2022-2

Linked Data Quality Assessment: A Survey

Aparna Nayak
Technological University Dublin, d19125691@mytudublin.ie

Bojan Bozic
Technological University Dublin, bojan.bozic@tudublin.ie

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

Recommended Citation

Nayak A., Božić B., Longo L. (2022) Linked Data Quality Assessment: A Survey. In: Xu C., Xia Y., Zhang Y., Zhang LJ. (eds) Web Services – ICWS 2021. ICWS 2021. Lecture Notes in Computer Science, vol 12994. Springer, Cham. DOI: 10.1007/978-3-030-96140-4_5

Funder: Science Foundation Ireland

Linked Data Quality Assessment: A Survey

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. Data is of high quality if it is fit for its intended use in operations, decision-making, and planning. There is a colossal amount of linked data available on the web. However, it is difficult to understand how well the linked data fits into the modeling tasks due to the defects present in the data. Faults emerged in the linked data, spreading far and wide, affecting all the services designed for it. Addressing linked data quality deficiencies requires identifying quality problems, quality assessment, and the refinement of data to improve its quality. This study aims to identify existing end-to-end frameworks for quality assessment and improvement of data quality. One important finding is that most of the work deals with only one aspect rather than a combined approach. Another finding is that most of the framework aims at solving problems related to DBpedia. Therefore, a standard scalable system is required that integrates the identification of quality issues, the evaluation, and the improvement of the linked data quality. This survey contributes to understanding the state of the art of data quality evaluation and data quality improvement. A solution based on ontology is also proposed to build an end-to-end system that analyzes quality violations' root causes.

Keywords: Data quality · Knowledge graphs · Linked data · Quality assessment · Quality improvement

1 Introduction

Data quality is often defined as "fitness for use" which signifies the term data quality is relative [6]. Thus, data with certain quality considered good for one use may not possess sufficient quality for another use. A massive amount of data is available in the public domain in the form of text, tables and linked data. However, most of these data are often incorrect, incomplete or ambiguous.

The term "Knowledge graph" refers to a set of best practices for publishing and connecting linked data on the web following Semantic Web principles. The main goal of Semantic Web is data interoperability, which allows data to be read and understandable both by humans and machine. A large number of published datasets (or sources) that follow linked data principles is currently

C. Xu et al. (Eds.): ICWS 2021, LNCS 12994, pp. 63–76, 2022.
https://doi.org/10.1007/978-3-030-96140-4_5

available and this number grows rapidly. Knowledge graph have a wide range of applications, including recommendation systems [23], semantic search based on entities and relationships, natural language disambiguation, deep reasoning, machine reading, entity consolidation for big data, and text analysis [8]. The semantic richness of knowledge graph can benefit explainable artificial intelligence, an emerging field of machine learning. However, large knowledge graphs such as DBpedia[1] and Wikidata[2] still suffer from different quality problems [21].

Data quality is being one of the major concern this paper aims to achieve the following objectives:

O1: Identification and survey existing data quality assessment/improvement framework/tools and data quality metrics.
O2: Investigate frameworks and tools that enable the quality assessment of data at A-box level.

Our contributions in this paper include identifying various ways to assess and improve problems associated with data quality. A preliminary framework that enables end-to-end systems for data assessment and improvement is also discussed. The rest of this paper is organized as follows. Section 2 discusses the literature present in data quality assessment and improvement. Section 3 provides an outlook for further research. Finally, Sect. 4 concludes the work.

2 Methods for Data Quality Assessment and Improvement

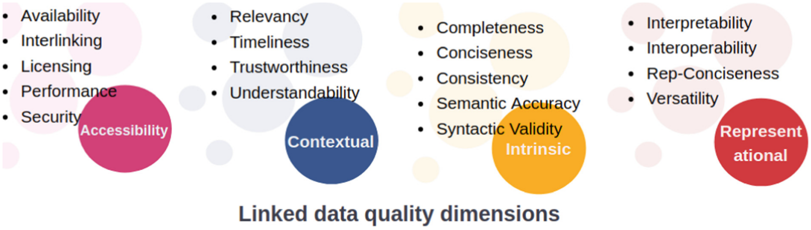

Fig. 1. Linked data quality dimensions

The objective of the data quality assessment activity is to analyze the relevance of a dataset to its consumers and to help publish better quality data. Analysts working with linked data must assess quality at various levels such as instance, schema and property. Data quality is a multidimensional concept. Various studies have classified data quality metrics into four dimensions intrinsic, accessible,

