SDQO and SFO, Ontologies for Spatial Data Quality Assessment

C. Yilmaz¹, C. Comert¹, D. Yildirim¹

¹KTU, Engineering Faculty, Dept. of Geomatics Engineering, 61080 Ortahisar, Trabzon Turkey - (cemre.yilmaz, ccomert, dyildirim)@ktu.edu.tr

KEY WORDS: Geo-ontology, specification ontology, spatial data quality ontology, spatial data quality management.

ABSTRACT:

Spatial quality assessment is based on the conformance of data to its specifications or fitness for users’ purpose. These specifications and the users’ purposes include the rules and constraints that a dataset should comply with. Assessing the compliance of data to the rules is still an active research subject and rule-based approach is the common method. For the efficient rule-based system implementation, it is desired to automate assessment process with a domain-independent and web-based approach. Reasoning capability and re-usability of semantic web components are expected to promote efficient implementation. In literature, many domains such as agriculture, music, Linked Data and geospatial domain etc. apply ontology-based methods for quality management. There is a need to model geospatial quality concepts and rules in a domain-independent way to automate the quality management process. In our model of rule formalism, we use Web Ontology Language (OWL) and Semantic Web Rule Language (SWRL). We devise two types of ontologies. These are: the specification ontologies (SFO) and the Spatial Data Quality Ontology (SDQO). SFO is to be created by domain experts/users to define rules according to specifications. SDQO is responsible with quality assessment; it is domain independent and makes assessment based on the rules defined by any SFO for the related domain. The quality elements are domain and toposemantic consistency that assessed by SWRL. In this paper, the design considerations of the ontologies for quality assessment are explained with an example.

1. INTRODUCTION

High quality spatial data is essential in providing better analyses and making better decisions involving such data. There have been studies such as ISO (2013), Zaveri et al. (2015), Fonte et al. (2017) to categorize and define the quality concepts, quality evaluation methods and solutions for producing better data.

In the modern age of web, Semantic Web (“meaningful web”) components such as ontologies are used for Rule-based system implementation because of expressivity, reasoning capability and re-usability. This is still an active research theme with recommended, accepted and upcoming standards; SWRL (W3C, 2004), SPARQL Inferencing Notation (SPIN) (W3C, 2011), Rule Interchange Format (RIF) (W3C, 2013) and a newer standard for schema validation Shapes Constraint Language (SHACL) (W3C, 2017).

Mostafavi et al. (2004), Fürber and Hepp (2011), Debattista et al. (2016), Degbelo (2012), Geisler et al. (2016), Zhu (2014) and Nash et al. (2011) are the examples to the studies that use ontologies for data quality management purposes.


In addition to quality management with OWL, SWRL rules are used in several studies for quality assessment process. (Wang et al., 2005); (Cheng et al., 2008); (Keßler et al., 2009); (Zhu, 2013); (Cherfif et al., 2017); (Varadharajulu et al., 2017); (Mobasheri, 2017); (Homburg and Boochs, 2019) can be given as example.

Wang et al. (2005) developed a system to detect inconsistent spatial data with the help of SWRL rules in a specific domain. Homburg and Boochs (2019), Varadharajulu et al. (2017), and Mobasheri (2017) propose rule-based approaches including SWRL rules for domain dependent solutions. Varadharajulu et al. (2017) design a framework to check the consistency of the transportation data against the rules that are created with SWRL.

In these studies, SWRL rules are used with a domain dependent quality management framework. Mobasheri (2017), proposes a rule -based system to increase the Quality of the OSM data with rules created by SQWRL.

Besides the previously explained studies, Nash et al. (2011) design a framework for the automatization of specification rules in agriculture domain with implementation of geospatial rules as Interchangeable Rule Format (RIF) and proposes, GeoRIF. The ontology they devised is specific to the agriculture regulations.

It is desired to make a robust framework that can be applicable to spatial data, independent from the domain. This will reduce redundancy and increase interoperability. Hence, there is a need to create ontology for the designed framework. The components and details of the framework is out of the scope of this paper. The main focus is the explanation of devised ontologies. These are; Specification Ontology (SFO) and Spatial Data Quality Ontology (SDQO). Following sections describe SDQO and SFO and briefly explain how to use them for quality assessment with an example rule.

