iterations of our software.

3.3 QUALITY AND PROVENANCE

Quality and provenance are important metadata which can enhance the value of a data set for its daily usage. The existence of quality and provenance parameters in the geospatial web and in the semantic web is often not standardized and not common. We therefore propose to extract and generate such parameters in all data sets we integrate into a common knowledge base.

3.3.1 PROVENANCE

Provenance parameters can tell us information about who was when providing which data using which process of data preparation and which original data source or measurements. Usually provenance information can be found along within accompanying metadata or on the homepage/service page where a particular data set has been published. Provenance information can be modeled using the Prov-O ontology defined by W3C (Lebo et al. 2013). Examples of such provenance information publication are as follows:

- Provenance information of a file: Creator, GPS measurement device, measurement method, date of creation, date of modification etc.
- Provenance information of the publishing institution: Name, email phone number etc.
- Provenance information of the publishing service: Domain, name, contact data, maintainer etc.

3.3.2 DATA QUALITY

The notion of data quality can be extended to various data quality dimensions. One definition of data quality could be

> Data quality is the degree to which data fulfills requirements.

Which requirements are important for the data we are working with depends on its use case. Every domain of knowledge can depend on various quality criteria. However, we can analyze as many quality criteria on our data as it is possible to prepare users to take qualified decisions about which data to use for their specific use case. In general, we categorize the goals associated with data quality in the following categories:

- Spatial data quality (Morrison 1995)
- Positional accuracy
- Completeness
- Logical consistency
- Semantic accuracy
- Semantic interpretability
- Temporal information
- Metadata quality
- Quality of service
- Open license/Cost of access

Examples of data quality parameters include:

- Positional accuracy of the geometry (with reference to a gold standard)
- Geocodability of the data set
- Amount of matchable attributes to a semantic concept
- Completeness of the dataset/attributes
- Completeness of metadata information and its verifiability
- Quality of service

We are hereby focusing on known data quality concepts from the semantic web, GIS research as well as data quality provided by the knowledge domains we are connected to through features.

3.3.3 EVALUATING PROVENANCE AND DATA QUALITY

When combining provenance information and quality parameters, datasets from specific resources can be associated with specific values of data quality. This allows not only to rank specific data sets but also to highlight data providers that are trustworthy because they have proven to provide data with a consistent data quality. If in doubt a reasoning system or the end user can take advantage of this information to choose the most trustworthy data set among several possible data sets for the fulfillment of his use case. In addition, other criteria to rank dataset of different quality evaluations can be considered such as high quality areas, high quality building types, ways or areas in which certain quality parameters are common as compared to areas in which the same parameters are not common at all.

3.4 DATA ACCESS AND REASONING

To access data we have imported using the process described in the previous sections we rely on a GeoSPARQL (Perry & Herring 2012) endpoint which allows us to use Egenhofer calculations in the semantic web. In addition we developed an extended vocabulary allowing us to use various PostGIS functions like geometry constructors to be used in GeoSPARQL. Queries that are often used or that lead to results that should be reused in a later stage of the development are standardized in so-called reasoning rules in languages like SWRL (Horrocks et al. 2004) or SPIN (Knublauch et al. 2009). At this stage of the project we are at the point of developing reasoning strategies together with our project partners. Therefore, first real world applications of reasoning are yet to be implemented in our research. To highlight a possible case of reasoning we refer to an example from Homburg et al. (2016) in which we highlighted the inference of nearest hospitals to a to-be-evacuated school as an example of automated reasoning in a disaster management case.

4 APPLICATIONS

The SemGIS project is aimed at applications in disaster management and energy. However correct application cases also require trustworthy and correct map data which needs to be ensured while executing the use case calculations or beforehand.

4.1 EVALUATION OF OPENSTREETMAP DATA

The largest repository of open geodata in the world is OpenStreetMap. It is used by various people around the world for many different purposes and is created by a vast amount of editors. To our knowledge a comprehensive analysis of the quality and provenance of OpenStreetMap data in Germany has not been undertaken yet. Therefore in our project, we would like to evaluate OpenStreetMap data by comparing them to the gold standard provided by the German national authorities for cartography and geodesy. By semantically interpreting and by extracting and adding provenance as well as quality information we can compare German official data to OpenStreetMap data in as many aspects as needed. We can highlight conflicts in a separate layer on top of OpenStreetMap, evaluate which parts of OpenStreetMap are good enough to serve for which use case we are aware of and can give hints to the OpenStreetMap community in which way to improve OpenStreetMap in the fu-

