Standards Conformance Metrics for Geospatial Linked Data

Beyza Yaman[1] and Kevin Thompson[2] and Rob Brennan[1]

[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 a set of new Geospatial Linked Data (GLD) quality metrics based on ISO and W3C spatial standards for monitoring geospatial data production. The Luzzu quality assessment framework was employed to implement the metrics and evaluate a set of five public geospatial datasets. For data publishers like Ordnance Survey Ireland (OSi), standards compliance of geospatial data is a key quality criteria that has often been overlooked. Despite the availability of metrics-based quality assessment tools for Linked Data, there is a lack of dedicated quality metrics for GLD and no metrics based on geospatial data standards and best practices. This paper provides nine new metrics and a first assessment of public datasets for geospatial standards compliance. Our approach also demonstrates the effectiveness of developing new quality metrics through analysis of the requirements defined in relevant standards.

1 Introduction

Geospatial data has long been considered a high value resource. Geospatial Linked Data (GLD) is ideally positioned to provide an open, web-based mechanism to exchange and interlink these geospatial entities for emerging national geospatial data infrastructures. As societal dependence on accurate real time geopositioning and contextualisation of data increases, so do the quality demands on geospatial data. All geospatial data is subject to a degree of measurement error. Further issues can occur during the data lifecycle: digitalization, curation, transformation and integration of geospatial measurements and metadata all have risks. Typically, spatial measurements must be integrated into a digital twin of an entity which can include or interlink topographical, political, historical, environmental and other factors. Thus, producing and updating geospatial data is expensive [10]. In the past, quantifying positional accuracy was sufficient, but now geospatial data must also comply with broader usage requirements such as the FAIR data principles [15].

There are a number of relevant standards and best practices for publishing high quality geospatial data and Linked Data. This includes ISO standards, Open Geospatial Consortium (OGC) standards, and W3C Spatial Data on the Web Best Practices (SDOTW). However, data publishers are not always able to assure that they produce data conforming to them. Thus, it is important for publishers to have quality assessment processes and tools to provide assurance and enable improvement of these complex data life-cycles. Unfortunately, no existing tools directly address standards compliance for GLD.

Several quality assessments of GLD have previously been conducted [10,9,12] but one of them relies on crowdsourced evaluations rather than automated metrics [9], another one provides a generic Linked Data quality assessments of the data that is not specific to geospatial concerns [10] and the other is tied to a custom ontology predating GLD standardisation [12]. In contrast, our starting point was to examine the relevant standards for geospatial data and to develop new quality metrics targeted at the GLD domain.

The research question investigated in this work is: To what extent can quality metrics derived from geospatial data standards be used to assess the standards compliance and quality of GLD. Thus, we reviewed the applicable standards for GLD from the ISO, OGC and W3C to identify a set of compliance points for each standard and a set of testable recommended best practices. Then a new set of metrics were developed to evaluate each compliance point. The metrics were implemented in the Luzzu open source quality assessment framework. A set of existing open GLD datasets were then evaluated for standards compliance quality by performing metric computation. All metrics developed here are described in the daQ vocabulary and published as an open resource for the community [3]. This paper is an extension of the previous work accepted from ISWC2020 poster&demo session [17] where an initial 3 metrics were proposed. In this work we describe 6 additional metrics which are developed as a part of LinkedDataOps project [16].

Our contributions are : i) Identification of standards conformance points for GLD across ISO, OGC and W3C standards; ii) Design and open source implementation of 9 new standards-based geospatial quality metrics in the Luzzu framework and providing a set of daQ ontology [3] describing the metrics; iii) providing a first comparative quality survey of public GLD datasets in terms of standards compliance. The remainder of this paper is organized as follows: Section 2 describes the motivational OSi use case, section 3 summarizes related works and background including GLD infrastructure and Luzzu framework. Section 5 discusses our approach including standardization proposals and defines the new metrics. We present the evaluation, our experiments and analysis of the results in Section 6. Conclusions and future work are discussed in Section 7.

