# (Linked) Data Quality Assessment: An Ontological Approach

Aparna Nayak, Bojan Božić, and Luca Longo

SFI Centre for Research Training in Machine Learning,
School of Computer Science,
Technological University Dublin, Dublin, Republic of Ireland
{aparna.nayak, bojan.bozic, luca.longo}@tudublin.ie

**Abstract.** The effective functioning of data-intensive applications usually requires that the dataset should be of high quality. The quality depends on the task they will be used for. However, it is possible to identify task-independent data quality dimensions which are solely related to data themselves and can be extracted with the help of rule mining/pattern mining. In order to assess and improve data quality, we propose an ontological approach to report data quality violated triples. Our goal is to provide data stakeholders with a set of methods and techniques to guide them in assessing and improving data quality.

**Keywords:** Data quality assessment · Data quality improvement · Linked data · Root cause analysis



## 1 Introduction

Data quality can be perceived as 'fitness for use' for a given application or a use case. Data quality is often determined by assessing if it meets the user's requirement. Assessing the data quality usually requires a large number of quality metrics to be computed rather than a single metric for a particular application. A broad range of data quality dimensions and categories of such dimensions as well as metrics for measuring these dimensions are defined in [20]. High-quality data leads to better decision-making across the application whereas poor quality data can be refined using multiple available techniques [14].

The linked data principles promote publishing data and interlinking them in a machine-readable format using Semantic Web standards. Knowledge graphs are seen as one of the essential components in envisioning the Semantic Web's idea. A knowledge graph is a graph-based knowledge base that can be considered as an RDF graph. Nodes in these graphs represent entities or literal, while edge represents the relation either between entities or between entities and literals. An RDF graph consists of a RDF triple where each triple (s,p,o) is an ordered set of the following RDF terms: a subject $s \in U \cup B$, a predicate $p \in U$, and an object

$o \in U \cup B \cup L$. An RDF term is either a Uniform Resource Identifier (URI, U), a blank node (B) or a literal (L). Nodes are usually associated with a type which a class in case of an entity or a datatype in the case of literal.

In the proposed model it is expected to design a unified approach to publish data quality along with improved quality dataset while understanding the root causes of data quality violated triples. The following is the general workflow for assessing quality and determining the root causes of violations: 1) identify a dataset, 2) uplift data in case of non-RDF data, 3) choose quality dimensions 4) detect root causes of data quality violated triples 5) apply data quality improvement techniques. Without the use of external knowledge bases, the model provides stakeholders with a set of techniques for identifying quality problems and automatic suggestions for improving the overall quality of the dataset.

In conclusion, the proposed method can be summarised as the definition of strategies to assess data quality. The remainder of this article is structured as follows. Section 2 describes the applicability of the proposed method. Existing methods and ontologies to assess data quality are discussed in section 3. Section 4 discusses the research questions that the proposed research aims to solve. Section 5 tries to answer the research questions in detail. Section 6 discusses preliminary proposed ontology along with gaps of existing frameworks. Section 7 mentions the evaluation plan. Finally, section 8 reflects the overall strengths of the framework.

## 2 Relevancy

This research proposal aims to deliver an end-to-end system that uplifts non-RDF to RDF, assesses its quality, and identifies root causes of quality violated triples to improve the quality. The effort and time required to preprocess the data will be reduced if methods to identify the root causes of data quality violated triples are provided. The proposed system is relevant for all data publishers, contributors, and consumers as the assessment of data quality will also locate the quality violated triples. Moreover, the proposed approach aims to publish the data in RDF which is a machine-understandable format. Users can decide whether or not to fix a problem in the dataset by looking at suggested quality improvement suggestions and the exact cause of the problem.

## 3 Related work

Data quality is described as a multidimensional concept [19]. Assessment of data quality involves inspecting multiple dimensions. Relevant dimensions for linked data quality have been elucidated exhaustively in literature [20], [15], [17]. Data quality assessment is followed by generating the quality report in a standard format. To manage data quality assessment reports effectively, several data quality ontologies such as Data Quality Management(DQM) [6], Reasoning Violation Ontology(RVO) [2], Data Quality Ontology (daQ) [4], Data Quality Vocabulary(DQV) [1] have been proposed. All the ontologies in the literature are helpful to represent data quality assessment reports except for RVO. RVO is a dedicated reasoning error ontology that helps to process errors.

[1] `https://www.w3.org/TR/vocab-dqv/`

There exist several frameworks such as Luzzu [3], SemQuire [10], LD Sniffer [13], TripleCheckMate [9], RDFUnit [8] and SWIQA [5] which aims to assess linked dataset. These frameworks differ in terms of scalability, generation of the quality report to publish results, total number of metrics assessed and use of external knowledge base. All the frameworks discussed here lack adoption of data quality improvement techniques as well as identifying root causes of quality violated triples. LiQuate [16], Sieve [12] and RDF improvements [7] are some frameworks that improves data quality after assessing couple of metrics.