2. ONTOLOGIES FOR SPATIAL DATA QUALITY MANAGEMENT

2.1 Ontology Design

Initial step to create ontology for quality management is to define the concepts. For this purpose, motivating scenario and
competency questions are defined. Following, existing ontologies are researched for both geospatial and data quality management concepts. The following sections explain the motivating scenario and questions, ontologies for quality management and geo-ontologies respectively.

2.1.1 Motivating Scenarios and Questions

Spatial data producers want to produce data compliant to the rules in the regulations/specifications. Producers should formalize the rules, assess the data and have the quality result as a report.

Data producers can select the rules that should be applied to a specific spatial task in a domain. Formalized rules for a domain can be stored and reusable for further assessments.

- Which instances of tested class have data quality problems according to predefined data quality rules?
- Which instances of tested class have spatial relation problems with the specified ones? e.g. mustBeWithin, mustNotCross.
  - Which instances of tested class have cross relation with the instances of another class when it is forbidden? (must not cross)
  - Which instances of tested class overlap with the instances of the same class when it is forbidden? (must not overlap)
  - Which instances of tested class with the type of Polygons/Lines overlaps any other feature in the same class when it is forbidden? (must not overlap with)
  - Which instances of tested class are spatially within any instance from the second feature class when it is forbidden? (must not be within)
  - Which instances of the tested class is not within any instance of the second class when it is forbidden? (must be within)
- Which instances of tested class have domain consistency problems?
  - Which instances of tested class have attributes different from a constant value while it is forbidden?
  - Which instances of tested class have “null” data value for its attributes while it is forbidden?
  - Which instances of tested class have an attribute greater than a specified one while it is not allowed?
- What is the percentage of the erroneous instances in the tested data?

2.1.2 Geospatial Ontologies

Mostly used ontologies for geospatial domain are; GeoNames (GeoNames Team, 2006), W3C Geo (W3C Semantic Web Interest Group, 2004), GeoOWL (W3C Geospatial Incubator Group, 2007) and OGC GeoSPARQL (OGC, 2012). The GeoNames Ontology and W3C Geo support only point type geometries. GeoOWL, the updated model of W3C Geo, is created compatible with GeoRSS Feature Model. Only GeoSPARQL supports other geometry types such as polygon and lines and basic spatial relations. The GeoSPARQL ontology has been selected for defining geospatial concepts and relations shown in Figure 1. Its classes are used as superclasses to the classes in the designed ontologies.

2.2 Devised Ontologies

In this study two types of ontologies are introduced. The SDQO ontology, contains the necessary rules and the concepts related with the quality assessment. The SfOs are simple ontologies, developed keeping in mind the reusability of rules with different kinds of datasets. There can be one or more SfOs (eg. for different scales) for each institution. Devised ontologies and their relations are shown on Figure 2. This section continues with the subsections to explain SDQO and SfO.

2.2.1 SDQO

According to the set of rules depending on appropriate and chosen elements for data quality, SDQO is also responsible of processing and integrating data quality elements with associated procedures and implementing the procedures in accordance with the geospatial data and the quality elements. SDQO prepares the resulting spatial data quality ontology which relates data quality results with tested data and prepares it for queries or publishing.

While creating the SDQO, terminologies which are used in other studies are considered. Although, all ontologies are created with the same intention, quality management, they are specific to different application domains. They have some common concepts such as “Quality Dimension”, “Quality Metric” or “Quality Result”.

In SDQO, there are three top classes directly below owl:Thing, one for data, one for data quality elements, and one for data quality results and processes. These are ogc: SpatialObject, sdqo:DataQualityElement, sdqo:DataQualityResult, respectively as shown on Figure 3.
In SDQO, ogc:Feature has three direct subclasses, other than the ones from imported ontologies. These are sdqo:GeomClassifiedFeature, sdqo:FixedRefFeature and sdqo:RestrictedFeature. These classes are represented in Figure 4.

sdqo:RestrictedFeature has four direct subclasses, as shown in Figure 5. These are sdqo:InterObjectPrRF, sdqo:IntraDataPrRF, sdqo:IntraObjectPrRF and sdqo:InterDataPrRF. Rules and given subclass relations determine which features these classes have. The labels and the descriptions are listed in Table. “Intra” classes are for restrictions within a single class and “Inter” classes are for restrictions in class pairs. sdqo:RestrictedFeature has the label “Features according to the property restrictions”.

