ONTOLOGY BASED QUALITY EVALUATION FOR SPATIAL DATA

C. Yılmaz[a], Ç. Cömert[a]

[a] KTU, Dept of Geomatics Engineering, 61080, Trabzon, TURKEY, (cemre.yilmaz, ccomert)@ktu.edu.tr

KEY WORDS: spatial data quality, rule based evaluation, data quality evaluation, spatial ontology, GeoSPARQL, SWRL

ABSTRACT:

Many institutions will be providing data to the National Spatial Data Infrastructure (NSDI). Current technical background of the NSDI is based on syntactic web services. It is expected that this will be replaced by semantic web services. The quality of the data provided is important in terms of the decision-making process and the accuracy of transactions. Therefore, the data quality needs to be tested. This topic has been neglected in Turkey. Data quality control for NSDI may be done by private or public "data accreditation" institutions. A methodology is required for data quality evaluation. There are studies for data quality including ISO standards, academic studies and software to evaluate spatial data quality. ISO 19157 standard defines the data quality elements. Proprietary software such as, 1Spatial's 1Validate and ESRI's Data Reviewer offers quality evaluation based on their own classification of rules. Commonly, rule based approaches are used for geospatial data quality check. In this study, we look for the technical components to devise and implement a rule based approach with ontologies using free and open source software in semantic web context. Semantic web uses ontologies to deliver well-defined web resources and make them accessible to end-users and processes. We have created an ontology conforming to the geospatial data and defined some sample rules to show how to test data with respect to data quality elements including; attribute, topo-semantic and geometrical consistency using free and open source software. To test data against rules, sample GeoSPARQL queries are created, associated with specifications.

1. INTRODUCTION

National Spatial Data Infrastructure (NSDI) is in developmental stage in Turkey. Current technical background of the NSDI is based on syntactic web services. It is expected that this will be replaced by semantic web services in future (Kara and Cömert, 2011). Many institutions will be providing data to NSDI. An institution will be a user and also a provider at the same time for NSDI. The data quality is important for both users and producers mainly for decision making applications and correct transactions. In data production process, quality controls are mainly done at the field according to regulations. There is no enforcement and regulation for data quality assessment of gathered data for NSDI. This leads to many problems in existing data. So a methodology is needed to assess the conformance of data to its specifications.

There are many studies done for spatial data quality including ISO standards. According to these standards, data quality elements provide quantitative quality information and describe how well a dataset meets the criteria set forth in its product specification (ISO, 2002). It is classified with five categories; completeness, logical consistency, positional accuracy, thematic accuracy, temporal quality (ISO, 2013). The latest standard is ISO 19157:2013 (ISO, 2013).

There are studies done and software produced to assess data quality. Mostafavi et al. (2004), Sanderson et al. (2009), Wang et al. (2005) can be given as examples to these studies. 1Spatial's 1Validate and ESRI's ArcGIS Data Reviewer are the main data quality check (QC) software. In this paper we seek for the methods and their implementations as a part of needed methodology. In Mostafavi et al. (2004), compliance to Canada National Topographic Database (NTDB) specifications was tested. The ontology of the NTDB was defined and translated into Prolog, then rules were created.

The second study, Sanderson et al. (2009) was made by the 1Spatial group. The 1Spatial group has developed a "standard-based rules language". They used Radius Studio software to implement this language. In INSPIRE (Infrastructure for Spatial Information in European Community) Annex 1 testing process, consistency of the datasets to its specifications have been tested. Wang et al. (2005) made a spatial data quality check while data gathering with mobile devices. They created a rule base using the Semantic Web Rule Language (SWRL). They used data in Geographic Markup Language (GML) which can be accessed with the Web Feature Service. SWRL was used to add some constraints to database schema defined with GML and with these restrictions system has the capability to warn the data collector if there is any inconsistent data gathered in the field.
One of the software for spatial data quality evaluation is 1Validate service. It validates spatial data against the different kinds of standard rules including, geometric, polygon, network and Irish MapRoad check rules.

ESRI, a GIS software company, has developed Data Reviewer as an extension for data QC (ESRI, 2013). It allows for topological checks, duplicate geometry checks etc. Topological checks "returns the geometry of features that violate the topology rules that have been defined for a feature dataset in the geodatabase" (ESRI, 2013). For example a parcel layer must not have dangles.