The root cause analysis technique aims at identifying triples that have violated data quality. In the literature, there exists multiple data validation techniques such as SHACL [2] and rule-based reasoning [11]. These validation techniques help to validate the shape of RDF rather than assessment and improvement. An extension of Luzzu [18] identifies data quality violated triples based on data quality metric evaluation that helps stakeholders to prioritize and fix the errors. However, all the data quality assessment or validation approaches require either an external knowledge base or an ontology or both. In our research, we found 14 quality dimensions in the literature that are both quantifiable and have quality violated triples. We focus on the intrinsic dimension as these are explicitly relevant to assess the quality at the A-Box level.

## 4 Research questions

Research questions (RQs) deal with uplifting structured data, assessment and the enhancement steps considering the quality dimensions. The RQ related to the proposed method can be summarised as follows.

***RQ1:*** *To what extent the comma separated values be uplifted to linked data format dependent/independent of the domain vocabulary?* Multiple mapping languages exist to uplifting the dataset to RDF. When the domain vocabulary is at hand, we aim to reuse the most promising mapping languages and define our ontology to uplift data to fill the gaps in missing vocabulary. On contrary, the input data can be used to learn ontology. Ontology learning helps to assess the consistency of the dataset.

***RQ2:*** *To what extent can the quality of linked data be assessed to identify the root causes of data quality violated triples without external knowledge base?* In order to identify the root causes of the data quality violated triples we have to investigate the quality dimensions in particular focusing on the intrinsic dimension can be defined independently of the external data sources.

***RQ3:*** *To what extent the performance of the proposed model can be significantly improved by identifying domain dependent characteristics?* The proposed model is compared with baseline methods with diverse datasets to analyze the scalability, domain-dependent/independent behaviour and applicability. It also requires building a synthetic dataset to showcase data quality metrics coverage.

***RQ4:*** *To what extent all the quality metrics under consideration can be improved without degrading existing data quality?* Data quality improvement technique for each metric is identified and applied to understand the overall data

[2] `https://www.w3.org/TR/shacl`

quality. This helps to identify correlated metrics and understand the efficacy of data quality improvement techniques.

## 5 Approach

The proposed approach aims to help stakeholders to i) assess data quality problems, ii) identify triples that have violated quality, iii) understand effective strategies in solving detected problems. The entire framework can be implemented as a linear approach, summarised as follows:

- Data quality assessment : Defines actions to acquire the values of the identified data quality dimensions.
  - Definition of the quality dimensions to assess
  - Measurement of the selected quality dimensions
  - Reporting the data quality measurement
- Root cause analysis : Analyses the triples for data quality violated triples.
  - Identification of quality dimensions
  - Designing an ontology to report root cause analysis
  - Representation of the results
- Suggestions to improve triples that have violated quality: Defines a process to rectify the violations when quality is poor.
  - Building a knowledge base to correct the common errors identified in root cause analysis
  - Implementation of rule-based method to identify appropriate data quality improvement techniques.

### 5.1 Data quality assessment

Data quality assessment computation usually requires an understanding of the metrics. One of the aims of the proposed method is to assess data at the A-Box level without considering the external knowledge base. Metrics that belong to the intrinsic dimension and some of the metrics that belong to the representational dimension focus on evaluating both A-box and T-box. Table 1 gives an overview of the metrics that will be considered to assess data quality. Rule mining algorithms will be used to learn frequent patterns in the linked data. This helps to assess the metrics such as no misuse of properties, correct domain and range definition and many more. The clustering algorithm helps assess the metrics such as no misplaced classes/properties and detection of outliers. Metrics that either does not require an external knowledge base or requires a static external knowledge base are identified and considered for the assessment.

Table 2 gives an overview of metrics that requires only computation. Assessment only column refers that the specified metric is used only for assessment purposes, and the root cause of quality violation is not applied. When the raw form of RDF data is considered, metrics that are listed under direct use of RDF data can be assessed.

