

Quality Metrics to Measure the Standards Conformance of Geospatial Linked Data

Beyza Yaman^{1*} and Kevin Thompson² and Rob Brennan¹

¹ ADAPT Centre, Dublin City University, Dublin, Ireland

² Ordnance Survey Ireland, Dublin, Ireland

{beyza.yaman,rob.brennan}@adaptcentre.ie, kevin.thompson@osi.ie

Abstract. This paper describes three new Geospatial Linked Data (GLD) quality metrics that help evaluate conformance to standards. Standards conformance is a key quality criteria, for example for FAIR data. The metrics were implemented in the open source Luzzu quality assessment framework and used to evaluate four public geospatial datasets that showed a wide variation in standards conformance. This is the first set of Linked Data quality metrics developed specifically for GLD.

1 Introduction

Geospatial data has long been considered a high value resource. As societal dependence on accurate real time geo-positioning and contextualisation of data increases, so do the quality demands on geospatial data. However, all geospatial data is subject to measurement error and variation in quality. Defects also arise during the data lifecycle: digitalization, curation, transformation and integration of geospatial measurements and metadata all have risks. In the past, quantifying positional accuracy was sufficient for geospatial data quality, but now it is essential to meet broader requirements like adherence to FAIR (Findable, Accessible, Interoperable and Reusable) data principles [7]. Also agencies such as the European Commission (EC) or the United Nations (UN) highlight the role of standards conformance for achieving the FAIR principles.

As Geospatial Linked Data (GLD) publication continues to grow, methods are needed to monitor the standards conformance of GLD. Quality assessments of GLD have been previously conducted, but none of them used GLD-specific metrics. These assessments instead reuse generic methods that cannot reveal the extent of GLD standards conformance besides there is no current tools to assess GLD standards conformance at present. One study relies on hard to scale crowdsourced evaluations rather than automated quality metrics [3], another uses generic Linked Data quality metrics [4] and the final study is tied to a custom ontology predating GLD standardisation [5].

Having said that managing data quality throughout the data pipeline and lifecycle is key to the organizations such as Ordnance Survey Ireland (OSi) due

* “Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).”

to having geospatial data printed as cartographic products or data sales and distribution at data.geohive.ie [2]. Moreover, the United Nations Global Geospatial Information Management (UN-GGIM) framework³ highlights the importance of standards conformance of data for quality. Thus, there is a need for monitoring and reporting on the standards conformance of OSi GLD. It is required to quantitatively measure, and to provide continuous upward reporting to the Irish government, European Commission and UN; enable more sophisticated data quality monitoring within the organisation and provide feedback to managers within OSi for engineering teams.

This paper investigates the research question: To what extent can quality metrics derived from geospatial data standards be used to assess the standards conformance and quality of GLD? Thus, a set of applicable standards were reviewed for GLD from the International Standards Organization (ISO), Open Geospatial Consortium (OGC) and World Wide Web Consortium (W3C) to identify a set of testable conformance points for each standard. Then, in consultation with the OSi quality team, a new set of metrics were prioritised and developed to evaluate each conformance point. The first three metrics which were implemented in the Luzzu open source quality assessment framework are presented here. A set of existing open GLD datasets were then evaluated for standards conformance quality by performing metric computation⁴.

This work was realized as a part of *LinkedDataOps* project⁵ [8] implementing an e2e quality assessment framework based on the Luzzu framework [1]. The project defines roles and responsibilities to ensure liability for data quality with policies and procedures. It supports the process by means of the proposed standards while maintaining the performance for good decision-making. Continuous validation of quality and standards conformance will be performed by data experts and engineers in the OSi data production pipeline using an e2e standards reporting tool developed in this project.

The contributions of this paper are three new GLD quality metrics for standards conformance and a study of their use on open GLD datasets. The rest of this paper is structured as a description of the new metrics and a discussion of an evaluation of open GLD datasets using them.