"Rule-based" approach is the common method in all these studies. In the semantic web context, ontologies and logic programming are used to implement rule based approach. Using expressiveness of ontology languages such as OWL-DL, OWL2 etc., reasoning capabilities and spatial transactions together allows efficient quality evaluation for spatial data. In this work, we present a rule-based approach as used in many applications. We look for how to implement a rule base for testing data conformance to its specification. In this study as a part of testing such a conformance, we consider some data quality elements

including attribute, topo-semantic and geometrical consistency (Goodchild, 1995; Servigne et al., 2000) using FOSS. Especially for spatial transaction required data quality element checking, there is a need for such an implementation in semantic web context.

For the case study, we used the base map of Trabzon as data, and we have identified several types of inconsistency related problems within data, taking into account the data specifications. Buildings that have more than seven floors, buildings which intersect with road data, parcels on which there are more than two buildings are some examples of these inconsistencies. We have developed an ontology associated with OGC "SpatialObject" class and devised four sample OGC GeoSPARQL queries and implemented them with BBN Parliament triple store. The rest of the paper follows with the explanation of the ontology and data quality. In this section ontology and current standards that can be used for ontology based spatial data quality evaluation is explained. Third section describes implementation of the method with a case study. In this section tools and queries are explained with sample data, one part of base map of Trabzon. Finally paper ends with conclusion.

2. ONTOLOGY AND DATA QUALITY EVALUATION

Spatial data should be produced in accordance with its data product specifications. To evaluate its conformance to specification, possible inconsistencies for data should be defined. This is a basis to find out the data that are not consistent with its specifications.

Ontologies are the main components for semantic web services. They can be defined in an abstraction level that allows for easy reusability by new users with eventually different datasets at hand.Ontologies include concepts from a specific domain with classes, attributes, relations and restrictions. In the geospatial data domain, ontology deals with the totality of geospatial concepts, categories, relations and processes and with their interrelations at different resolutions (Mostafavi et al., 2004). Spatial domain ontologies can be used to define all the "concepts", "attributes", and "interrelations" of concepts. In a specification or regulation these concepts can be defined such as "road class has attribute "roadType" as String". In this example "road" is a class for road data, "roadType" is datatype property with "String" data type. To define ontology, it is necessary to decide which semantic web language and which ontology editor will be used. Every ontology language has its own way to define the concepts, properties of concepts and relations between them in a different way. RDF, RDFS, OWL1 and OWL2 are examples for ontology languages. Main differences of these languages are their expressiveness. OWL2 is more expressive than OWL1 and RDF and difference between their object property characteristics can be given as an example. While OWL 1 allows assertions that an object property is symmetric or transitive, it is impossible to assert that the property is reflexive, irreflexive or asymmetric (Bao, 2012). Such expressiveness enable to define more assertions. In spatial ontologies spatial relations between each class can be defined with object properties. For example, in spatial domain, spatial relation between "road" and "building" classes should be "disjoint" which can be characterized as symmetric. "If a property is symmetric, then if the pair (x,y) is an instance of the symmetric property P, then the pair (y,x) is also an instance of P" (Golbreich et al., 2009). If a road x is disjoint from a building y, then that building y is disjoint from the road y. Grau et al. (2008) makes an overview of OWL 2—an extension to and revision of OWL, which is developed within W3C OWL Working Group.

For data quality evaluation using ontologies, we employ constraints such as "Buildings that have more than seven floors are inconsistent". This exceeds OWL expressivity. To implement such constraints/rules on OWL ontologies, currently, a rule language is needed. Semantic Web Rule Language (SWRL), a W3C submission, is a language for specifying rules to be applied on Semantic Web ontologies.

SWRL makes use of built-ins, but it does not have spatial built-ins, and the rules cannot be geospatial; we cannot define a rule about "buildings which intersect with road data". To work with geospatial data, SWRL needs to be extended. Karmacharya et al. (2010) attempts to introduce spatial built-ins for SWRL. They use PostGIS, a spatial extension of PostgreSQL. However, currently there is not a spatial built-in support for SWRL in existing ontology editors, in our case Protégé.

There are query languages for Semantic Web Data. The basic, W3C recommended one is SPARQL Query Language for RDF. SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL also supports extensible value testing and constraining queries by source RDF graph (Prud'Hommeaux and Seaborne, 2008).