The root cause violations of the triples will be identified using data quality metrics focused on evaluating intrinsic dimensions. Of the 24 quality metrics identified by [20], eleven metrics neither requires an ontology nor dynamic knowledge

**Table 1.** Metrics for data quality assessment

| **Metric** | **A-Box level** | **Root causes analysis** | **External KB** |
|---|---|---|---|
| No syntax errors | N | Y | N |
| Accurate values | Y | Y | Y |
| No malformed datatypes | Y | Y | N |
| No outliers | Y | Y | N |
| No misuse of properties | N | Y | N |
| No misplaced classes or properties | N | Y | N |
| Correct domain and range definition | N | Y | N |
| Keeping short URI | N | Y | N |
| No use of prolix RDF features | N | Y | Y |

**Table 2.** Metrics that requires only computation

| **Metric** | **Assessment only** | **Direct use of RDF data** |
|---|---|---|
| Intensional conciseness | ✓ | ✗ |
| Extensional conciseness | ✓ | ✗ |
| Reuse of existing terms | ✓ | ✗ |
| Reuse of existing vocabulary | ✓ | ✗ |
| Detecting the use of blank nodes | ✓ | ✗ |
| No misuse of owl:DatatypeProperty/ owl:ObjectProperty | ✓ | ✓ |
| Members of owl:DerecatedClass/ owl:DeprecatedProperty | ✓ | ✓ |
| No use of entities as members of disjoint classes | ✓ | ✓ |

base. The remaining dimensions are not considered mainly due to the factors 1) Input data requires to be stored on client side. 2) Need of external data sources. For example, quality dimensions in the accessibility class (availability, security, performance, interlinking, licensing) are not taken into account because this dimension focuses on and evaluates metrics when data is stored on the server. Contextual dimension focus on the context of the task at hand is also not considered for evaluation. The supported dimensions are intrinsic and representational. Hence the strengths of the proposed model are in enforcing intrinsic and representational data quality dimensions.

### 5.2 Root cause analysis

Root cause analysis helps to identify the reason for the violations. To develop an effective correctable measure to correct and prevent such adverse outcomes in the future, it is critical to first understand the cause. The evaluation of the root cause is followed by the formulation of recommendations. For example, datatype mismatch usually is a quality problem. When the system detects it, a

recommendation can be given for the desired datatype along with subject and object information.

### 5.3 Data quality improvement

This stage aids in the improvement of data quality. The identification of quality violated triples is the result of root cause analyses. Depending on the data quality assessment value for each assessed metric, quality improvement techniques can be applied. Data quality improvement method helps to remove outliers, correct malformed datatype literals, infer missing triples and many more. The proposed method also aims to notify the suggestion in the form of add/delete/modify the erroneous triple, which could lead to an overall data quality improvement.

## 6 Discussion

Existing data quality assessment approaches compute the quality metric, and the assessment result is given as a numeric value. Prevailing assessment methods require an external knowledge base in the form of vocabulary/ontology or a dictionary. Linked data is verified against a given external knowledge base to generate an assessment report. The proposed approach makes use of ontology learning, rule mining, and static dictionaries to assess the quality of the data. Ontology learning and rule mining algorithm help to understand the frequent patterns that can be used to assess metrics such as consistency, syntactic validity and outliers. Dataset users should have the privilege to know the root causes of the data quality violated triples. Root cause analysis helps to identify the location of the problem, such as subject, predicate, or object, along with the type of the quality problem. Quality problem type refers to a metric that has failed on a particular triple. The frameworks presented in section 3 does not identify the quality violated triples rather only helps in assessment of metrics. An ontology will be proposed to report data quality violated triples as well as data quality assessment. A preliminary proposed ontology is as shown in figure 1. Moreover, suggestions over quality violated triples will be given that helps to improve data quality. The suggestions in terms of add/modify/delete help to improve data quality.

## 7 Evaluation Plan

The outcome of this research project will be a tool for supporting data scientists to uplift datasets to RDF format, assess data quality and analysing root causes of data quality violated triples to improve data quality. Multiple features must be evaluated to understand the success of the proposed approach. However, the main contribution of the work focuses on assessing data quality and analysing the root causes of data quality violated triples to improve the quality. The model's performance is also evaluated considering scalability and time taken to identify the root causes of violations of the knowledge graph. Moreover, metrics coverage can be evaluated by considering a) synthetic dataset, b) the results obtained by other tools to compare and c) results which has to be verified manually.

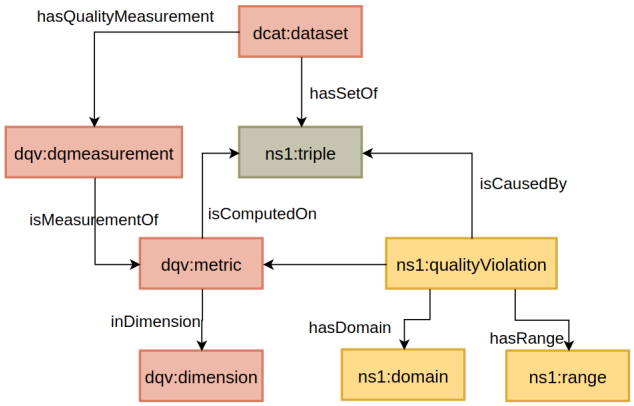


**Fig. 1.** Ontology for quality violated triples

The synthetic dataset is carefully designed to verify the coverage of all the implemented metrics. Apart from the synthetic dataset, the model is evaluated by considering multiple datasets to verify applicability. The correctness of the model is verified by comparing the proposed model with existing tools and with the help of manual evaluation. Thus, our hypothesis that answers all the research questions defined in section 4 is stated as follows.