2 Three New GLD Metrics for Standards Conformance

A year-long series of internal workshops with stakeholders across OSi identified and evaluated a set of relevant standards for GLD. The standards were in two main groups: Geospatial datasets and metadata (ISO/TC 211 Geographic information/Geomatics committee ISO 19000 series) and Geospatial Linked Data (OGC's GeoSPARQL and W3C Best Practises for Spatial Data and Web of Data recommendations) totaling 15 standards including data, metadata and

³ http://ggim.un.org/meetings/GGIM-committee/8th-Session/documents/Standards_Guide_2018.pdf

⁴ <https://github.com/beyzayaman/standard-quality-metrics>

⁵ linkeddataops.adaptcentre.ie

schema definitions. Suitable conformance points were identified, e.g., OGC’s GeoSPARQL defines 30 requirements for GLD and there are 14 best practices identified for GLD by W3C [6]. Each of these could become a target conformance point and metrics developed to measure it.

In consultation with OSi, the most essential conformance points were identified. Metric ease of implementation was also used to guide the choices to enable rapid prototyping. Three new metrics were implemented (M1 from OGC GeoSPARQL⁶ and ISO 19125-1⁷, M2 from W3C Best Practices [6] and M3 from ISO 19157⁸). Together these metrics enable the assessment of a dataset in terms of standards conformance including metadata, spatial reference systems and geometry classes. Note that each metric described below must compute their rate (Equation 1) over the whole dataset to give values in the range of [0-1] as is best practice for quality metrics and this is not repeated in each definition, instead the base calculation method (\bar{e}) is the set of instances conforming the conditions for each metric which must then be inserted into Equation 1 for each metric ((Prefix for geo: <http://www.opengis.net/ont/geosparql#>):

$$M = \sum_{i=1}^e \frac{\bar{e}(i)}{\text{size}(e)} \quad (1)$$

M1, Geometry Extension Object Consistency Check: This metric addresses the requirement “All RDFS Literals of type `geo:wktLiteral` shall obey a specified syntax and ISO 19125-1.”. According to the OGC GeoSPARQL requirements, WKT serialization regulates geometry types with ISO 19125 Simple Features, and GML serialization regulates them with ISO 19107 Spatial Schema. *Metric Computation:* If the entity in the dataset is a member of class `geo:Geometry` then this metric checks the rate of employed `geo:asWKT` or `geo:asGML` properties in the dataset.

$$\bar{e} := \{e | \forall e \in \text{class}(\text{geo} : \text{Geometry}) \cdot \text{hasWKT}(e) \vee \text{hasGML}(e)\}$$

M2, Links to Spatial Things Check: This metric addresses the requirement “Use appropriate relation types to link Spatial Things where source and target of the hyperlink are Spatial Things”. Thus, W3C suggests using appropriate relation types to link *Spatial Things* which is any object with spatial extent, (i.e. size, shape, or position) such as people, places [6]. *Metric Computation:* The metric detects the rate of entities having links to external spatial things in other datasets and internal spatial links within the dataset.

$$\bar{e} := \{e | \forall e \in \text{class}(\text{geo} : \text{Geometry}) \cdot \text{hasST}(e)\}$$

M3, Consistent Polygon and Multipolygon Usage Check: This metric addresses the requirement “Polygons and multipolygons shall form a closed circuit”. Polygons are topologically closed structures, thus, the starting point and end point of a polygon should be equal to provide a consistent geometric

⁶ <https://www.ogc.org/standards/geosparql>

⁷ <https://www.iso.org/standard/40114.html>

⁸ <https://www.iso.org/standard/32575.html>

Table 1. Quality Assessment Results for datasets

Metric	OSi	OS UK	LinkedGeoData	Greek GLD
M1	1	0	1	0
M2	0.36	0.84	1	0
M3	1*	1	0	0.5

shape. *Metric Computation:* This metric checks the equality of the starting and end points of polygons. Each polygon in a multipolygon must be checked.

$$\bar{e} := \{e | \forall e \in class(geo : Geometry) \cdot (hasClosedPolygon(e))\}$$

3 Evaluation

Four open GLD datasets were assessed using the new metrics implemented in Luzzu (Table 1). This section discusses the performance of each dataset w.r.t. the given metrics. The datasets were: OSi’s Irish national mapping Linked Open Data⁹. Ordnance Survey UK’s United Kingdom mapping Linked Open Data¹⁰. LinkedGeoData¹¹ is provided by the University of Leipzig by converting OpenStreetMap data to Linked Data. Greece LD¹² is provided by the University of Athens as part of the TELEIOS project. The metric values shown in Table 1 are the mean value of the metric for all GLD resources in the dataset. Specific discussion on each metric’s results is provided below.