[1] https://wiki.dbpedia.org/.
[2] https://www.wikidata.org/wiki/Wikidata:Main_Page.

representational, and contextual [44] as shown in Fig. 1. Data quality metrics that belong to intrinsic dimensions focus on whether the information correctly and completely represents the real world and whether the information is logically consistent in itself. The accessible dimension encompasses the aspects of data access, authentication, and retrieval in order to retrieve all or a portion of the data required for a particular use case. Representational dimensions capture information about the data's design. Contextual dimensions are those that are highly context-dependent, such as relevance, trustworthiness, comprehendibility, and timeliness. Zaveri et al. [51] discusses a comprehensive survey that includes multiple metrics for evaluating each dimension. It examines 68 quality metrics for linked data and provides a detailed explanation of how each metric is calculated. On the other hand, data quality metrics are divided into baseline and derived by incorporating the metrics defined in Zaveri et al. [51] and ISO 25012[3].

The most frequently encountered issues such as missing data, missing entity relationships, and erroneous data values have a direct impact on data quality. Additionally, converting data from one format to linked data may degrade data quality due to various problems such as errors introduced at the source, parsing values, interpreting, and converting units [48]. Integration of data from multiple sources does not always result data quality improvement; rather, if the sources contain contradictory information, the quality may deteriorate [31]. Regardless of the total number of integrated data sources, quality issues persist at the schema and instance levels [39]. In the following subsections various methods to assess and improve the data quality are discussed.

2.1 Ontologies Based on Data Quality

This section discusses the ontologies that have been modeled in order to identify data quality issues and generate a report on data quality. Data Quality Management (DQM) vocabulary, conceptualizes data quality requirements by focusing on the intrinsic quality of the data [20]. This ontology aids in the description of data quality assessment results and data cleaning rules in a Semantic Web architecture. Data Cleaning Ontology (DCO), one more ontology that represents the data cleaning process [4]. DCO is an advanced version of DQM that assists domain experts with data cleaning. However, these ontologies do not directly help to assess data quality. Data Quality Vocabulary (daQ), helps to represent results of data quality assessment in machine-readable format [15]. This ontology defines a core vocabulary that enables the uniform definition of specific data quality metrics, which data publishers can include in their metadata. W3 has published Data Quality Vocabulary (DQV) [3] to represent data quality assessment in Semantic Web format[4]. Data publisher or consumer can use this vocabulary to represent their data quality assessment report. Fuzzy Quality Data Vocabulary (FQV) extends DQV to represent the fuzzy concepts. Fuzzy

[3] https://iso25000.com/index.php/en/iso-25000-standards/iso-25012.

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

ontology assesses the data quality using fuzzy inference systems based on user-defined fuzzy rules [5]. The aforementioned ontologies do not help to assess the quality of the data, rather publish quality reports in a machine-readable manner. Data quality is assessed at various levels such as perception, data, processed and, rules. This helps to differentiate validation report of the data quality from the different point of view [35]. Reasoning Violations Ontology (RVO) is an ontology used to validate the triples and reason out the violations if any [9].

Table 1. Ontologies based on data quality

Ontology	Richness	Dataset	Evaluation method
DQM	64	Synthetic data	SPARQL queries
FQV	13	Peel, DBLP (L3S), DBPedia, EIONET	Compared proposed method with Sieve [31]
DQV	10	–	–
RVO	14	Dacura schema manager	Integrated RVO in multiple ontology to identify errors
Grounding based ontology	4	OpenStreetMap data	Domain experts and external dataset such as Google maps

Table 1 compares various ontologies that focus on data quality. Richness of the ontology is computed based on total number of classes in the ontology. Dataset column indicates the dataset used to validate the ontology and evaluation method depicts how the ontology is evaluated.