2 Use Case

National mapping agencies such as Ordnance Survey Ireland (OSi) are now geospatial data publishers more than cartographic institutions. The United Nations Secretariat: Global Geospatial Information Management publishes detailed advice for national Integrated Geospatial Information systems. They identify the importance of managing data quality and the role of standards in effective geospatial information systems. OSi's national geospatial digital infrastructure

[3] https://github.com/beyzayaman/standard-quality-metrics

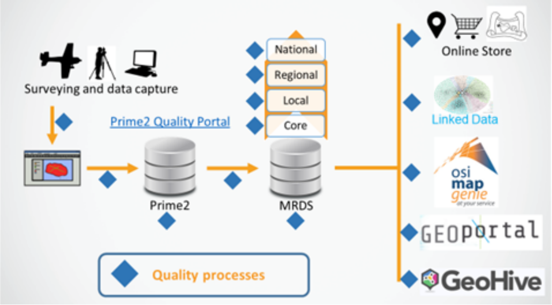

Fig. 1. OSi Geospatial Information Publishing Pipeline with Quality Control Points

(Fig. 1) encompasses surveying and data capture, image processing, translation to the Prime2 object-oriented spatial model of over 50 million spatial objects tracked in time and provenance, conversion to the multi-resolution data source (MRDS) database for printing as cartographic products or data sales and distribution at data.geohive.ie [5]. These services run on a state of the art Oracle Spatial and Graph installation supporting both relational and RDF models. Managing data quality throughout the data pipeline and lifecycle is key to OSi (Fig. 1) and there are already quality checks on the data quality dimensions of positional accuracy, logical consistency, completeness, representational conciseness, syntactic validity, positional validity and semantic accuracy. Current data quality assessment within OSi depends on i) two automated tools: the rules-based 1Spatial 1Integrate and Luzzu for Linked Data [4] and ii) manual or semi-automated techniques by domain experts.

Moreover, the United Nations Global Geospatial Information Management (UN-GGIM) framework 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. For example, to provide continuous upward reporting to the Irish government, European Commission, UN; and provide feedback to managers within OSi for engineering team. Through a year-long series of internal workshops with stakeholders across the organisation the following requirements were addressed: ***i)*** **Req 1:** Identification of the relevant standards conformance points for GLD. ***ii)*** **Req 2:** Metrics for assessment of GLD for geospatial standards conformance. ***iii)*** **Req 3:** Quality assessment tools and processes for geospatial standards conformance. ***iv)*** **Req 4:** Fusion with existing quality metrics for visualization, analysis and reporting. Requirement showed a necessity for a new GLD standards conformance quality monitoring approach as for the first time quality is being measured in terms of standards compliance and specifically for the geospatial domain.

3 Related Work

GLD allows us to link between spatial objects. GLD enables richer models of real world entities with spatial dimensions and locations, that can be accessed with spatial references or queries. This section discusses traditional and Linked Data solutions for geospatial data quality and standards conformance assessment.

Non-Semantic Approaches to Geospatial Data Quality Assessment Rules-based quality assessment, as implemented in the 1Integrate tool suite for spatial systems such as Oracle Spatial and Graph, is flexible and often used for implementing data cleansing as well as quality assessment. In practice, rules definitions are expensive to develop and maintain. Luzzu framework is useful as it generates self-describing plug and play metrics and quality observations metadata. Scalability is another area in which rules-based assessment can fail. Execution of the explicit rules over 50 million spatial objects can take days, even on custom high end hardware,like an Oracle exadata platform. Especially when large-scale data transformations must be carried out (for example for schema updates or to fix systematic errors identified in older releases) then the time required is unsustainable. Using probabilistic (sampling-based) metrics, as deployed in Luzzu, for computationally expensive metrics is an advantage.

GLD Quality Assessment The LinkedGeoData [2] and GeoLinkedData [11] projects study spatial features in their datasets, however, the quality assessment step is not addressed in their work. The most notable project is GeoKnow [10] which assesses spatial data quality using standard quality metrics i.e. no metrics addressing the specific requirements or models of GLD were used. However, most of the implemented metrics provide statistical summaries for the data by looking at the coverage of the data instead of specific geospatial measures.

Semantic approaches have significant advantages as interpretative frameworks for quality results. Quality assessments are published using the W3C data quality vocabulary [14] or dataset quality vocabulary [3]. The results are categorised into specific quality dimensions (e.g. consistency) using the taxonomy (e.g. Zaveri et al. [18]) into hierarchies when analysing geospatial data quality [6]. Another advantage is the ability of semantic models to encompass multiple data quality models through R2RML mappings or data quality observations. They allow observations easily consumed in tool chains (e.g. analysis dashboards) compared to the non-standard outputs of proprietary tools, such as 1Integrate, that require domain expert knowledge to develop extraction rules.

The Luzzu framework is employed in this study to take advantage of the advantages of a semantic data quality approach. Luzzu [4] is an open-source, Java based, Linked Data quality assessment framework which allows users to create custom quality metrics to produce a time series of quality observations about datasets. This is an interoperable tool allowing ontology driven backend to produce machine readable quality reports and metadata about the assessment results. The quality metadata is represented by domain independent daQ ontology based on W3C RDF Data Cube and PROV-O vocabularies.

