Linked Open Data Vocabularies for Semantically Annotated Repositories of Data Quality Measures

Franz-Benjamin Mocnik
Heidelberg University, Institute of Geography
Im Neuenheimer Feld 348, 69120 Heidelberg, Germany
mocnik@uni-heidelberg.de
https://orcid.org/0000-0002-1759-6336

Abstract

The fitness for purpose concerns many different aspects of data quality. These aspects are usually assessed independently by different data quality measures. However, for the assessment of the fitness for purpose, a holistic understanding of these aspects is needed. In this paper we discuss two Linked Open Data vocabularies for formally describing measures and their relations. These vocabularies can be used to semantically annotate repositories of data quality measures, which accordingly adhere to common standards even if being distributed on multiple servers. This allows for a better understanding of how data quality measures relate and mutually constrain. As a result, it becomes possible to improve intrinsic data quality measures by evaluating their effectivity and by combining them.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases data quality, measure, semantics, Linked Open Data (LOD), vocabulary, repository, reproducibility, OpenStreetMap (OSM)

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.50

Category Short Paper

Supplement Material `http://purl.org/data-quality`, `http://purl.org/osm-data-quality`

Funding This work was supported by the DFG project *A framework for measuring the fitness for purpose of OpenStreetMap data based on intrinsic quality indicators* (FA 1189/3-1).

1 Introduction

Data quality and fitness for purpose are major issues for many applications. Are the data of use for a certain application because they are capable of delivering the desired result? Applications each have their own requirements: certain aspects of the data might be more important than other ones for a specific application. Data quality measures quantify how usable the data are in respect to a certain aspect of the data. Among such aspects are the completeness of the data, logical consistency, positional and thematic accuracy, temporal quality, etc. [6] As in many cases no reference data are available – the reference data would then be used instead of the considered data – one aims for *intrinsic measures*, which evaluate aspects of data quality by, for the most part, only referring to the data themselves.

While often examining different aspects of data quality independently, a holistic view is needed in many practical examples. In case of vehicle routing, the completeness of the representation of the road network and the topological quality play a major role, but the geometric quality and the thematic accuracy have an impact as well. The same is true for many other applications: whether a dataset is fit for a certain purpose can only be evaluated

when assessing all concerned aspects of the data. Therefore, a repository of data quality measures should ideally address the following needs:

(N1) **Formal harmonization of measures.** Measures can often not be related because they are implemented independently. Their results are semantically incompatible and their descriptions in publications stay often unrelated. Common standards, including semantic descriptions, allow for harmonizing and combining measures.

(N2) **Situational interpretation of measures.** When assessing data quality, the results need interpretation. Measures often presume a certain context and work only in a certain setting – they mutually constrain. A repository allows for relating measures to gain a situational interpretation of their results if the relations and dependencies are formally described.

(N3) **Traceability of complex results.** Data quality measures are described and evaluated in scientific publications but their algorithms are often not properly documented. The publication in a repository under an open license and the semantic annotation allow for tracing how individual measures lead to a complex assessment of data quality.

In this article, we discuss how data quality measures and their dependencies can be described as Linked Open Data (LOD). First, we shortly summarize related work (Section 2). Subsequently, we discuss properties of data quality measures, including relations between different measures (Section 3). These properties are formalized in two vocabularies, which can be used to annotate data quality measures as LOD (Section 4). Such annotations allow for a harmonization of data quality measures and, accordingly, for examining them as a whole. The structure of a repository of data quality measures is discussed by referring to the role that the LOD vocabularies may take in this context (Section 5).

2 Related Work

Numerous data quality measures have been discussed in literature. Senaratne et. al. [13] list measures for Volunteered Geographic Information in general, and Mocnik et. al. [10] for OSM in particular. Such measures can be classified by their grounding, i. e., by the source of information used to assess data quality. A corresponding ontology has been introduced by Mocnik et. al. [10]. Data quality aspects have been discussed by Wand and Wang [14] and been published as a norm [6]. Descriptions of data quality by the properties of the data have been complemented by descriptions of how the data can be used, the fitness for purpose [2, 5]. The concepts of fitness for purpose and data quality have been related by Devillers et. al. [4]. Couclelis has discussed differences between information and knowledge in respect to imperfection [3], which emphasizes the need to relate several data quality measures. Mocnik et. al. have discussed the comparison of intrinsic and extrinsic measures [11]. The importance of traceability has, among others, been discussed by Popper [12].