2.2 Data Quality Assessment

Existing data quality assessment tools differ on various characteristics such as the number of metrics to assess quality, approaches to process data, type of data used to evaluate, user flexibility to choose metric & corresponding weight and assessment report. Luzzu [16] is a stream-oriented data quality assessment framework that requires domain experts to explicitly mention the metrics using either a programming language or declarative statements. Semquire [27], a software tool for linked data quality assessment, implements the quality metrics mentioned in [51] based on user/application requirement. Despite that the framework provides a cyclical process to define quality metrics and evaluate a dataset, it does not address the defects' root causes. A number of other data quality assessment tools focus on either a specific data set or a specific metric mentioned in Table 2.

A plethora of research focus primarily on various levels of linked data. These levels include schema, instance and properties. One of the sources for linked data is (Semi-) structured data. The mapping languages used to convert semi-structured data into linked format impacts the data quality due to incorrect usage of schema in the mapping definitions, mistakes in the original data source [18,41]. Various quality deficiencies at schema and instance level and resolution strategy have been listed in [7]. One more method to assess the data quality is to use of external sources. All RDF triples are compared with external sources to identify inaccurate information present in the knowledge graph [29]. The correctness of RDF triples can be measured by a confidence score that is generated based on the reliability score of each triple. Other works analyze the quality of DBpedia available in different language editions such as Spanish [34], and Arabic [26] by comparing different versions of DBpedia or comparing various language editions. The results of the research can be used by the DBpedia community (publisher) to eliminate the errors in its further editions.

Table 2. Data quality assessment tools

Tool	Data source	Goal	Evaluation method
Sieve [31]	DBpedia	Identify the quality and integrate data from multiple sources to get improved data set	Not mentioned
TripleCheck Mate [25]	DBpedia	Assess and improve DBpedia data	Crowdsourcing
Databugger [24]	DBpedia	Test driven data debugging framework based on SPARQL queries	Used same queries against 5 different data set to show case the tool re-usability
Luzzu [16]	Real world dataset	To identify the quality of the linked dataset	Evaluated the tool for scalability
LD Sniffer [32]	DBpedia	To analyze the availability of the given URI and assess the retrieved data using LDQM	Not mentioned
Semquire [27]	Real world dataset	To identify the quality of given linked dataset	Compared various publicly available KG

Data quality assessment tools such as ABSTAT [36], Loupe [33], DistQualityAssessment [42], Roomba [2] focus on understanding statistical information which include number of triples, and implicit vocabulary information. The information derived from these tools help the user get insight into the dataset that includes detecting outliers in the vocabulary usage, most frequent patterns in linked data, and thus interpreting data quality. Data quality framework KBQ [40,43] help in evolution analysis of linked data by comparing all the triples of two consecutive releases of the dataset. Other related work [18,47] assess the data quality; however, it fails to mention any technique to improve the identified data quality problem. In addition, some methods involve manual work to evaluate each fact for correctness [1,50].

2.3 Data Quality Improvement

Data quality improvement can make use of either external data or the knowledge graph itself. The presence of illegal values, typographical errors and missing information may lead to poor data quality [39]. Knowledge graph refinement [37], and reasoning is a technique used to refine existing data and add missing hidden information. Reasoning methods are based on logical rules, neural networks, and continuous vector space that can be used to infer missing knowledge by refining the given knowledge graph [12]. Sieve [31] compares two different data sources and chooses the accurate value based on time-closeness and preference. Sieve is a data fusion approach that enriches the DBpedia data by comparing English and Portuguese wikipedia editions. Conceptnet, one of the publicly available knowledge graph is improved by adding more triples that are extracted from news and tweets [49]. Though the accuracy of the relation extraction model is low, authors haven't mentioned anything about the quality of the added information.

Quality of the data can be improved by using supervised methods [10,11,30], or unsupervised methods [17,38,45]. Data quality can improve by resolving range violation [28], outlier detection [17], tensor factorization [45] and link prediction [10,11,30]. Statistical relational learning plays a significant role in knowledge graph as it also studies the graph structure of knowledge graph [22].