sdqo:GeomClassifiedFeature is defined to classify features classified according to their geometries. It has three subclasses, sdqo:CalcPoly, sdqo:CalcLine and sdqo:CalcPoint, declared to be owl:disjoint. Any feature that is associated with a geometry that has a valid ogc:asWKT value, is sdqo:CalcPoint, declared to be owl:disjoint. Any feature that is use of reference individuals.

sdqo:FixedRefFeature has OWL named individuals to be used in rules as reference markings. For instance, attribute tests can make subclass relations determine which features these classes have. The labels and the descriptions are listed in Table 1. “Intra” classes are for restrictions within a single class and “Inter” classes are for restrictions in class pairs.

sdqo:DataQualityElement is defined as abstract class to classify the elements of data quality. It has subclasses as follows,

sdqo:GeometricAccuracy has sdqo:GeometryValidity as a subclass.

sdqo:TopoSemanticConsistency is a subclass to sdqo:SemanticAccuracy and sdqo:LogicalConsistency. sdqo:LogicalConsistency also has sdqo:DomainConsistency.

sdqo:DataQualityResult is defined as abstract class to classify the results of data quality assessment.

SDQO has the following own datatype properties. Several of them are to be used as a super property to data properties in SfOs.

Other datatype properties are schema:startTime and OGC datatype properties such as ogc:asWKT.

sdqo:errorCode datatype property is defined to give a code to results to identify the problem of data.

sdqo:hasErrorCode datatype property is defined to relate data with the resultant sdqo:errorCode.

sdqo:dataHasErrorWithCode datatype property connects erroneous data to error codes.

Every feature must have sdqo:featureID datatype property which is to be unique to every feature.

sdqo:hasMessage datatype property connects results and processes to error messages.

sdqo:elementHasMessage datatype property is a shortcut from elements to error messages (sdqo:hasResult and sdqo:hasMessage).

There are SDQO datatype properties for attribute tests such as sdqo:dataProp01.

sdqo:subnr property is the main data property for assigning the classes to be tested with the subclass number. It has subproperties such as sdqo:subnrOverlap and sdqo:subnrMustWithin. These subproperties are used in defining top classes in SfO.

sdqo:hasQueryString datatype property is for SPARQL query strings.

While sdqo:resultForData object property create a relationship between quality results and data, sdqo:hasResult relates data quality elements to results. sdqo:hasResult is inverse property to sdqo:resultForData.

Furthermore, GeoSPARQL object properties such as ogc:sfOverlaps are used for relations between spatial classes. ogc:sfIntersects is declared to be a super property to intersection type properties such as ogc:sfOverlaps. SDQO as in Protégé ontology editor is shown in Figure 6.

**Table 1. SDQO superclasses**

<table>
<thead>
<tr>
<th>Class</th>
<th>rdfs:label</th>
</tr>
</thead>
<tbody>
<tr>
<td>sdqo:InterObjectPrRF</td>
<td>Feature restricted wrt relations between classes</td>
</tr>
<tr>
<td>sdqo:InterDataPrRF</td>
<td>Feature restricted by attributes between classes</td>
</tr>
<tr>
<td>sdqo:IntraDataPrRF</td>
<td>Feature restricted by attributes within class</td>
</tr>
<tr>
<td>sdqo:IntraObjectPrRF</td>
<td>Feature restricted wrt relations within class</td>
</tr>
</tbody>
</table>
2.3 SfO, Specification Ontology

SfO is designed to be manageable by domain experts to be even without Semantic Web expertise. A GUI is created for users to input rules according to the specifications. Once rules are defined by the help of GUI, an OWL file (SfO) is created as a result of translations from GUI to CSV and OWL files respectively. The translations and the assessment framework are out of the scope of this study. The SfO and its components will be explained in detail.

A SfO has the class hierarchy reflecting the associated specifications. Furthermore, these specification hierarchy classes have superclasses, still in SfO. These classes are created to deanonymize the restrictions. They are the top classes with SfO’s IRI. They are subclasses to SDQO classes and establish the connection to SDQO. Furthermore, these top classes should have associated geometries under the appropriate simple features class, such as sf:Polygon or sf:Linestring.