3 Properties of and Relations Between Measures

In this section, we discuss the semantic foundations of a repository of data quality measures. Both intrinsic and extrinsic measures are often constrained by a context or other measures, creating the need to formally capture such constrains and relations. In the following, we discuss how to describe measures and their interrelations formally. OpenStreetMap will serve as an example while the definitions apply to data quality measures in general.

Measures assign meta information to a dataset. As an example, the *saturation principle* can measure the completeness of a road network represented in some dataset [1]. Thereby,

it is measured whether the length of the road network still increases or already stagnates – stagnation occurs when the road network is (more or less) completely represented in the data. The measure assigns to the dataset meta information about its completeness, e. g., by the increase of the road network's length. In general, measures can be conceptualized as follows:

▶ **Definition 1.** A *measure* $\mu\colon D \to R$ is a function or algorithm that assigns to each dataset $d \in D$ a result $\mu(d) \in R$. A measure is called a *data quality measure* if the result refers to the quality of the dataset.

In geographical applications, measures are of particular interest if they describe a dataset spatially. Many datasets explicitly expose a spatial dimension while others include them implicitly [9, 7]. We call a measure spatial if its result explicitly exposes a spatial dimension, aggregated by a discrete grid. The saturation principle can, e. g., be applied independently to a collection of grid cells for assessing the completeness of the road network for each of them.

▶ **Definition 2.** A measure $\mu\colon D \times G \to R$ is called *spatial* in case of G being a discrete grid that tessellates some region in $\mathbb{R}^n$ or S^n.

The saturation principle works in case of a road network for OSM [1] but it remains unclear whether it also works in other contexts, e. g., for the electrical grid. In addition, the principle only works in case that the increase of road length is in a meaningful interval. This fact can be expressed as a condition ξ to the information resulting from the measure: if the increase is outside a certain range, the measure cannot be expected to deliver meaningful information[1]. Similar concepts even apply to other measures. We accordingly define:

▶ **Definition 3.**

(a) A measure $\mu\colon D \to R$ is called to be *valid in a context* c if the result $\mu(d)$ has a meaningful interpretation in respect to c.

(b) A spatial measure $\mu\colon D \times G \to R$ is called to be *valid in an area* $G' \subset G$ if $\mu(d, g)$ has a meaningful interpretation for all $(d, g) \in D \times G'$. The measure is called to *meet condition* ξ if μ is valid in the area $G' := \{g \mid \xi(g)\} \subset G$.

Many conditions cannot be provided in general but depend on the examined place. The saturation principle, e., g., only works if volunteers contribute data about the examined area. Otherwise, the length of the road network does not increase, independent of its completeness. A second measure can be used to examine the presence of mapping activity in a particular area and, in turn, to determine in which areas the saturation principle provides meaningful information. Such relations between measures can, more formally, be described as follows:

▶ **Definition 4.** A spatial measure $\mu\colon D \times G \to R$ is said to *presume another spatial measure* $\nu\colon \tilde{D} \times G \to \tilde{R}$ *under a condition* ξ if μ is valid in the area $G' \subset G$ where ν meets ξ.

Even in before evaluating a spatial measure by computing its result for some region, one might want to know what to expect from the measure. The saturation principle might, e. g., not be able to properly distinguish between a completeness of 95 and 100 per cent. If the repository contains information about such limits of the expected results, one can decide in before whether to evaluate the saturation principle. We define:

▶ **Definition 5.** Assume R to be a totally ordered set. Then, the *minimum/maximum* of a spatial measure $\mu\colon D \times G \to R$ is defined as the minimum/maximum for both components:

$$\min \mu := \min_{d,g} \mu(d, g) \qquad \text{and} \qquad \max \mu := \max_{d,g} \mu(d, g).$$

[1] It needs to be discussed in detail and in respect to each measure what *meaningful information* refers to.

Table 1 Linked Open Data vocabulary for describing data quality measures.