For a geospatial query such as "Find the buildings which intersect with road data", relationship between the classes should be calculated at instance level. A Geographic Query Language for RDF Data (GeoSPARQL) is published as an OGC standard specification (OGC, 2012). The OGC GeoSPARQL standard supports representing and querying geospatial data on the Semantic Web. GeoSPARQL defines a vocabulary for representing geospatial data in RDF, and it defines an extension to the SPARQL query language for processing geospatial data (OGC, 2012). Its ontology has three basic classes; geo:SpatialObject, geo:Feature, geo: Geometry as illustrated in Figure 1.

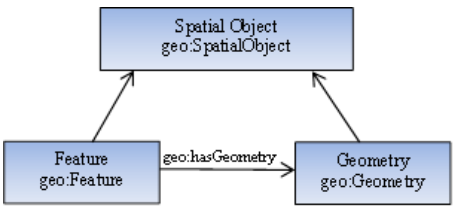

Figure.1. Main concepts and their relations in GeoSPARQL

It uses two types of geometries; Geography Markup Language (geo:GML) and Well Known Text (geo:asWKT) for geospatial data.The geo:asWKT and geo:asGML properties link the geometry entities to the geometry literal representations. Values for these properties use the sf:wktLiteral and gml:gmlLiteral data types respectively (Battle and Kolas, 2012). Furthermore, Simple Features, Egenhofer and RCC8 relations are used as topological relations (as seen in Table.1). GeoSPARQL standard is implemented by Parliament, Oracle Spatial and Strabon.

We use GeoSPARQL to take advantage of its spatial relations support for making query of RDF.

Simple Features	Egenhofer	RCC8
equals	equal	EQ
disjoint	disjoint	DC
intersects	¬disjoint	DC
touches	meet	EC
within	inside+covered by	NTTP+TPP
contains	contains+covers	NTTPi+TPPi
overlap	overlap	PO

Table 1. Topological relations; Simple Features, Egenhofer (Egenhofer and Herring, 1990), RCC8 (Battle and Kolas, 2012).

Additionally, using reasoning with the expressiveness of OWL, it is possible to check the consistency of the ontology and the rules with the help of spatial reasoner such as PelletSpatial (Stocker and Sirin, 2009). It provides consistency checking and query answering over spatial data represented with the Region Connection Calculus (RCC).

3. CASE STUDY

For the case study, we used the base map of Trabzon as data, and we have identified several types of inconsistency related problems within data, taking into account the specifications. Building, cadastral parcel and road data are used as sample data. We use MySQL as relational database for geospatial data. We created a subcollection of geometries conforming to OGC standards with QGIS, and exported them to MySQL. We have developed an ontology associated with OGC SpatialObject class.

As the ontology editor, we used Protégé (versions 4.3 and 3.4.2) with its plugins. We created an ontology associated with GeoSPARQL SpatialObject class and subclasses. Then, to import spatial data into Protégé (Figure.2), DataMaster plug-in is used with version 3.4.2. (Nyulas et al., 2007). Each table in the database is imported as a subclass of "Feature" class which is subclass of "Spatial Object", and associated with elements in Geometry subclass. We use geometries as "asWKT" for each type of feature.

Finally, to find out inconsistencies, we make queries with Parliament, a triple store that implements GeoSPARQL (Battle and Kolas, 2012). First, we inserted the data, and created spatial indices. When we converted our ontology and saved it as RDF, queries supported with OGC functions, e.g. sfIntersects, sfOverlaps, started to work.

We have implemented four types of queries taking into account classifications provided in literature (Goodchild, 1995; Servigne et al., 2000).

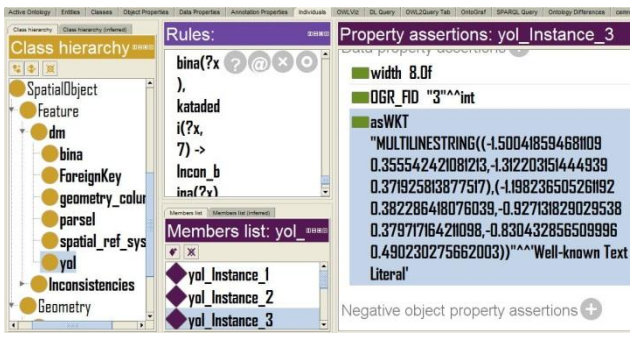

Figure.2. Protégé ontology editor with the ontology

Query 1) "Find buildings that have more than seven floors" (where this is restricted).

```
SELECT ?x
{?x rdf:type ouront:bina .
?x ouront:bina.katadedi ?y .
FILTER( ?y > 7 ) .}
```

Only one instance satisfied this simple query as seen in Figure 3. (bina@tr = building@en, katadedi@tr = numfloors@en). The first query is an example for a general, non-spatial, given attribute-related situation. It features an attribute accuracy problem (Goodchild, 1995).