Geometry Extension Object Consistency Check (M1): OS UK and Greek LGD does not conform to the standards due to the use of non-standard, specialized ontologies in the dataset (e.g., `strdf:WKT` (Prefix for `strdf: http://strdf.di.uoa.gr/ontology#`) instead of `geo:wktLiterals`). OSi and LinkedGeoData conform to the standards for every geospatial entity in the dataset. **Links to Spatial Things (M2):** It was seen that while LinkedGeoData dataset has links to the GADM dataset¹³, the OSi has links to Logainm dataset¹⁴. OS UK provides two different granularities in county and Europe within the dataset. This shows that every LinkedGeoData instance has a connection with another spatial thing and the dataset has the highest interoperability between datasets. **Polygon and Multipolygon Check (M3):** In particular, it was seen that OSi, OS UK, Greek GLD have polygons and multipolygons included in their dataset, whereas entities are only represented by points in LinkedGeoData, and waterlinestring by some Greek GLD (note that full OSi was computed with sampling so it is estimated and denoted with *). This means that the data in LinkedGeoData and Greek GLD were not represented (or partially represented) as boundaries. This is due to having different spatial dimensions rather than

⁹ <http://data.geohive.ie/downloadAndQuery.html>

¹⁰ <https://data.ordnancesurvey.co.uk/datasets/boundary-line>

¹¹ http://linkedgeodata.org/Datasets?show_files=0

¹² <http://linkedopendata.gr/dataset>

¹³ <http://gadm.geovocab.org/>

¹⁴ <https://www.logainm.ie/en/inf/proj-machines>

geospatial polygons data whereas it is very important for GIS applications e.g. Polygons are essential to building things, as otherwise it is not known where anything begins and ends.

4 Conclusions

Three new GLD quality metrics have been defined based on analysis of GLD standards. They have been implemented in the Luzzu quality assessment framework and used to assess four open GLD datasets. This has shown that *i*) it is fruitful to use standards conformance points as a basis for new quality metrics and *ii*) that despite the availability of best practice advice and standards for GLD, there is still a very low level of conformance to GLD standards in the GLD cloud. The ability to make this standards conformance assessment in an objective, quantitative, automated way is an advance in the state of the art. The metrics have limitations due to their simplicity and the flexibility of Linked Data and hence the heterogeneity of real datasets. However, this approach is still useful for publishers like OSi who wish their data to conform to the requirements and best practices published by standardization organisations.

Acknowledgements. This research received funding from the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 801522, by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology [grant number 13/RC/2106] and Ordnance Survey Ireland.

References

1. J. Debattista, S. Auer, and C. Lange. Luzzu—a methodology and framework for linked data quality assessment. *Journal of Data and Information Quality (JDIQ)*, 8(1):1–32, 2016.
2. C. Debruyne, A. Meehan, É. Clinton, L. McNerney, A. Nautiyal, P. Lavin, and D. O’Sullivan. Ireland’s authoritative geospatial linked data. In *International Semantic Web Conference*, pages 66–74, 2017.
3. R. Karam and M. Melchiori. Improving geo-spatial linked data with the wisdom of the crowds. In *Proceedings of the joint EDBT/ICDT 2013 workshops*. ACM, 2013.
4. J. Lehmann, S. Athanasiou, A. Both, A. García-Rojas, G. Giannopoulos, D. Hladky, J. J. Le Grange, A.-C. N. Ngomo, M. A. Sherif, C. Stadler, et al. Managing geospatial linked data in the geoknow project., 2015.
5. M.-A. Mostafavi, G. Edwards, and R. Jeansoulin. An ontology-based method for quality assessment of spatial data bases. 2004.
6. J. Tandy, L. van den Brink, and P. Barnaghi. Spatial data on the web best practices. *W3C Working Group Note*, 2017.
7. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.
8. B. Yaman and R. Brennan. Linkeddataops:linked data operations based on quality process cycle. In *EKAW (Posters & Demos)*, 2020.