




Linked data quality assessment: A survey

Aparna Nayak , Bojan Božić , and
Luca Longo 

ML-Labs, Technological University Dublin, Dublin, Republic of Ireland
{aparna.nayak, bojan.bozic, luca.longo}@tudublin.ie

Abstract. Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. There is a colossal amount of linked data available on the web. However, it is difficult to know how well the linked data fit the modeling tasks due to the defects present in the data. Faults present in the linked data spread 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

In recent years, machine learning has received increased interest as a solution for real-world business problems. However, the deployment of machine learning models can present a number of issues and concerns that triggers from the input data quality [36]. The term “data quality” can be defined as “fitness for use” which signifies the term data quality is relative [6]. Thus data with quality considered appropriate for one use may not possess sufficient quality for another use. Data quality in machine learning/artificial intelligence domain has largely been neglected in order to focus more specifically on learning algorithms and methods. Most research in these fields begins with the assumption that the data feeding the algorithms is of high quality – accurate, complete and timely [44]. 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 structured data on the Web. Semantic Web, aims to publish data that is human-readable as well as machine-readable. A large number of published datasets (or sources) that follow linked data principles is currently 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 DBpedia¹ and Wikidata² still suffer from different quality problems [21].

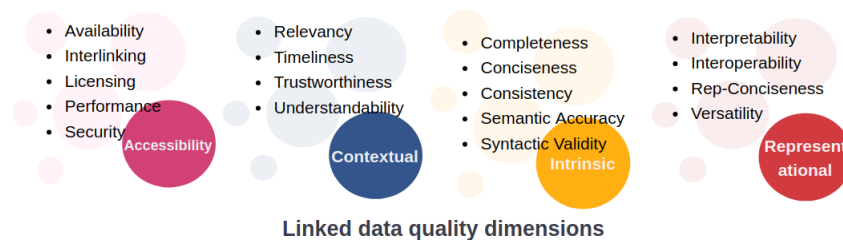
Data quality is being one of the major concern this paper aims to achieve the following objectives:

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 quality 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, section ?? concludes the work.

2 Methods for data quality assessment and improvement



The objective of the data quality assessment activity is to analyze the relevance of data to its consumers and to help publish better quality data. Analysts working

¹ <https://wiki.dbpedia.org/>

² https://www.wikidata.org/wiki/Wikidata:Main_Page

with linked data must assess data quality at various levels such as instance, schema and property. Data quality is a multidimensional concept. Various data quality metrics are clustered by state of the art on data quality into four dimensions based on its usage. These dimensions (refer Fig. 2) are broadly classified as intrinsic, accessible, representational, and contextual [46]. 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. Accessible dimension involves aspects related to the access, authenticity and retrieval of data to obtain either the entire or some portion of the data for a particular use case. Representational dimensions capture aspects related to the design of the data. Contextual dimensions are those that highly depends on the context, such as relevancy, trustworthiness, understandability and timeliness. A comprehensive survey that covers multiple metrics to evaluate each dimension is discussed in Zaveri et al. [53]. It addresses 68 quality metrics of the linked data with a detailed explanation for the calculation of each metric. On the other hand, data quality metrics are divided as a baseline and derived by integrating the metrics defined in [53] and ISO 25012 ³.

The most common problems that directly influence data quality are missing information, missing entity relationships, and erroneous data value. In addition, data conversion from one format to linked data may deteriorate the data quality due to the errors present at the source, parsing values, interpreting and converting units [50]. The integration of data from multiple sources does not guarantee data quality improvement; on the contrary, quality may deteriorate if the sources contain conflicting information [31]. Irrespective of the total number of integrated data sources, quality problems prevail at schema and instance-level [40].