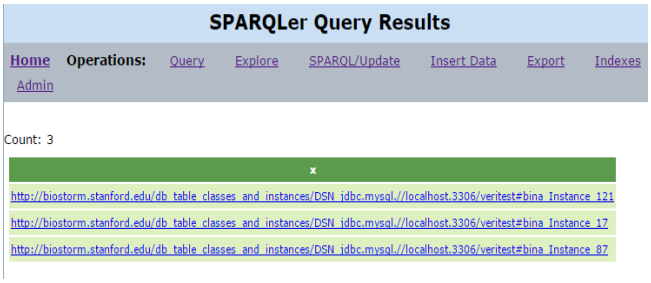

Figure. 3. Result of the first query.

Query 2) "Find buildings that intersect with roads".

```
SELECT
?z
WHERE {
?x ouront:yol.asWKT ?y.
?z ouront:bina.asWKT ?b.
FILTER(geof:sfIntersects(?y,?b)) }
```

There are two buildings that intersect with roads as seen in Figure 4 (yol@tr = road@en). The second query is a simple query that has geospatial component, using OGC simple feature relations. It features a topo-semantic inconsistency, a type of logical inconsistency (Servigne et al., 2000).

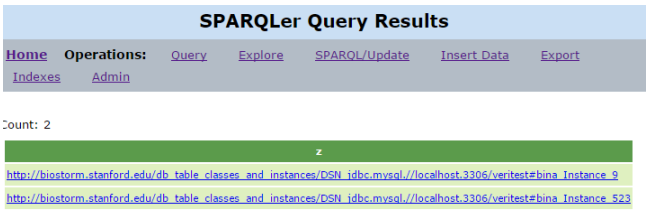

Figure. 4. Result of the second query.

3) "Find parcels that intersect with more than 2 buildings"

```
SELECT
?x (COUNT(?x) as ?xc)
WHERE {
?x ouront2:parsel.asWKT ?p .
?z ouront:bina.asWKT ?b .
FILTER(geof:sfIntersects(?p, ?b))}
GROUP BY ?x
HAVING ( ?xc >2)
```

There is only one such parcel, and it intersects with three buildings (parsel@tr = parcel@en) as seen in Figure 5.

The third query uses more advanced functions related to SPARQL (ARQ type aggregates, here), also present in GeoSPARQL. This also features a topo-semantic error (Servigne et al., 2000).

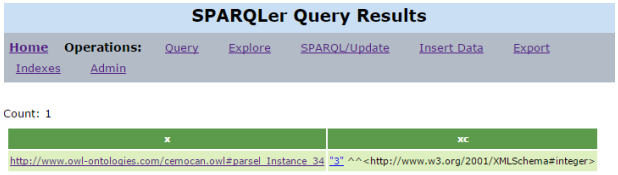

Figure. 5. Result of the third query.

4) "Find roads whose endpoints are close to the boundary of some building smaller or equal to 0.5".

```
SELECT
?y
WHERE {

?y ouront:yol.asWKT ?yw .
?b ouront:bina.asWKT ?bw .
BIND (geo:boundary(?yw) as ?n ) .
BIND (geo:boundary(?bw) as ?k ) .
BIND (geo:buffer(?k, 0.5) as ?kb ) .
FILTER (geof:sfIntersects(?n, ?kb) ) .
}
```

No such road satisfied the query.

When the bound for distance is selected correctly, the result of fourth query helps us solve the following problem that occurs occasionally. The surveyors sometimes do not complete line detail when it is nearby a border of a polygon detail. This is a geometric error (Servigne et al., 2000), which is visually hard to detect. Our data, subcollection of geometries, did not have such a situation according to the query (that should return a superset), and if the bound is correct, the afore-mentioned situation did not occur in our data.
These queries did not work with the whole dataset on an i5 PC with 4GB memory. We reduced the dataset to a town, and created a subcollection of geometries.