SfOs directly import SDQO and the directly or indirectly import data ontologies associated with that SfO or other SfOs.

With SfO, more general relations are translated into is-a relations in the scope of SfO, if possible. The feature pairs causing the errors are identified. Below, a sample translation is given.

<table>
<thead>
<tr>
<th>Specification rule</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features in class A must not overlap with features in class B</td>
<td>A is a subclass of C, an SfO class; B is a subclass D, an SfO class; these are subclasses of some SDQO classes that take part in SWRL rules establishing the relevant error properties involving faulty features in A, B and forbidden overlaps relation.</td>
</tr>
</tbody>
</table>

SfOs define specification classes and generic classes for data to be tested. SfO generic classes are defined according to the spatial relations related to data. Spatial relations should be defined between feature classes. In specifications, a feature class might have topological relations with; itself, a feature class, more than one feature class.

Specification classes are SfO classes such as “sfo:Road” and “sfo:Building”, that exist in the specifications. They can have subclasses; sfo:PermanentLake can be a subclass to sfo:Lake.

Examples include sfo:ClassOv01, sfo:ClassOv02, etc. and sfo:ClassOv01, sdqo:ClassOv02, etc. Here Ov is for Overlaps, following f is for ‘first’, ‘s’ is for ‘second’. The numbers represent the subclass numbers, the order in pairs of specification classes. A path of subclass relations in a SfO is shown in Figure 7.

2.4 SDQO and SfO relation

The data quality elements are linked to individuals for processes and results. These individuals are used in SWRL rules and SPARQL queries. Classes for features with constraints are defined. SfOs are devised to define which restrictions will be applied to dataset according to specifications by a domain expert. They are mostly hierarchical. SfOs import SDQO and data ontologies. The data ontologies are likely to have some faults. The system should be robust and stay consistent. The system is designed to be robust and user-friendly, therefore usable. The common parts of the SfOs are moved to SDQO. Most of the rules are in the SDQO. The use of SDQO enables domain-independent, easy-to-update quality assessment with the SfOs. Especially with an ontology editor, it is expected that domain experts can quickly understand how to manipulate and update their SfOs, when necessary. SfOs typically do not use the whole capability of the SDQO. When the rulesets change, even without needing to change the SDQO significantly, it is expected that the new SfO can be implemented. Thanks to the prevailing OWA, SDQO is easy-to-update, when necessary. With OWA, truth of statements does not change.

Case 1: Geometric classes with Forbidden relation (Inter-object Must Not).

Example: “c1, c2, Forbidden, Crosses, class1, class2; class3, 6210.12”

Meaning: “class1 is of line-type, class2 and class3 are of polygon-type, features in class1 cannot have the Crosses relation with the features in the classes class2 and class3” 6210.12 is the timestamp, which is the same throughout the process (the minimal one).

Two generic classes are created in the SfO; they are set as subclasses of the relevant classes in SDQO. If there are already six “c1,c2,Forbidden,Crosses” pairs, the created classes are sfo:ClassCrf07 and sfo:ClassCrs07. Here “Cr” stands for “Crosses”, “f” stands for “first” and “s” stands for “second”. Restrictions are among anonymous classes in OWL. sfo:ClassCrf07 is set as a subclass of SDQO class sdqo:ClassCross1 and sfo:ClassCrs07 is set as a subclass of sdqo:ClassCross2.

The subclass relations are established. sfo:class1 is created as a subclass of sfo:ClassCrf07. sfo:class2 and sfo:class3 are created as subclasses of sfo:ClassCrs07. A sample for relevant class hierarchies are given below with the associated SWRL rule. It is also shown in Figure 8. “~” denotes the subclass relation. Case 1 type cases (“Inter-object Must Not”) are easily updated by updating the subclass relations of the specification classes or creating new classes with the appropriate subclass relations.
SWRL rules for this kind of restriction is as follows, 
crossesOrOverlaps(?x, ?y) ^ CalcLine(?x) ^ CalcPoly(?y) -> 
sfCrosses(?x, ?y)

RestrictedFeature(?x) ^ ogc:hasGeometry(?x, ?g) ^ 
ogc:asWKT(?g, ?w) ^ swrlb:contains(?w, "POLYGON") -> 
CalcPoly(?x)