4 A New Assessment Method for Standards Conformance of Geospatial Linked Datasets

In this section, we will discuss the process of creating the new geospatial metrics and introduce the new metrics. The followed process was to *i)* identify a list of geospatial data standards conformance points or best practice recommendations and *ii)* prioritise the list (Section 4.1); *iii)* devise a set of new quality metrics for automated dataset assessment in a quality framework (Section 4.2). Each metric was then assigned to an appropriate quality dimension as defined by Zaveri *et al.* [18] and a set of semantic metric descriptions created using the daQ ontology. The descriptions facilitated collection of rich dataset quality observations as W3C data cubes for further analysis, integration and visualisation.

4.1 Identifying Geospatial Data Quality Standards Conformance Points

In this section we identify, evaluate and compare a set of relevant standards and recommendations for GLD quality proposed by the Open Geospatial Consortium (OGC), ISO and W3C. The ISO/TC 211 Geographic information/Geomatics committee defines geographic technology standards in the ISO 19000 series [1] as well as the OGC creates open geospatial standards. The both organizations have close connections such that some documents prepared by OGC are adopted by ISO or implemented by the collaboration of both parties. We evaluate the standards in 3 main groups:

Geospatial datasets: ISO 19103, 19107, 19108, 19109, 19112, 19123, 19156[1] are published to describe the data, in particular the schema, spatial referencing by geospatial data, and methods for representing geographical data and measurements. OGC equivalence of the documents can be seen on the right hand side of the table. Old ISO 19113/19114/19138 are combined to 19157 data quality standards. Thus, while ISO 8000 defines data quality concepts and processes for generic information systems, ISO 19157 and ISO 19158 provide more detailed guidence on data quality practices for geospatial data. ISO 19158 specifies metrics and measurements for evaluation of data quality elements at different stages of the geospatial data lifecycle. It also defines quality metric evaluation by using aggregation methods and thresholds. ISO 19157 defines a set of data quality measures when evaluating and reporting data quality of geospatial data.

Geospatial metadata: ISO 19111 and 19115 describe the metadata standards for geospatial data. While ISO 19115 focuses on metadata for cataloging and profiling purposes with the extensions for imagery and gridded data; ISO 19111 describes appropriate metadata for a Coordinate Reference System.

Geospatial Linked Data: There are three relevant types of documents for data quality. *i)* ISO 19150 which guides high level ontology schema appropriate for geospatial data and rules for using OWL-DL. *ii)* OGC's GeoSPARQL standard that define a set of SPARQL extension functions for geospatial data, a set of RIF rules and a core RDF/OWL vocabulary for geographic information based on the General Feature Model, Simple Features, Feature Geometry and SQL MM [13]. *iii)* W3C has two documents, first the Data on the Web Best Practices recommendation for improving the consistency of data management and secondly the SDOTW working group note which complements the earlier recommendation but is specialized for geospatial data.

In total OGC's GeoSPARQL defines 30 requirements for geospatial data and there are 14 best practices identified for geospatial data by W3C (Table 1).

Table 1. OSi Priority Standards Compliance Points Identified

Origin	Req.	Description
OGC	R1	Implementations shall allow the RDF property geo:asWKT or geo:asGML to be used in SPARQL graph patterns.
OGC	R2	All RDFS Literals of type geo:wktLiteral shall obey a specified syntax and ISO 19125-1.
OGC	R3, R4	Implementations shall allow the RDFS class geo:Geometry with geo:hasGeometry property to be used in SPARQL graph patterns.
OGC	R5	Implementations shall allow the RDFS class geo:Geometry with geo:hasDefaultGeometry property used in SPARQL graph patterns.
W3C	R6	Use spatial data encodings that match your target audience.
W3C	R7	Use appropriate relation types to link Spatial Things where source and target of the hyperlink are Spatial Things.
ISO	R8	Polygons and multipolygons shall form a closed circuit.
ISO	R9	Provide information on the changing nature of spatial things.

Each of these may be used to construct standards compliance quality metrics. A possible set of metrics was proposed and discussed in a series of workshops with OSi staff drawn from the Geospatial Services, Data Governance & Quality department. First of all, OSi data quality system requirements and background discovered in meetings and workshops. It served as a basis for further development of OSi data quality governance in project. We first described the background in the form of the existing OSi data publishing pipeline and the available quality assessment points.