Classes (selection)	**Definition**
`dq:measure`, `:dataQualityMeasure`, `:result`	Definition 1
`dq:spatialMeasure`	Definition 2
`dq:context`	Definition 3(a)
`dq:grounding`	grounding of a data quality measure [10]
Individuals (selection)	**Definition**
`dq:extrinsicPerceptionBasedGrounding`	perception-based grounding [10]
`dq:intrinsicDataBasedGrounding`, `:extrinsic...`	data-based grounding [10]
`dq:intrinsicGroundingInProcessedData`, `:ext...`	grounding in processed data [10]
`dq:intrinsicGroundingInRulesPatternsKnowledge`, `:extrinsicGroundingInRulesPatternsKnowledge...`	grounding in rules/patterns/knowledge [10]
Predicates (selection)	**Definition**
`dq:implementedBy`	who implemented the measure
`dq:documentedBy`	who documented the measure
`dq:api`	URL of the REST API
`dq:typeOfResult`	Definition 1
`dq:assesses`	assessed data quality aspect [6]
`dq:validInContext`, `:validInArea`	Definition 3
`dq:usesGrounding`	refers to the grounding-based ontology of data quality measures [10]
`dq:presumes`	Definition 4
`dq:maximumResult`, `:minimumResult`	Definition 5

These formal definitions describe how measures relate and which properties they have. In the next section, we discuss how these formal definitions can semantically be expressed by the use of Linked Open Data (LOD) vocabularies.

4 Semantic Annotation Using Linked Open Data Vocabularies

The semantic annotation of a measure allows for a better interpretation of the measure's results and for an understanding of the context of the measure. When being able to relate measures by their semantics, one can make sense of them as a whole. Here, we discuss two new LOD vocabularies for semantically annotating measures, with the aim of expressing the definitions of the preceding section and of further properties.

The first of the two vocabularies describes data quality measures and their relations (`dq`; `http://purl.org/data-quality`; Table 1). The class `measure` represents measures in general; its subclass `dataQualityMeasure`, data quality measures; and its subclass `spatialMeasure`, spatial measures. If a measure is only valid in a certain context or area, this can be described by `validInContext` and `validInArea`, respectively. The predicate `presumes` expresses that a spatial measure presumes another one.

The vocabulary can also be used to represent the source of the data quality information when evaluating a data quality measure. Data refers to the environment by relating symbols to objects and processes, i. e., the data are grounded in the environment. When data is assessed, the original grounding is compared to an additional one, which is described by the

Table 2 Linked Open Data vocabulary for describing data quality measures for OpenStreetMap.

Classes (selection)	Definition
`osmdq:spatialMeasure`	spatial measure (Definition 2) related to OSM
`osmdq:spatialDataQualityMeasure`	spatial data quality measure (Definition 2) related to OSM
`osmdq:elementType`	type of the OSM element (node, way, area, relation)
`osmdq:node`, `:way`, `:area`, `:relation`	OSM node, OSM way, OSM area, OSM relation
`osmdq:tag`, `:key`, `:value`	OSM tag, and corresponding key and value
Predicates (selection)	**Definition**
`osmdq:assessesElementType`	type of element that is assessed in particular
`osmdq:assessesTag`	tags of the elements assessed

grounding-based ontology of data quality measures [10]. The vocabulary allows for a formal representation of this ontology, by which data quality measures can be classified.

OSM-related data quality measures can be characterized by which elements they assess in the OSM dataset. This characterization is captured by a second LOD vocabulary (`osmdq`; `http://purl.org/osm-data-quality`; Table 2). In particular, `assessesElementType` describes whether a particular type of element is assessed (node, way, area, or relation). The predicate `assessesTag` refers to the tags of the elements that are assessed by the measure.

The two vocabularies described in this section can be used to annotate data quality measures and OSM-related data quality measures in particular. This allows for making sense of such measures as a whole, in particular when combining them. In the next section, we discuss the structure of a repository that contains semantically annotated measures.

5 A Repository of Quality Measures

A repository needs to expose executable algorithms as well as semantic information if it shall address the needs (N1)–(N3) of the introduction. Accordingly, different techniques have to be combined. Here, we exemplarily discuss which techniques can practically be used to build a repository[2] of data quality measures for addressing the needs (N1)–(N3).

The algorithm related to a measure is in many cases simple to understand, but its evaluation is often more complex than the central parts of the algorithm would suggest. For instance, the dataset needs to be distributed among a number of machines for efficient processing, the data need to be indexed, the history of the data might be made accessible, etc. The use of a common query language ensures the traceability of the results when the algorithms are made publicly available.