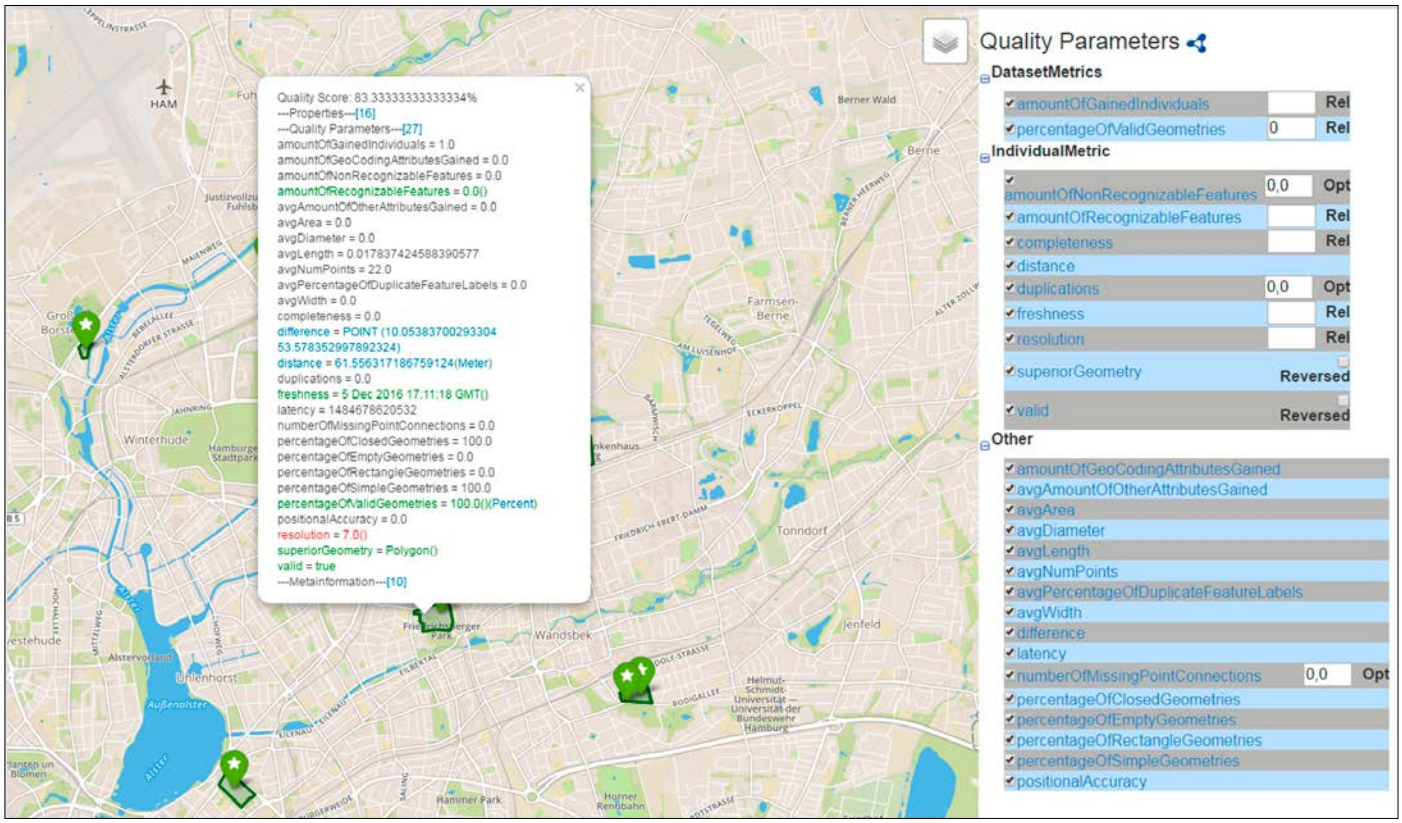

Figure 2: Preliminary quality comparison screen of OpenStreetMap data vs. open data

ture. A preliminary development of this approach can be seen in figure 2 in which we gather first statistics on corresponding geometries such as the completeness of the attributes in the dataset, the distance between such geometries, the interpretability of attributes, the validity of geometries and the information gain of merging the two geometry representations together. Further information is stored in order to recognize correlations between evaluated datasets, such as the average number of points, its resolution and meta information provided with the dataset. In the future this will allow for classifications of similar aspects of data and to create maps of areas of similar quality aspects as described in section 3.3.

While importing various amounts of data we also create a huge amount of quality-annotated data in the semantic web. This data serves the semantic web community which is becoming increasingly interested in geospatial topics. In the context of disaster management we ensure that the resources we use to do flood simulations are correct to the extent we need them, so that predictions of flood and the consequences thereof are accurate. Lastly comparing datasets of different quality helps us to consolidate different features that are present in the different datasets. By knowing quality and provenance requirements of the end-user the system can end up with a merged dataset of high quality and the maximum amount of features possible.

4.2 MULTI-AGENT NATURAL DISASTER SIMULATIONS

Disaster management consists of various steps that can be highlighted in the so-called disaster management cycle (Coppola 2011). During an event of disaster for example a flood, various actors need to cooperate in order to prevent further damages, evacuate people and rescue endangered areas. The efficiency of these activities depends on many elements which need to be prepared. Three of these elements are the resources needed for this activities, activity planning and common data set shared by all actors. Each of these elements have an impact in the disaster management response. The activity planning improves the organization and allows for knowing what you need to do according to the situation, and thus, to act quickly. Resources are key elements for the activity. If the resources are not enough to achieve the activities goal, the activity may be slowed down or even fail. The coordination between different actors becomes simplified when they are able to work with a common data set. A main problem in this field is that all actors do not have same rights and the same access of data. The identification of a data sets which could be used as a common data set for all actors (even if some of them can have more information) would be a good point for the coordination of the response activities. In order to assess these three elements, our project has aiming to simulate agents corresponding to real persons acting in a disaster event according to a rule-based system using gathered and interpreted data as described above our project. The simulation has aiming to support the preparation of disaster management response in assessing activity planning, resources and data sets.