Following, the concepts and architecture of the end to end data quality portal initiative are described and both existing metrics and new sources of metrics for the OSi end to end quality monitoring framework are discussed. We evaluated the quality of the existing data in OSi with Luzzu framework. We used a set of generic quality metrics and appointed a threshold to ensure the conformance of the datasets to the given indicators. However, due to generic structure of the metrics it was not possible to evaluate the datasets according to the geospatial dimensions. Also the legislations, regulations and standardization requirements by organizations such as OECD, UN, and EC are needed to comply to ensure the reliant governance of the data in the public agencies. The aim is efficiency in the provision of public services. The requirements and best practices in Table 1 were identified as high priority for the initial deployment.

A major focus of developing an end to end quality governance system for OSi is to establish a set of new metrics that will give OSi the ability to monitor and report on quality in a way that can satisfy their customers and unique requirements of the geospatial domain. Thus, initial candidate standards which are discussed in Section 4.2 were considered as a set of standards for OSi to measure its compliance with. It was seen most crucial to enable publication of the associated data in-line with agreed standards. In accordance with OSi staff, most essential metrics were chosen *i)* to check the usage of chosen standards to reduce the heterogeneity in representing data, *ii)* to measure the discoverability

Table 2. New Geospatial Standards Conformance Quality Metrics

Req.	**ID**	**Metric Name**	**Dimension**
R1	CS-M1	Geometry Extension Property Check	Completeness
R2	CS-M2	Geometry Extension Object Consistency Check	Completeness
R3,R4	CS-M3,CS-M4	Geometry Classes and Properties Check	Completeness
R5	CS-M5	Spatial Dimensions Existence Check	Completeness
R6	I-M6	Links to Spatial Things (internal&external)	Interlinking
R7	I-M7	Links to Spatial Things from popular repositories	Interlinking
R8	CY-M8	Polygon and Multipolygon Check	Consistency
R9	T-M9	Freshness Check	Timeliness

and freshness of the data to see the impact on the usage by the LOD cloud users *iii)* to measure the consistency of the data to provide high uniformity *iv)* to measure the completeness of the data to provide high coverage to the users. Thus, the metrics in Table 2 are chosen representing each dimension with the feedback of the OSi staff. We have chosen to implement different standardization metrics to demonstrate the potential of developing any metrics *w.r.t.* the required standards. Furthermore, we collaborated with OSi using different means of communications such as "basecamp" and "gogs" for the efficient development. In the following section we will introduce the implemented metrics.

4.2 New Geospatial Data Quality Metrics

Nine new metrics are defined here for the nine priority conformance points identified in the last section. Design principles were used for effective data quality metrics for both decision making under uncertainty and economically oriented data quality management[8]. Together these metrics enable the assessment of a dataset in terms of standards conformance including metadata, spatial reference systems and geometry classes. Each metric is identified by their quality dimension, summarised in Table 2 and discussed below in detail.

Geometry Extension Property Check (CS-M1): This metric addresses requirement R1 "Implementations shall allow the RDF property geo:asWKT or geo:asGML to be used in SPARQL graph patterns". Thus, conformant GLD datasets must have at least one geometry property associated with individuals which are geospatial features. Two properties are allowed by the GeoSPARQL standard, well known text (WKT) or geography markup language (GML). Both OGC and ISO standards rely on WKT geometries and GML serialization.

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. This is evaluated using functions as *hasWKT(e)* or *hasGML(e)* which return a boolean value. The metric is computed as a rate over the whole dataset as follows (Note that the following metrics also compute their rate over the whole dataset and thus Equation 1 will not be repeated in each metric definition):

$$\sum_{i=1}^{e} \frac{\overline{e}(i)}{size(e)} \tag{1}$$

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

Geometry Extension Object Consistency Check (CS-M2): This metric addresses requirement R2 "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 [ISO 19125-1], and GML serialization regulates them with ISO 19107 Spatial Schema.

Metric Computation: This metric checks the conformance of the dataset to the serialization requirement of OGC GeoSPARQL by checking the conformance of objects in terms of the order of use of coordinate system URI, spatial dimension and literal URI. Geometry data should consist of an optional URI identifying the coordinate reference system (e.g., CRS84, WGS 84) followed by WKT describing a geometric value. Spatial dimension may include polygon, multipolygon, line, point, or multilinestring shapes. Finally, the syntax should include the `geo:wktLiteral` URI declaring the object is a literal.