2.4 Root Cause Identification

Data contains errors that need to be identified and resolved. Identification of the location of the data quality problem is possible by root cause analysis. Various datasets published by the government have been evaluated for quality defects such as missing data, format issues, logical duplication and many more. Some of the common mistakes that is often generated by publisher side that affect quality problems and suggestions to improve the same are listed by [13]. However, they have not mentioned the fine-grained level of quality analysis. In another related study, [46] root causes of data quality violations are identified with the help of cause and effect diagram. The experiment comprises of quantitative metrics to analyze the data quality. The research shows that analysis of errors is helpful both for novice and domain experts. However, there is a lack of research that suggests an improvement over identified quality problems. Authors in [14] have validated RDF dataset using constraints that give detailed root cause explanations for all the errors present in the given RDF triple. The framework is validated against SHACL[5] and covers most of the constraints SHACL can validate.

[5] https://www.w3.org/TR/shacl/.

3 Recommendations and Future Work

The findings from this survey are (i) lack of end-to-end systems that assess and refine data quality of knowledge graphs, (ii) lack of evaluation methods. The end-to-end system requires a complete understanding of data quality metrics assessment, root causes of violations, and suggestions to refine the triples that do not obey the data quality. The proposed data quality refinement lifecycle, as shown in Fig. 2 includes the following:

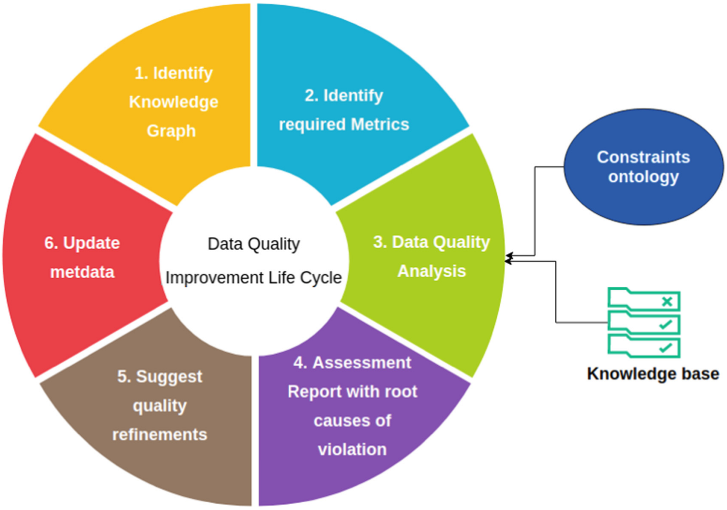
1. Identify Knowledge Graph
2. Identify required Metrics
3. Data Quality Analysis
4. Assessment Report with root causes of violation
5. Suggest quality refinements
6. Update metdata
Data Quality Improvement Life Cycle
Constraints ontology
Knowledge base

Fig. 2. Stages of ontology based data quality improvement

A ontology Data Quality Assessment and Improvement (DQAI) is proposed and has to be modeled by considering all the stages of lifecycle shown in Fig. 2. Figure 3 describes initial version of the proposed ontology which describes the dataset along with data quality assessment, root causes of violations and improvement classes. Each dataset is assessed using multiple metrics (M1 ...) that belongs to accessible, intrinsic, contextual and representational dimension. Metric is associated with quality violation which describes type of violation associated with the triple. Each type of quality violation is associated with improvement technique.

1. **Identify the Knowledge graph.** The first step is to select a knowledge graph, whose quality has to be analysed. Knowledge graphs follow some structure to store data which is referred as domain ontology. In case of absence of domain ontology, it can be learned from the knowledge graph. DQI stores the knowledge graph under graph class.

   For example, consider Microsoft Academic Knowledge Graph (MKAG) [19]. MKAG ontology has eight classes that are Paper, Affiliation, Field of study etc.