5 CONCLUSION

Working towards a unified endpoint for semantically interpreted and quality assured geospatial data is a profitable approach for both the geospatial web of data as well as for the semantic web. In this article we

have shown our efforts on how to approach this goal and the progress we have achieved on the way. We have also shown how provenance and data quality parameters can be used in our system in the future to evaluate and append other sources of open data like OpenStreetMap or to act as a beneficial knowledge base for disaster management optimizations using multi-agent simulations. Our future work will continue on said use cases with our project partners and to investigate on how our concepts will help to improve the workflows of the several actors in disaster management.

ACKNOWLEDGEMENTS

The SemGIS project was funded by the German Federal Ministry of Education and Research under Project Reference: 03FH032IX4.

References

Coppola, D. P. (2011): Introduction to international disaster management. Elsevier, Amsterdam.

Das, S.; Sundara, S.; Cyganiak, R. (2012): R2RML: RDB to RDF Mapping Language. W3C Recommendation, 27 September 2012. World Wide Web Consortium (W3C) (www.w3.org/TR/r2rml).

Fürber, C.; Hepp, M. (2011): Towards a vocabulary for data quality management in semantic web architectures. In: Proceedings of the 1st international workshop on linked web data management, March 21 – 24, 2011, Uppsala, Sweden. ACM, New York, NY, pp. 1-8.

Hartig, O. (2009): Provenance information in the web of data. In: Proceedings of the Linked Data on the Web Workshop (LDOW) at WWW, April 20, 2009, Madrid, Spain.

Homburg, T.; Prudhomme, C.; Würriehausen, F.; Karmacharya, A.; Boochs, F.; Roxin, A.; Cruz, C. (2016): Interpreting heterogeneous geospatial data using semantic web technologies. In: International Conference on Computational Science and its Applications, July 4 – 7, 2016, Beijing, China, pp. 240-255.

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, May 21, 2004.

Knublauch, H.; Hendler, J.; Idehen, K. (2009): Spin-SPARQL inferencing notation. W3C Member Submission.

Lebo, T.; Sahoo, S.; McGuinness, D. (Eds.) (2013): PROV-O: The PROV ontology. W3C Recommendation 30 April 2013.

Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D. et al. (2015): DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. In: Semantic Web, 6 (2), pp. 167-195.

Mendes, P. N.; Mühleisen, H.; Bizer, C. (2012): Sieve: Linked data quality assessment and fusion. In: Proceedings of the 2012 joint edbt/icdt workshops. ACM, New York, NY, pp. 116-123). http://doi.acm.org/10.1145/2320765.2320803; doi:10.1145/2320765.2320803

Morrison, J. L. (1995): Spatial data quality. In: Elements of spatial data quality, 202, pp. 1-12.

Navigli, R.; Ponzetto, S. P. (2010): Babelnet: Building a very large multilingual semantic network. In: Proceedings of the 48th annual meeting of the association for computational linguistics, July 11 – 16, 2010, Uppsala, Sweden, pp. 216-225.

Nogueras-Iso, J.; Zarazaga-Soria, F. J.; Béjar, R.; Álvarez, P.; Muro-Medrano, P. R. (2005): OGC catalog services: a key element for the development of spatial data infrastructures. In: Computers & Geosciences, 31 (2), pp. 199-209.

Pellicer, F. J. L. (2011): Semantic linkage of the invisible geospatial web. Doctoral dissertation, Universidad de Zaragoza (unpublished).

Perry, M.; Herring, J. (2012): OGC GeoSPARQL – A Geographic Query Language for RDF Data. OGC Implementation Standard, OGC 11-052r4, September 2012.

Prudhomme, C.; Homburg, T.; Ponciano, J. J.; Boochs, F.; Roxin, A.; Cruz, C. (2017): Automatic Integration of Spatial Data into the Semantic Web. WebIST 2017, April 2017, Porto, Portugal.

Redman, T. C. (2001): Data quality: the field guide. Digital press, Boston.

Shi, W.; Fisher, P.; Goodchild, M. F. (2003): Spatial data quality. CRC Press, Boca Raton, FL.

Stadler, C.; Lehmann, J.; Höffner, K.; Auer, S. (2012): Linkedgeodata: A core for a web of spatial open data. In: Semantic Web, 3 (4), pp. 333-354.

Vrandečić, D.; Krötzsch, M. (2014): Wikidata: a free collaborative knowledgebase. In: Communications of the ACM, 57 (10), pp. 78-85.