$\overline{e} := \{e | \forall e \in class(geo : Geometry) \cdot hasCRSURI(e) \wedge hasSpatialDimension(e) \wedge hasWKTLiteral(e))\}$

Geometry Classes and Properties Check (CS-M3,CS-M4): These metrics address requirements R3 and R4 "Implementations shall allow the RDFS class `geo:Geometry` with `geo:hasGeometry` and `geo:hasDefaultGeometry` properties to be used in SPARQL graph patterns". OGC requires that each geometry object is an individual of the root geometry class `geo:Geometry`. In addition, a `geo:Feature` should be related to a geometry describing its spatial extent via the `geo:hasGeometry` property. The `geo:hasDefaultGeometry` property is also required to link a feature with its default geometry.

Metric Computation: This metric checks the rate of declaration of geometry classes and properties in the datasets. The *hasGeometry(e)* and *hasDefaultGeometry(e)* functions check each entity and return a boolean value for property existence. The metric checks each entity which is an individual of the `geo:Geometry` class.

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

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

Spatial Dimension Existence Check(CS-M5): This metric addresses requirement R5 "Use spatial data encodings that match your target audience". W3C SDOTW suggests encoding in a useful way such that machines can decode and process the encoded data using *Spatial Dimension* which is the measure of spatial extent, especially width, height, or length.

Metric Computation: This metric assesses the rate of spatial dimension properties related to each entity in the dataset. It compares the total number of spatial dimensions (multipolygon, polygon, line, point, multilinestring) described for each entity in the dataset to the overall number of entities.

$\overline{e} := \{e | \forall e \in class(geo : Geometry) \cdot (isMultipolygon(e) \vee isPolygon(e) \vee isLine(e) \vee isPoint(e) \vee isMultilinestring(e))\}$

Links to Spatial Things Check (I-M6, I-M7): This metric addresses requirement R6 "Use appropriate relation types to link Spatial Things where source and target of the hyperlink are Spatial Things". Thus, W3C SDOTW 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 [14]. W3C SDOTW suggests two types of links for Spatial things: i) links to other spatial things using an object with its own URI within dataset or to other datasets decreasing the computational complexity and enriching the data semantically ii) links to spatial things from popular repositories which increases the discoverability of the dataset. However, the challenge in this metric is that it is not possible to understand if a link has spatial extent without visiting the other resource. Thus, first a set of different pay-level-domains are detected manually and according to the used schema, the rate of the links are computed as an efficient approximation.

Metric Computation: First the metric detects the rate of entities having links to external spatial things in other datasets and internal spatial links within dataset. In I-M6, the *hasST(e)* function checks the entities with these links and later this number is divided into the overall number of entities.

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

Metric Computation: This metric detects the rate of entities having links to external spatial links in popular and highly referenced datasets. In this work, we specifically looked at the usage of DBpedia, Wikidata and Geonames datasets. We counted the entities with these links and divided to the overall entity number.

$\overline{e} := \{e | \forall e \in class(geo : Geometry) \cdot (isDBpedia(e) \vee isWikidata(e) \vee isGeonames(e)))\}$

Consistent Polygon and Multipolygon Usage Check(CY-M8): This metric addresses requirement R7 "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 shape.

Metric Computation: This metric checks the equality of the starting and end points of polygons. Each polygon in a multipolygon must be checked. We measure the rate of correctly described polygons and multipolygons in a dataset. In metric CY-M8 the function `hasClosedPolygon(e)` detects the correct usage for each entity in the dataset.

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

Freshness Check (T-M9): This metric addresses requirement R8 "Provide information on the changing nature of spatial things". According to ISO and W3C it is crucial to provide the provenance information about when data has changed during their lifecycle.

Metric Computation: This metric checks the age of the data (f) by looking at the creation time and when it was last updated to the recent version. This metric was used as an updated version from [4]. In this formula, Volatility (v) is "the length of time the data remains valid" which is analogous to the shelf life of perishable products; Currency (c) is "the age of the data when it is delivered to the user" [7]. This metric is computed at the dataset and not instance level level due to lack of information in the entity level.

$f = (max(1 - c/v, 0))$

4.3 Semantic Metric Models

Semantic metric models were created for all of the metrics described above as follows. The daQ ontology was extended with the new metrics by inheriting upper daQ concepts. Then each metric was classified under Linked Data quality dimensions and categories as presented in Zaveri *et al.* and descriptive metadata added. For example the WKT Property metric is classified under the Completeness dimension in the Intrinsic category. This allowed us to produce and publish daQ machine readable metadata as Linked Data for further processing such as metric fusion, visualisation or root cause analysis.