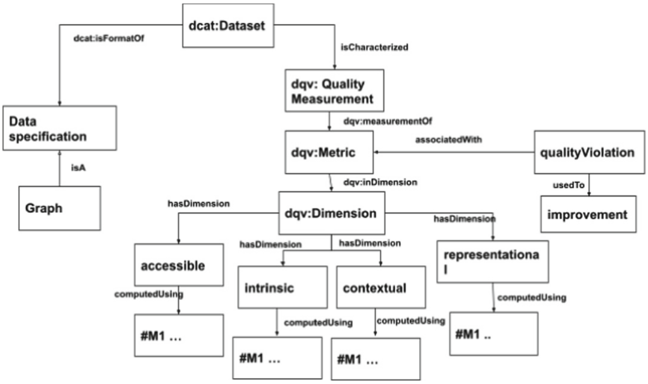

Fig. 3. Proposed ontology for data quality assessment and improvement

2. **Identify required metrics.** The quality assessment requirement varies according to the dataset. For instance, if the considered knowledge graph is an RDF dump, users do not need to concern about the SPARQL endpoint and server being accessible. Additionally, the user must have the option of selecting the required metric. This enables flexibility of the system. This step invokes required metrics under each dimension in DQAI ontology.
   From the MKAG example, let us consider a user who wants to assess two quality metrics on MKAG that are syntactically accurate values and no malformed datatype literals.
3. **Data quality analysis.** During this stage, the knowledge graph is assessed against the quality metrics identified and defined in the previous lifecycle stage. The metric implementation can be assisted with domain ontology and a knowledge base. Domain ontology and a knowledge base helps evaluate multiple metrics of intrinsic dimension. These act as rules to evaluate data quality. With the help of reasoning engine, all quality violated triples are stated and stored for further analysis by mapping them to axioms in the data quality ontology. Metrics that are of interest are computed and stored in the ontoloty DQAI.
   From the MKAG example, domain ontology of MKAG can be considered. A reasoner based on description logic will infer problematic triples that does not obey the rules mentioned in the knowledge base and ontology. Consider a class 'author' that has properties orcidId and paperCount. PaperCount has a datatype integer that means any value other than integer for this attribute is quality violation as per the definition of the metric 'no malformed datatype'. 'Syntactically accurate values' is computed either with the help of clustering/syntactic rules. Clustering on orcidId would cluster similar id into one/multiple clusters leaving out the wrongly mapped orcidId. Similarly all other metrics that are of interest to the user are computed on all the properties in the knowledge graph and quality values are computed based on the definitions given to each metric in [51].

4. **Assessment report with root causes of violations.** Data quality assessment report describes the data quality of the knowledge graph for all the metrics chosen in step 2. This report can make use of the data quality vocabulary(DQV) approved by W3 consortium to report data quality assessment score. It will also elucidate triples violating quality constraints along with the precise reason for the violation. While evaluating the data quality metric, it is possible to identify triples that violate the quality metric. These violations are stored in the ontology for subsequent analysis.
5. **Suggest quality refinements/improvements.** Resolving the violations requires refinement process by the framework. Improvement of data quality requires to add/modify/remove the triple violating quality constraint. These automatic suggestions help the user to make a decision. Improvement techniques can be applied on quality violated triples that are stored in DQAI ontology.

   From the MKAG example, for quality violated triples a suggestion should be given to the user. It helps the user to take a decision that helps to improve the quality of the available data.
6. **Update metadata.** In this stage, the knowledge graph is appended with a quality assessment report along with all triples violating quality constraints and suggestions. It helps the user to understand their knowledge graph quality and root causes of triples violating quality constraints before using knowledge graph.

   MAKG example of sample input and expected output for step 4 is as shown in Listing 1.1. Assume that there is wrongly mapped datatype for paperCount, syntactically invalid value for orcidId. Quality violated triples are identified with the help of knowledge base that validates the triples with the given ontology and facts stated by domain experts. The output must identify all triples that do not obey the constraints mentioned in the knowledge base. Expected output shows ill-typed literal and the data quality associated dataset. The further step involves refinement that can make suggestions to add/modify/remove a particular triple.

Listing 1.1. Expected input and output of the proposed method

```
Input:
mk: https://mkag.org/class.
mag: https://makg.org/property.
foaf: http://xmlns.com/foaf/0.1/.
dbo: http://dbpedia.org/ontology.
: http://dataqualityviolation.com/violations.

mk:author dbo:orcidId ‘‘1234−2345−1234−43’’;
          mag:paperCount 12.3;

Expected output:
:violation :type :datatypemismatch ;
           :triple mk:author ;
           :value mag:paperCount ;
           :datatype xsd:decimal ;
           :expectedDT xsd:integer .

:myDataset a dcat:Dataset ;
           dcterms:title MAKG ;
           dqv:hasQualityMeasurement :somemeasurement .

:somemeasurement a dqv:QualityMeasurement ;
                 dqv:computedOn :myDataset ;
                 dqv:isMeasurementOf :inverseFuncmismatch ;
                 dqv:value ‘‘12”^^xsd:int .
```

Most of the literature have evaluated their model by considering various knowledge graphs rather than comparing their model with similar other models. One of the most significant issues is a diverse format of the quality assessment report because of which it is highly challenging to compare quality assessment results of the models. W3 has defined the data quality vocabulary to describe the results of data quality assessment. Researchers can make use of this vocabulary while publishing data quality assessment results. Another problem is the number of metrics used to assess the model. A solution for such problem requires benchmarking standard collection of metrics as well as an evaluation method with the help of domain experts.