2.1 Ontologies for data cleaning and quality report

This section discusses the ontologies that have been modeled in order to identify the problems associated with data and provide a quality assessment report. Data Quality Management (DQM) vocabulary ⁴, represents better data quality requirements by focusing on the intrinsic quality of the data [20]. This ontology helps to describe the results of data quality assessment and the rules for data cleaning in semantic web architecture. Another ontology representing the data cleaning process, Data Cleaning Ontology (DCO), is described in [4]. DCO is a special case analysis of DQM that is evaluated for data cleaning operations. Data Quality Vocabulary (daQ) ⁵, helps to represent results of data quality assessment in machine-readable format[15]. This ontology provides a core vocabulary to allow the uniform definition of specific data quality metrics that allows data publishers to attach the quality information as part of metadata. W3 has published Data Quality Vocabulary (DQV) [3] to represent data quality assessment in

³ <https://iso25000.com/index.php/en/iso-25000-standards/iso-25012>

⁴ <http://semwebquality.org/dqm-vocabulary/v1/dqm>

⁵ <http://theme-e.adaptcentre.ie/daq/daq.html>

semantic web format⁶. A potential user can use this to represent their data quality assessment report. Fuzzy Quality Data Vocabulary (FQV) extends DQV to represent the fuzzy concepts. Fuzzy 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 available ontologies. 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 compute data quality, approach 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 [53] 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

⁶ <https://www.w3.org/TR/vocab-dqv/>

quality assessment tools focus on either a specific data set or a specific metric mentioned in the Table 2.

A plethora of research focus primarily on various levels of linked data. These levels include schema, instance and properties. (Semi-) structured data is one of the sources for linked data, mapping languages that will be used to map semi-structured data into linked format also impact the data quality [18], [42]. Various quality deficiencies at schema and instance level and resolution strategy have been listed in [7]. Data quality can be assessed with the help 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]. 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

Tools such as ABSTAT [37], Loupe [33], DistQualityAssessment [43], Roomba [2] focus on understanding statistical information which include number of triples, implicit vocabulary information, etc. The statistical information derived from the

tools help the user get insight into the dataset that includes knowing outlier in the vocabulary usage, most frequent patterns in linked data, etc and thus interpreting data quality. Tools/frameworks such as KBQ [41], [45] help in evolution analysis of linked data by comparing all the triples of two consecutive releases of the dataset. Other related work [18], [49] 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], [52].

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 [40]. Knowledge graph refinement [38], 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 consider news and tweets [51]. Though the accuracy of the relation extraction model is low, authors haven't mentioned anything about the quality of the added information.

Techniques such as resolving range violation [28], outlier detection [17], tensor factorization [47] and statistical distribution [39] will help to improve the internal data quality of linked data without referring to external sources. Supervised methods have been discussed in [30], [11] and [10] that tries to add missing links between subject and object. 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 made by publishers that affect quality problems and suggestions on improvement over the entire data set are listed by [13]. However, they have not mentioned about the fine-grained level of quality analysis. In another related study, [48] location of a problematic triple is 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 ⁷ and covers most of the constraints SHACL can validate.

3 Recommendation and future work

Some findings of interest 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 Figure 1 includes the following:



Fig. 1. Stages of ontology based data quality improvement

1. Identify the Knowledge graph A user who wishes to refine their knowledge graph in terms of quality can select the metrics they wish to validate along with ontology if available.

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

2. Identify required metrics Quality assessment requirement varies between datasets. For example, if the considered knowledge graph is an RDF dump, users need not bother about the accessibility of the SPARQL endpoint and server. Also, the user has the flexibility to choose the required metric.

⁷ <https://www.w3.org/TR/shacl/>

From the MKAG example, let us consider a user who wants to verify two quality metrics on MKAG that are syntactically accurate values and no malformed datatype literals.

3. Data quality analysis This stage requires an external domain ontology and a knowledge base. Domain ontology should describe all possible constraints that a knowledge graph may exhibit during the quality assessment process. This ontology relates the data quality metrics to constraints that trigger whenever there is a quality problem. The model applies logical reasoning of the knowledge base over all the triples in the knowledge graph. It will help to identify and locate the problematic triples for further analysis by mapping them to constraints in the ontology.

From the MKAG example, domain ontology of MKAG is present. A reasoner based on description logic will infer all missing triples. 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 is computed on all the properties in the knowledge graph and quality values are computed based on the definitions given to each metric in [53].

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.

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. 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 :datatype mismatch ;
           :triple mk:author ;
           :value mag:paperCount ;
           :datatype xsd:decimal ;
           :expectedDT xsd:integer .

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

:sommemeasurement 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. Some evaluation methods conclude 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 test dataset including all possible violations can evaluate the proposed model.