5 Evaluation

This section describes a first study showing our new metrics in operation with experimental set-up in Section 5.1 followed by results in Section 5.2, the usability of the defined metrics in Section **??** and the lessons learned in Section 5.3.

5.1 Experimental Setup

Experiments were executed to measure the metrics' ability to detect the standards compliance of GLD datasets, as well as, the extent of standards compliance of published Open GLD to meet OSi's requirements. Investigation was performed by implementing new metrics as scalable Luzzu plug-ins in Java and assessing a set of four open GLD datasets. We used a computer with Intel i7 8th generation processor and 8GB memory.

Datasets: Major open topographical geospatial datasets describing political or administrative boundaries were chosen to ensure geometrical features were represented in each dataset. Despite this selection, there is considerable variation in the datasets in number of triples, size, languages and used coordinate reference systems (CRS) as depicted in Table 4. Ordnance Survey Ireland (OSi) is the national mapping agency of Ireland and they publish a subset of their data as Linked Open Data. The OSi boundaries dataset describes political and administrative boundaries in Ireland. Ordnance Survey UK is the national mapping agency of the United Kingdom and they also publish their data partially

Table 3. Dataset Summary

Dataset	#Triple	Size	Languages	CRS
OSi	1936763	274M	EN,GA	IRENET95 / ITM
OS UK	64641	224.1M	EN	WGS 84
LinkedGeoData	464193	1.5G	EN,Various	WGS 84
Greece LD	24583	183M	EN,GR	WGS 84

as Linked Data. LinkedGeoData is provided by the University of Leipzig by converting OpenStreetMap data to Linked Data. Greece LD is provided by the University of Athens as part of the TELEIOS project.

Method: Assessments were performed on each dataset using the Luzzu framework. In addition to assessing the full datasets, subset were also assessed to provide a common baseline for comparison between datasets. Observations for the nine metrics presented in Section 4.2 were collected as quality metadata using the daQ vocabulary[4] as mentioned in Section 4.3.

5.2 Results and Discussion

This section discusses the performance of each dataset w.r.t. the given metrics. The metric values shown are the average value of the metric for all GLD resources in the dataset (Table 4). The table also shows the mean observed values of each metric across all datasets (last column) and the mean of all metrics for each dataset as simple aggregated quality indicator (last row).

In general we see that most datasets either conform or do not conform to specific standards and hence individual metrics score 1 or 0. Nonetheless the aggregated metric value gives an insight into the overall level of standards compliance for a specific dataset. However these relative scores should not be interpreted as an absolute statement of quality. Choice of which metrics are relevant for a specific application or dataset is always a key quality management decision. It can be said that OSi have selected these metrics as important for their datasets and thus these metrics help OSi monitor quality. Note that standards compliance is not the same as functional capability, thus using a non-standard ontology to express GLD may grant the same or better capabilities but from the user's perspective it may be more difficult to use (requiring mappings, query-re-writing etc.) and thus having a lower quality from the perspectives of "fitness for use" or "adherence to standards" [18].

It is interesting to note that sub-datasets such as the OSi parishes sample can have quite different standards compliance metrics scores than their parent datasets. This partially due to the scale and complexity of national spatial data collection which is an ongoing task with evolving requirements, methods and teams contributing to maintaining an overall dataset composed of many contributions over time. Specific results for each metric are discussed below.

Geometry Extension Property Check (CS-M1): The Greek LGD and OS UK score zero (non compliant). The Greek LGD doesn't use required properties (`geo:asWKT`[5] and `geo:asGML`) and OS UK uses a property from their specialized ontology instead. OSi and LinkedGeoData use the standard properties. A drawback of this metric is that it requires a specific vocabulary, but that reflects what the standards require for conformance. Adding support for inference like property inheritance is useful in theory but given the practicalities of closed world data quality assessment and Linked Data publishing practices it is not necessary for a useful implementation.

[4] https://github.com/beyzayaman/standard-quality-metrics

[5] Prefix for geo: `http://www.opengis.net/ont/geosparql#`

Table 4. Quality Assessment Results for datasets

Metric Name	**OSi Full**	**OSi parishes**	**OS UK parishes**	**LinkedGeoData boundaries**	**Greek GLD coastlines**	**Greek GLD water bodies**	**Mean**
CS-M1	1	1	0	1	0	1	0.66
CS-M2	1	1	0	1	0	0	0.5
CS-M3	1	1	0	0	0	1	0.5
CS-M4	0	0	0	0	0	0	0
CS-M5	1	1	1	1	1	1	1
I-M6	0.36	0.94	0.84	1	0	0	0.52
I-M7	0.142	0	0	0.0004	0	0	0.024
CY-M8	1*	1	1	-	1	-	1
T-M9	0	0	1	*	0	0	0.33
Agg. Metric	0.50	0.66	0.52	0.57	0 0.22	0.33	

Geometry Extension Object Consistency Check (CS-M2) Again 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`[6] instead of `geo:wktLiterals`). OSi and LinkedGeoData conform to the standards for every geospatial entity in the dataset.