An assessment framework that works on any knowledge graph is a requirement. However, to the best of our knowledge the knowledge graph used for most of the existing research is DBpedia. Researchers have tried to solve quality issues related to DBpedia rather than giving a generic approach. One can use their proposed model on multiple RDF dumps to understand whether the model can identify problems associated with RDF data.

4 Conclusion

This paper presents a survey on knowledge graph assessment and improvement approaches. It can be seen that a larger body of work exists on data quality assessment techniques ranging from an assessment based on a single metric to multiple metrics with different goals. The survey has revealed that there are, at the moment, rarely any approaches which simultaneously assess and refine the knowledge graphs. Most of the literature considers scalability performance as an evaluation method rather than defining the model's accuracy by considering test dataset.

This survey's future work involves modeling an ontology to capture all the data quality violations. It also includes building a knowledge base that can logically reason out violations to locate the quality violated triples. This helps data publishers and consumers understand their data quality along with quality violated triples, if any. A gold standard dataset has to be prepared to all possible violations which can be used to for evaluation purposes.

Acknowledgements. This work was funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183).

References

1. Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Flöck, F., Lehmann, J.: Detecting linked data quality issues via crowdsourcing: a DBpedia study, vol. 9, pp. 303–335. IOS Press (2018)
2. Assaf, A., Troncy, R., Senart, A.: Roomba: an extensible framework to validate and build dataset profiles. In: Gandon, F., Guéret, C., Villata, S., Breslin, J., Faron-Zucker, C., Zimmermann, A. (eds.) ESWC 2015 (LNAI and LNB). LNCS, vol. 9341, pp. 325–339. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25639-9_46
3. Albertoni, R., et al.: Data quality vocabulary (DQV). W3C interest group note. World Wide Web Consortium (W3C) (2015)
4. Almeida, R., Maio, P., Oliveira, P., Barroso, J.: Ontology based rewriting data cleaning operations, vol. 20–22-July-2016, pp. 85–88. Association for Computing Machinery (2016)
5. Arruda, N., et al.: A fuzzy approach for data quality assessment of linked datasets, vol. 1, pp. 387–394. SciTePress (2019)
6. Ballou, D.P., Tayi, G.K.: Enhancing data quality in data warehouse environments. Commun. ACM **42**(1), 73–78 (1999)
7. Behkamal, B., Kahani, M., Bagheri, E.: Quality metrics for linked open data. In: Chen, Q., Hameurlain, A., Toumani, F., Wagner, R., Decker, H. (eds.) DEXA 2015. LNCS (LNAI and LNB), vol. 9261, pp. 144–152. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22849-5_11
8. Bonatti, P.A., Decker, S., Polleres, A., Presutti, V.: Knowledge graphs: new directions for knowledge representation on the semantic web (Dagstuhl seminar 18371). Dagstuhl Rep. **8**(9), 29–111 (2019)
9. Bozic, B., Brennan, R., Feeney, K., Mendel-Gleason, G.: Describing reasoning results with RVO, the reasoning violations ontology. In: MEPDaW/LDQ@ ESWC. pp. 62–69 (2016)
10. Caminhas, D., Cones, D., Hervieux, N., Barbosa, D.: Detecting and correcting typing errors in DBpedia, vol. 2512. CEUR-WS (2019)
11. Chen, J., Chen, X., Horrocks, I., Jiménez-Ruiz, E., Myklebust, E.B.: Correcting knowledge base assertions. ArXiv abs/2001.06917 (2020)
12. Chen, X., Jia, S., Xiang, Y.: A review: knowledge reasoning over knowledge graph. Expert Syst. Appl. **141**, 112948 (2020)
13. Csáki, C.: Towards open data quality improvements based on root cause analysis of quality issues. In: Parycek, P., et al. (eds.) EGOV 2018. LNCS (LNAI and LNB), vol. 11020, pp. 208–220. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98690-6_18
14. De Meester, B., Heyvaert, P., Arndt, D., Dimou, A., Verborgh, R.: RDF graph validation using rule-based reasoning. Semant. Web J. **12**(1), 117–142 (2020)
15. Debattista, J., Lange, C., Auer, S.: daQ: an ontology for dataset quality information. In: Central Europe Workshop Proceedings, vol. 1184. CEUR-WS (2014)
16. Debattista, J., Auer, S., Lange, C.: Luzzu-a methodology and framework for linked data quality assessment. J. Data Inf. Qual. **8**(1), 1–32 (2016)