4. CONCLUSION

In this study, we seek how ontologies can be used to evaluate spatial data quality. We consider three types of quality elements; attribute, topo-semantic and geometrical consistency. A part of base map of Trabzon is used as data and a case study is implemented with current standards. We have developed an ontology associated with OGC SpatialObject class using Protégé ontology editor, and devised sample GeoSPARQL queries and implemented them with BBN Parliament triple store. These help us measure the quality of spatial data provided by institutions in semantic web context.

We have used a subcollection of geometries within data for reduced memory consumption. As a result, we can say these tools can be used for data quality evaluation for ontologies which are created as RDF. RDF has the least expressive power of ontology languages. The OGC GeoSPARQL functions did not work on more expressive OWL ontologies. This is a disadvantage of the implementation.

Ontologies are good means for data quality evaluation because of their reusability. To take advantage of expressiveness of OWL2 ontologies a query language for OWL2 is needed. There is not a standard query language such as GeoSPARQL for OWL2 ontologies that supports spatial functions.

REFERENCES

Bao, J., 2012. OWL 2 Web Ontology Language document overview. W3C Recommendation. World Wide Web Consortium, December 2012.

Battle, R., Kolas, D., 2012. Enabling the geospatial Semantic Web with Parliament and GeoSPARQL. *Semantic Web*, 3(4), pp. 355-370.

Egenhofer, M.J., Herring, J., 1990. A mathematical framework for the definition of topological relationships. *Fourth international symposium on spatial data handling*, Zurich, Switzerland, pp. 803-813.

ESRI, 2013. ESRI ArcGIS Data Reviewer Extension.

Golbreich, C., Wallace, E.K., Patel-Schneider, P., 2009. OWL 2 Web Ontology Language: new features and rationale. W3C working draft, W3C.

Goodchild, M.F., 1995. Attribute accuracy, Elements of spatial data quality, pp. 59-79.

Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U., 2008. OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6 (4), 309-322.

ISO, 2002. Geographic Information – Quality Principles, 19113: 2002.

ISO, 2013. Geographic information – Data Quality, 19157: 2013.

Kara, G., Cömert, Ç., 2011. ULUSAL KONUMSAL VERİ ALTYAPISI İÇİN SEMANTİK VERİ TANIMLAMA, *UCTEA GIS Congress*, Antalya, Turkey.

Karmacharya, A., Cruz, C., Boochs, F., Marzani, F., 2010. Spatial Rules through Spatial Rule built-ins in SWRL. *Journal of Global Research in Computer Science*, 1(2), pp. 1-4.

Mostafavi, M.-A., Edwards, G., Jeansoulin, R., 2004. An ontology-based method for quality assessment of spatial data bases, *Third International Symposium on Spatial Data Quality*, Bruck/Leitha, Austria, pp. 49-66.

Nyulas, C., OConnor, M., Tu, S., 2007. DataMaster–a plug-in for importing schemas and data from relational databases into Protégé, *10th International Protege Conference*, Budapest, Hungary.

OGC, O.G.C., 2012. OGC GeoSPARQL-A geographic query language for RDF data. OGC Candidate Implementation Standard 2.

Prud'Hommeaux, E., Seaborne, A., 2008. SPARQL query language for RDF. W3C recommendation 15.

Sanderson, M., Ramage, S., Van Linden, L., 2009. SDI Communities: Data quality and knowledge sharing, *11th GSDI Conference*, Rotterdam, The Netherlands.

Servigne, S., Ubeda, T., Puricelli, A., Laurini, R., 2000. A methodology for spatial consistency improvement of geographic databases. *GeoInformatica*, 4(1), 7-34.

Stocker, M., Sirin, E., 2009. PelletSpatial: A Hybrid RCC-8 and RDF/OWL Reasoning and Query Engine. *OWLED*, 529.

Wang, F., Mäs, S., Reinhardt, W., Kandawasvika, A., 2005. An Ontology-based Quality Assurance Method for Mobile Data Acquisition. *19th international conference on Informatics for Environmental Protection: Networking Environmental Information*, Brno, Czech Republic, pp.334-341.