Geometry Classes and Properties Check (geo:hasGeometry Property (CS-M3) and geo:hasDefaultGeometry Property (CS-M4): OS UK entities do not have any geometry property or class, Greek LGD use a property from their own ontology, and LinkedGeoData have used NeoGeo geometry ontology[7] all of which are different from the OGC standard. OSi is the only dataset that used OGC features but it is not complete as well because `geo:hasDefaultGeometry` was not used by any of the datasets. Even though using open standards is a requirement for 5-star Linked Data publishing, this doesn't seem to be followed by most of the publishers.

Spatial Dimensions Existence Check (CS-M5) All the datasets performed well as all the entities in the datasets have spatial dimensions provided as points, polygons, multipolygons and waterlinestrings.

Links to Spatial Things (internal&external) (I-M6): [14]. It was seen that while the LinkedGeoData dataset has links to the GADM dataset[8], the OSi full dataset has links to Logainm dataset[9], but for the parishes it doesn't have any external links. OS UK provides two different granularities in county and Europe within the dataset.

Interlinking & Links to Spatial Things from popular repositories (I-M7) DBpedia, Wikidata and Geonames were considered as popular knowledge graphs [14] and we have discovered that OSi has links to DBpedia, LinkedGeoData has links to Wikidata with the rates given in Table 4. Considering LinkedGeoData provides a wide range of properties it would have been expected to see links to DBpedia or higher ratio of links for Wikidata. Thus publishers

[6] Prefix for strdf: `http://strdf.di.uoa.gr/ontology#`

[7] `http://geovocab.org/geometry`

[8] `http://gadm.geovocab.org/`

[9] `https://www.logainm.ie/en/inf/proj-machines`

can consider using interlinking tools such as Silk[10] or LIMES[11] to enrich their data and increase the discoverability on the web.

*Aggregated Results for Interlinking:*This metric is very similar to the Debattista etal. [4] external link data providers metric which calculates all the datasets in the LOD cloud where LOD cloud has average 27% external links to other datasets [4]. Our results show that compared to the LOD cloud, these datasets have a higher rate of external spatial links but a much lower rate of links to popular datasets. If we consider the aggregated result for the Interlinking dimension, the rate is similar to LOD cloud rate with a mean of 27%.

Polygon and Multipolygon Check (CY-M8) As can be seen from the Table 4 all the datasets conform to this standard (note that full OSi was computed with sampling so it is estimated and denoted with *). 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 Greek GLD thus, we kept them outside of the computation (denoted with -). This indicates there is currently too little geospatial polygon data on the web whereas it is very important for GIS applications e.g. historians working on historic roads and boundaries[12].

Freshness Check (T-M9) : OS UK provides both creation and modification metadata for the dataset with the date of November 2019 which makes the dataset quite fresh. LinkedGeoData provides a modification date but no creation date. Hence freshness was not computed as it is based on creation time. No creation time was available in the OSi boundaries dataset but this has been fixed for the newer release of buildings data. This result confirms that provenance information is not given a high importance when publishing datasets[4].

In summary, the last row of Table 4 shows the mean aggregated GLD standards quality metric for each dataset. This could be considered as an estimate for the overall quality dimension of standards compliance for each dataset. Also it can be seen from Table 4 that different subsets of the datasets result in different scores even in the same dataset such as OSi or Greek GLD. Adoption of non-standard vocabularies decreases the scores. Overall the aggregated metric values are in the mid-range for most datasets, showing that usage of GLD standards and best practices are not widely applied by the publishers yet. A lack of standardisation has increased the heterogeneity of GLD and this makes it more difficult to use the datasets or to compare them with standardized metrics.

5.3 Lessons Learned from OSi Deployment

The adoption of semantic technology for quality metric specification and assessment in Ordnance Survey Ireland has shown the following: (1) Initially we believed that Linked Data quality assessment techniques were far advanced of the mainstream state of the art due to the obvious enhancements of the work by Zaveri et al. [18] compared to generic data quality standards like ISO 8000.