17. Debattista, J., Lange, C., Auer, S.: A preliminary investigation towards improving linked data quality using distance-based outlier detection. In: Li, Y.-F., et al. (eds.) JIST 2016. LNCS, vol. 10055, pp. 116–124. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50112_39
18. Dimou, A., et al.: Assessing and refining mappings to RDF to improve dataset quality. In: Arenas, M., et al. (eds.) ISWC 2015. LNCS (LNAI and LNB), vol. 9367, pp. 133–149. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25010-6_8
19. Färber, M.: The Microsoft academic knowledge graph: a linked data source with 8 billion triples of scholarly data. In: Ghidini, C., et al. (eds.) ISWC 2019. LNCS, vol. 11779, pp. 113–129. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30796-7_8
20. Fürber, C., Hepp, M.: Towards a vocabulary for data quality management in semantic web architectures. In: Proceedings of the 1st International Workshop on Linked Web Data Management, LWDM 2011, pp. 1–8. Association for Computing Machinery, New York (2011)
21. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semanti. Web **9**(1), 77–129 (2018)
22. Hadhiatma, A.: Improving data quality in the linked open data: a survey, vol. 978, p. 012026. Institute of Physics Publishing (2018)
23. Heitmann, B., Hayes, C.: Using linked data to build open, collaborative recommender systems. In: AAAI Spring Symposium: Linked Data Meets Artificial Intelligence, vol. SS-10-07, pp. 76–81 (2010)
24. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R.: Databugger: a test-driven framework for debugging the web of data, pp. 115–118. Association for Computing Machinery, Inc. (2014)
25. Kontokostas, D., Zaveri, A., Auer, S., Lehmann, J.: TripleCheckMate: a tool for crowdsourcing the quality assessment of linked data. In: Klinov, P., Mouromtsev, D. (eds.) KESW 2013. CCIS, vol. 394, pp. 265–272. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41360-5_22
26. Lakshen, G., Janev, V., Vraneš, S.: Challenges in quality assessment of Arabic DBpedia. Association for Computing Machinery (2018)
27. Langer, A., Siegert, V., Göpfert, C., Gaedke, M.: SemQuire - assessing the data quality of linked open data sources based on DQV. In: Pautasso, C., Sánchez-Figueroa, F., Systä, K., Murillo Rodríguez, J.M. (eds.) ICWE 2018. LNCS (LNAI and LNB), vol. 11153, pp. 163–175. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-03056-8_14
28. Lertvittayakumjorn, P., Kertkeidkachorn, N., Ichise, R.: Resolving range violations in DBpedia. In: Wang, Z., et al. (eds.) JIST 2017. LNCS (LNAI and LNB), pp. 121–137. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-70682-5_8
29. Liu, S., d'Aquin, M., Motta, E.: Measuring accuracy of triples in knowledge graphs. In: Gracia, J., Bond, F., McCrae, J.P., Buitelaar, P., Chiarcos, C., Hellmann, S. (eds.) LDK 2017. LNCS (LNAI and LNB), vol. 10318, pp. 343–357. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59888-8_29
30. Melo, A., Paulheim, H.: Automatic detection of relation assertion errors and induction of relation constraints. Sprachwissenschaft, pp. 1–30 (2020)
31. Mendes, P., Mühleisen, H., Bizer, C.: Sieve: linked data quality assessment and fusion. In: ACM International Conference Proceeding Series, pp. 116–123 (2012)