The measures in the repository should be semantically annotated by the vocabularies that have been discussed in the preceding section. Without semantics, it is hard to combine different measures and make sense of them as a whole. The use of the vocabularies, however, allows for a formal representation of the information necessary to combine different results and for taking account of mutual constraints between measures. When several measures and their results are combined, there is a need to trace how these results have been concluded. The use of formal vocabularies in combination with executable algorithms makes the evaluation of single measures and their interrelations between measures more transparent and traceable.

[2] see `https://osm-measure.geog.uni-heidelberg.de` for an exemplary implementation of these ideas

Both the algorithms and the semantic annotation can be stored in a repository using a version control system. In addition, they need to be offered on a website, where the semantic information is available as LOD. The algorithms can be run on a REST server[3] that executes the code, aggregates by the ISEA3H Discrete Global Grid System[4] [8], and caches the result. This setup ensures the effective use of the LOD vocabulary in the context of a repository.

6 Outlook

We have discussed how measures can relate and mutually constrain. In addition, we have introduced vocabularies for representing these relations and further properties. The vocabularies integrate well into a repository of data quality measures.

Intrinsic data quality measures only consume the data themselves. Despite this advantage, they can be unreliable because they cannot rely on any additional source of information. When comparing intrinsic and extrinsic measures by the use of a repository, one is able to trace the mutual dependencies of these measures. This allows for a better understanding of their relations and, as a consequence, improves the applicability of intrinsic measures.

Reasoners can take advantage of semantic annotations when relating measures. The formal representation of mutual dependencies allows thus for computationally combining data quality measures by their potentially similar (or dissimilar) results as well as by their mutual constraints, which renders synergy effects. As a result, more stable measures can be derived and data quality and fitness for purpose can be assessed more situationally.

References

1 Christopher Barron, Pascal Neis, and Alexander Zipf. A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, 18(6):877–895, 2014.

2 Nicholas R. Chrisman. The role of quality information in the long-term functioning of a geographic information system. *Cartographica*, 21(2):79–87, 1984.

3 Helen Couclelis. The certainty of uncertainty: GIS and the limits of geographic knowledge. *Transactions in GIS*, 7(2):165–175, 2003.

4 Rudolphe Devillers, Yvan Bédard, and Roberg Jeansoulin. Multidimensional management of geospatial data quality information for its dynamic use within GIS. *Photogrammetric Engineering and Remote Sensing*, 71(2):205–215, 2005.

5 Andrew U. Frank. Metamodels for data quality description. In Robert Jeansoulin and Michael F. Goodchild, editors, *Data quality in geographic information. From error to uncertainty*, page 15–29. Hermès, Paris, 1998.

6 International Organization for Standardization. ISO 19157:2013. Geographic information. Data quality, 2013.

7 Franz-Benjamin Mocnik. *A scale-invariant spatial graph model*. PhD thesis, Vienna University of Technology, 2015.

8 Franz-Benjamin Mocnik. A novel identifier scheme for the ISEA Aperture 3 Hexagon Discrete Global Grid System. *Cartography and Geographic Information Science*, 2018.

9 Franz-Benjamin Mocnik and Andrew U. Frank. Modelling spatial structures. *Proceedings of the 12th Conference on Spatial Information Theory (COSIT)*, page 44–64, 2015.

[3] e. g., using `http://github.com/giscience/measures-rest`

[4] e. g., using `http://github.com/giscience/geogrid` and `http://github.com/giscience/geogrid.js`

10 Franz-Benjamin Mocnik, Amin Mobasheri, Luisa Griesbaum, Melanie Eckle, Clemens Jacobs, and Carolin Klonner. A grounding-based ontology of data quality measures. *Journal of Spatial Information Science*, 16, 2018.

11 Franz-Benjamin Mocnik, Alexander Zipf, and Hongchao Fan. The inevitability of calibration in VGI quality assessment. *Proceedings of the 4th Workshop on Volunteered Geographic Information: Integration, Analysis, and Applications (VGI-Analytics)*, 2017.

12 Karl Popper. *The logic of scientific discovery*. Routledge, London, 1992.

13 Hansi Senaratne, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri, and Mordechai Haklay. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1):139–167, 2017.

14 Yair Wand and Richard Y. Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.