**Hypothesis:** The time required to improve data quality after identifying the root causes of violations is less than applying random data quality improvement technique directly on the raw data.

## 8 Reflections

To the best of our knowledge, quality assessment, root cause analysis and quality improvement are rarely managed simultaneously in linked data. Therefore, our goal is to fill this gap by proposing a framework that helps data providers and consumers to assess and improve data quality. Moreover, the features offered by this framework will be integrated into the ML framework to understand data quality and test the applicability of our proposal.

**Acknowledgements** This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

1. Albertoni, R., Isaac, A.: Introducing the data quality vocabulary (DQV). Semantic Web **12**(1), 81–97 (2021)

2. Bozic, B., Brennan, R., Feeney, K., Mendel-Gleason, G.: Describing reasoning results with rvo, the reasoning violations ontology. In: MEPDaW and LDQ co-located with ESWC. CEUR Workshop Proceedings, vol. 1585, pp. 62–69. CEUR-WS.org (2016)
3. Debattista, J., Auer, S., Lange, C.: Luzzu - A methodology and framework for linked data quality assessment. ACM J. Data Inf. Qual. **8**(1), 4:1–4:32 (2016)
4. Debattista, J., Lange, C., Auer, S.: daq, an ontology for dataset quality information. In: Proceedings of the Workshop on Linked Data on the Web co-located with WWW. CEUR Workshop Proceedings, vol. 1184. CEUR-WS.org (2014)
5. Furber, C., Hepp, M.: Swiqa - a semantic web information quality assessment framework. In: 19th European Conference on Information Systems, ECIS. p. 76 (2011)
6. Fürber, C., Hepp, M.: Towards a vocabulary for data quality management in semantic web architectures. In: Proceedings of the 2011 EDBT/ICDT Workshop on Linked Web Data Management. pp. 1–8. ACM (2011)
7. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: Proceedings of the WWW 2010 Workshop on Linked Data on the Web. CEUR Workshop Proceedings, vol. 628. CEUR-WS.org (2010)
8. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.: Test-driven evaluation of linked data quality. In: 23rd International World Wide Web Conference, WWW. pp. 747–758. ACM (2014)
9. Kontokostas, D., Zaveri, A., Auer, S., Lehmann, J.: Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. In: Knowledge Engineering and the Semantic Web. Communications in Computer and Information Science, vol. 394, pp. 265–272. Springer (2013)
10. Langer, A., Siegert, V., Göpfert, C., Gaedke, M.: Semquire - assessing the data quality of linked open data sources based on DQV. In: Current Trends in Web Engineering - ICWE. Lecture Notes in Computer Science, vol. 11153, pp. 163–175. Springer (2018)
11. Meester, B.D., Heyvaert, P., Arndt, D., Dimou, A., Verborgh, R.: RDF graph validation using rule-based reasoning. Semantic Web **12**(1), 117–142 (2021)
12. Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops. pp. 116–123. ACM (2012)
13. Mihindukulasooriya, N., García-Castro, R., Gómez-Pérez, A.: LD sniffer: A quality assessment tool for measuring the accessibility of linked data. In: Knowledge Engineering and Knowledge Management - EKAW. Lecture Notes in Computer Science, vol. 10180, pp. 149–152. Springer (2016)
14. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web **8**(3), 489–508 (2017)
15. Radulovic, F., Mihindukulasooriya, N., García-Castro, R., Gómez-Pérez, A.: A comprehensive quality model for linked data. Semantic Web **9**(1), 3–24 (2018)
16. Ruckhaus, E., Vidal, M., Castillo, S., Burguillos, O., Baldizan, O.: Analyzing linked data quality with liquate. In: The Semantic Web: ESWC. Lecture Notes in Computer Science, vol. 8798, pp. 488–493. Springer (2014)
17. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Communications of the ACM **40**(5), 103–110 (1997)
18. Vaidyambath, R., Debattista, J., Srivatsa, N., Brennan, R.: An intelligent linked data quality dashboard. In: Proceedings for the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science. CEUR Workshop Proceedings, vol. 2563, pp. 341–352. CEUR-WS.org (2019)
19. Wand, Y., Wang, R.Y.: Anchoring data quality dimensions in ontological foundations. Commun. ACM **39**(11), 86–95 (1996)
20. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web **7**(1), 63–93 (2016)