RestrictedFeature(?x) ^ ogc:hasGeometry(?x, ?g) ^ 
ogc:asWKT(?g, ?w) ^ swrlb:contains(?w, "LINESTRING") -> 
CalcLine(?x)

Sample class hierarchy paths:
sfo:Road < sfo:ClassCrf07 < sdqo:ClassCross1 < sdqo:Cross < 
sdqo:InterObjectPrRF < sdqo:RestrictedFeature < ogc:Feature < 
ogc:SpatialObject < owl:Thing

sfo:Road < sfo:ClassCrf07 < [ sdqo:subnrCross = 7 ]

sfo:Building < sfo:ClassCrs07 < sdqo:ClassCross2 < sdqo:Cross < 
sdqo:InterObjectPrRF < sdqo:RestrictedFeature < ogc:Feature < 
ogc:SpatialObject < owl:Thing

sfo:Building < sfo:ClassCrs07 < [ sdqo:subnrCross = 7 ]

3. CONCLUSION

The aim of this study was to create ontologies that can be used 
for spatial data quality assessment process. To support 
reusability, domain independence, extensibility and spatial 
quality rules two types of ontologies are devised. SfO is created 
by the domain experts to represent rules. Once a 
company/institution or a user create the SfO for its domain 
( agriculture, transportation etc.) then this ontology can be saved 
to represent spatial problems.

4. REFERENCES

Cheng, G., Du, Q., Ma, H., 2008. The Design and 
Implementation of Ontology and Rules Based Knowledge Base 
for Transportation, in: 2008 International Conference on 
Computer Science and Software Engineering. Presented at the 
2008 International Conference on Computer Science and 
https://doi.org/10.1109/CSSE.2008.1405

Cidália Costa Fonte, Vyon Antoniou, Lucy Bastin, Jacinto 
Estima, Jamal Jokar Arsanjani, Juan-Carlos Laso Bayas, Linda 
See, Rumiana Vatseva, 2017. Assessing VGI Data Quality, in: 
Mapping and the Citizen Sensor.

for Linked Data Quality Assessment, in: 2016 IEEE Tenth 
124–131. https://doi.org/10.1109/ICSC.2016.48

Spatial Data Quality Characterization in the Semantic Sensor 

Fei Wang, Stephan Mäs, Wolfgang Reinhardt, Admire 
Kandawasvika, 2005. Ontology Based Quality Assurance for 
Mobile Data Acquisition. Presented at the Informatics for 
Environmental Protection - Networking Environmental Information.

Fürber, C., Hepp, M., 2011a. Towards a vocabulary for data 
quality management in semantic web architectures, in: 
Proceedings of the 1st International Workshop on Linked Web 

Data Quality Management for Data Streams. J. Data Inf. Qual. 7, 
1–34. https://doi.org/10.1145/2968332


Homburg, T., Boochs, F., 2019. Situation-Dependent Data 
Quality Analysis for Geospatial Data Using Semantic Technologies, in: 
Abramowicz, W., Paschke, A. (Eds.), Business Information Systems Workshops. Springer International 


Context-Aware Geographical Information Retrieval, in: 
Barnaghi, P., Moessner, K., Presser, M., Meissner, S. (Eds.), 
Smart Sensing and Context. Springer Berlin Heidelberg, Berlin, 
Heidelberg, pp. 77–92. https://doi.org/10.1007/978-3-642- 
04471-7_7

Quality Assessment. University of Waterloo, Waterloo, Ontario, 
Canada.

Miri-Abolfazl Mostafavii, Geoffrey Edwards, Robert Jeansoulin, 
2004. An ontology-based method for quality assessment of 
spatial data bases. Presented at the Third International 
Symposium on Spatial Data Quality, Bruck an der Leitha.

Mobasher, A., 2017. A Rule-Based Spatial Reasoning Approach 
for OpenStreetMap Data Quality Enrichment; Case Study of 
https://doi.org/10.3390/s17112498


W3C, 2017. Shapes Constraint Language (Scha1).

W3C, 2013. RIF- Basic Logic Dialect.

W3C, 2011. SPIN - Overview and Motivation.

W3C, n.d. SWRL: A Semantic Web Rule Language Combining OWL and RuleML.

W3C Geospatial Incubator Group, 2007. GeoOWL.