[10] https://github.com/silk-framework/silk

[11] https://github.com/dice-group/LIMES

[12] https://github.com/silknow

However geospatial data quality has a long tradition, and this reflected in the relative maturity of the ISO 19157 standard which has an extensive taxonomy of quality dimensions that go beyond the Linked Data work. We are currently working to reconcile and map between all of these standards. (2) The flexibility and self-describing nature of Linked Data for expressing data quality assessment results is very useful and this is an area in which semantic technology facilitates the unification of quality assessments from many different tools across the data production pipeline assessing diverse technologies. (3) In addition to the standard data quality dimensions it would be useful to have the ability in tools to assign metrics to custom dimensions, for example on a per standard or standards organisation basis, to enable more fine-grained, deployment-specific reporting. (4) The current dominant approach of Linked Data assessment tools addressing the entire dataset with a single set of observations is limiting when it comes to further analysis and it would be useful to have standard ways to assign metric observations to sub-sets of a dataset.

6 Conclusion and Future Work

This paper investigated to what extent quality metrics derived from geospatial data standards can be used to assess the standards compliance and quality of GLD. Nine new metrics have been defined and implemented in the Luzzu quality assessment framework. The metrics have been used to assess four open GLD datasets. This has shown that, despite the availability of best practice advice and standards for GLD, there is still little standards conformance in the GLD Linked Data cloud. The ability to make this standards compliance assessment of GLD in an objective, quantitative, automated way is an advance in the state of the art. Standards conformance was not viewed equally important by all publishers of the test datasets. However, it is hoped that this study is still informative for publishers who wish their data to conform to the requirements and best practices published by standardization organisations. It should be noted that the Greek LGD and OS UK datasets were largely created before the standardization efforts we check for and thus conformance is not expected.

Ordnance Survey Ireland has seen the utility of this approach and started to roll out this new standards-based assessment for its own datasets. This work could have a longer term impact through exposure at the Eurogeographics consortium of European national mapping agencies with consequent potential impact on EU INSPIRE data collection practices.

In future work we intend to develop additional standards conformance metrics and integrate them into our end to end quality dashboard for the OSi data publishing pipeline. The idea of standards compliance data quality metrics has a much wider scope than just geospatial data and the basic approach could be applied in any domain where standards are available.

Acknowledgement This research received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-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. International standardization organization. https://ec.europa.eu/eip/ageing/standards/ict-and-communication/data/iso-19000-series_en. Ac:15.09.2020.
2. S. Auer, J. Lehmann, and S. Hellmann. Linkedgeodata: Adding a spatial dimension to the web of data. In *International Semantic Web Conference*, pages 731–746. Springer, 2009.
3. J. Debattista, C. Lange, and S. Auer. daq, an ontology for dataset quality information. In *LDOW*, 2014.
4. J. Debattista, C. Lange, S. Auer, and D. Cortis. Evaluating the quality of the lod cloud: An empirical investigation. *Semantic Web*, (Preprint):1–43, 2018.
5. 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. Springer, 2017.
6. R. Devillers, Y. Bédard, and R. Jeansoulin. Multidimensional management of geospatial data quality information for its dynamic use within gis. *Photogrammetric Engineering & Remote Sensing*, 71(2):205–215, 2005.
7. O. Hartig and J. Zhao. Using web data provenance for quality assessment. CEUR Workshop Proceedings, 2009.
8. B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz. Requirements for data quality metrics. *Journal of Data and Information Quality (JDIQ)*, 9(2):1–32, 2018.
9. 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*, pages 68–74. ACM, 2013.
10. 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.
11. H. Moellering. A draft proposed standard for digital cartographic data, national committee for digital cartographic standards. In *American Congress on Surveying and Mapping Report*, volume 8, 1987.
12. M.-A. Mostafavi, G. Edwards, and R. Jeansoulin. An ontology-based method for quality assessment of spatial data bases. 2004.
13. M. Perry and J. Herring. Ogc geosparql-a geographic query language for rdf data. *OGC implementation standard*, 40, 2012.
14. J. Tandy, L. van den Brink, and P. Barnaghi. Spatial data on the web best practices. *W3C Working Group Note*, 2017.
15. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.
16. B. Yaman and R. Brennan. Linkeddataops:linked data operations based on quality process cycle. In *EKAW (Posters & Demonstrations)*, 2020.
17. B. Yaman, K. Thompson, and R. Brennan. Quality metrics to measure the standards conformance of geospatial linked data. In *(TBA) Proceedings of the ISWC 2020 Satellite Tracks (Posters & Demonstrations)*, 2020.
18. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016.