32. Mihindukulasooriya, N., García-Castro, R., Gómez-Pérez, A.: LD sniffer: a quality assessment tool for measuring the accessibility of linked data. In: Ciancarini, P., et al. (eds.) EKAW 2016. LNCS (LNAI and LNB), vol. 10180, pp. 149–152. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58694-6_20
33. Mihindukulasooriya, N., Poveda-VillaÍon, M., García-Castro, R., Gómez-Pérez, A.: Loupe-an online tool for inspecting datasets in the linked data cloud, vol. 1486. CEUR-WS (2015)
34. Mihindukulasooriya, N., Rico, M., García-Castro, R., Gómez-Pérez, A.: An analysis of the quality issues of the properties available in the Spanish DBpedia. In: Puerta, J.M., et al. (eds.) CAEPIA 2015. LNCS (LNAI and LNB), vol. 9422, pp. 198–209. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24598-0_18
35. Mocnik, F.B., Mobasheri, A., Griesbaum, L., Eckle, M., Jacobs, C., Klonner, C.: A grounding-based ontology of data quality measures. J. Spat. Inf. Sci. **2018**(16), 1–25 (2018)
36. Palmonari, M., Rula, A., Porrini, R., Maurino, A., Spahiu, B., Ferme, V.: ABSTAT: linked data summaries with ABstraction and STATistics. In: Gandon, F., Guéret, C., Villata, S., Breslin, J., Faron-Zucker, C., Zimmermann, A. (eds.) ESWC 2015. LNCS, vol. 9341, pp. 128–132. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25639-9_25
37. Paulheim, H.: Knowledge graph refinement: a survey of approaches and evaluation methods. Semant. Web **8**(3), 489–508 (2017)
38. Paulheim, H., Bizer, C.: Improving the quality of linked data using statistical distributions, vol. 3. IGI Global (2018)
39. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. IEEE Data Eng. Bull. **23**(4), 3–13 (2000)
40. Rashid, M., Rizzo, G., Mihindukulasooriya, N., Torchiano, M., Corcho, O.: KBQ - a tool for knowledge base quality assessment using evolution analysis, vol. 2065, pp. 58–63. CEUR-WS (2017)
41. Rico, M., Mihindukulasooriya, N., Kontokostas, D., Paulheim, H., Hellmann, S., Gómez-Pérez, A.: Predicting incorrect mappings: A data-driven approach applied to dbpedia. In: Proceedings of the 33rd annual ACM symposium on applied computing, pp. 323–330. Association for Computing Machinery (2018)
42. Sejdiu, G., Rula, A., Lehmann, J., Jabeen, H.: A scalable framework for quality assessment of RDF datasets. In: Ghidini, C., et al. (eds.) ISWC 2019. LNCS (LNAI and LNB), vol. 11779, pp. 261–276. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30796-7_17
43. Spahiu, B., Maurino, A., Palmonari, M.: Towards improving the quality of knowledge graphs with data-driven ontology patterns and SHACL. In: Conference of 9th Workshop on Ontology Design and Patterns, pp. 103–117. CEUR-WS (2018)
44. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Commun. ACM **40**(5), 103–110 (1997)
45. Trouillon, T., Dance, C., Gaussier, E., Welbl, J., Riedel, S., Bouchard, G.: Knowledge graph completion via complex tensor factorization. J. Mach. Learn. Res. **18**, 4735–4772 (2017)
46. Vaidyambath, R., Debattista, J., Srivatsa, N., Brennan, R.: An intelligent linked data quality dashboard. In: AICS 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, pp. 1–12 (2019)
47. Weiskopf, N., Weng, C.: Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J. Am. Med. Inform. Assoc. **20**(1), 144–151 (2013)

48. Wienand, D., Paulheim, H.: Detecting incorrect numerical data in DBpedia. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS (LNAI and LNB), vol. 8465, pp. 504–518. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07443-6_34
49. Yoo, S., Jeong, O.: Automating the expansion of a knowledge graph. Expert Syst. Appl. **141**, 112965 (2020)
50. Zaveri, A., et al.: User-driven quality evaluation of DBpedia. In: Proceedings of the 9th International Conference on Semantic Systems, pp. 97–104 (2013)
51. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: a survey. Semant. Web **7**(1), 63–93 (2016)