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. Ahmad, A., Troncy, R., Senart, A.: Roomba: An extensible framework to validate and build dataset profiles. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9341**, 325–339 (2015)
3. Albertoni, R., Isaac, A., Guéret, C., Debattista, J., Lee, D., Mihindukulasooriya, N., Zaveri, A.: 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., Alcântara, J., Vidal, V., Brayner, A., Casanova, M., Pequeno, V., Franco, W.: 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. Communications of the ACM **42**(1), 73–78 (1999)
7. Behkamal, B., Kahani, M., Bagheri, E.: Quality metrics for linked open data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9261**, 144–152 (2015)
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 Reports **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](https://arxiv.org/abs/2001.06917) (2020)
12. Chen, X., Jia, S., Xiang, Y.: A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications* **141** (2020)
13. Csáki, C.: Towards open data quality improvements based on root cause analysis of quality issues. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11020 LNCS**, 208–220 (2018)
14. De Meester, B., Heyvaert, P., Arndt, D., Dimou, A., Verborgh, R.: Rdf graph validation using rule-based reasoning. *Semantic Web Journal* (2020)
15. Debattista, J., Lange, C., Auer, S.: Daq: an ontology for dataset quality information. vol. 1184. *Central Europe Workshop Proceedings, CEUR-WS* (2014)
16. Debattista, J., Auer, S., Lange, C.: Luzzu—a methodology and framework for linked data quality assessment. *J. Data and Information Quality* **8**(1) (Oct 2016)
17. Debattista, J., Lange, C., Auer, S.: A preliminary investigation towards improving linked data quality using distance-based outlier detection. In: *Semantic Technology*. pp. 116–124. Springer International Publishing, Cham (2016)
18. Dimou, A., Kontokostas, D., Freudenberg, M., Verborgh, R., Lehmann, J., Mannens, E., Hellmann, S., Van de Walle, R.: Assessing and refining mappings to rdf to improve dataset quality. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9367**, 133–149 (2015)
19. Färber, M.: The microsoft academic knowledge graph: a linked data source with 8 billion triples of scholarly data. In: *International Semantic Web Conference*. pp. 113–129. Springer (2019)
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*. p. 1–8. LWDM '11, Association for Computing Machinery, New York, NY, USA (2011)
21. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic 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. *Communications in Computer and Information Science* **394**, 265–272 (2013)
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. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11153**, 163–175 (2018)
28. Lertvittayakumjorn, P., Kertkeidkachorn, N., Ichise, R.: Resolving range violations in dbpedia. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10675 LNCS**, 121–137 (2017)
29. Liu, S., d’Aquin, M., Motta, E.: Measuring accuracy of triples in knowledge graphs. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10318 LNAI**, 343–357 (2017)
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. pp. 116–123. In: *ACM International Conference Proceeding Series*. (2012)
32. Mihindukulasooriya, N., García-Castro, R., Gómez-Pérez, A.: Ld sniffer: A quality assessment tool for measuring the accessibility of linked data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10180 LNAI**, 149–152 (2017)
33. Mihindukulasooriya, N., Poveda-Villaón, 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 Mariano, M., García-Castro, R., Gómez-Pérez, A.: An analysis of the quality issues of the properties available in the spanish dbpedia. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9422**, 198–209 (2015)
35. Mocnik, F.B., Mobasher, A., Griesbaum, L., Eckle, M., Jacobs, C., Klöner, C.: A grounding-based ontology of data quality measures. *Journal of Spatial Information Science* **2018**(16), 1–25 (2018)
36. Paleyes, A., Urma, R.G., Lawrence, N.D.: Challenges in deploying machine learning: a survey of case studies. *arXiv preprint arXiv:2011.09926* (2020)
37. Palmonari, M., Rula, A., Porrini, R., Maurino, A., Spahiu, B., Ferme, V.: Abstat: Linked data summaries with abstraction and statistics. In: *The Semantic Web: ESWC 2015 Satellite Events*. pp. 128–132. Springer International Publishing, Cham (2015)
38. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web* **8**(3), 489–508 (2017)
39. Paulheim, H., Bizer, C.: Improving the quality of linked data using statistical distributions, vol. 3. IGI Global (2018)
40. Rahm, E., Do, H.H.: Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin* **23**(4), 3–13 (2000)
41. Rashid, M., Rizzo, G., Mihindukulasooriya, N., Torchiano, M., Corcho, O.: Kmq - a tool for knowledge base quality assessment using evolution analysis. vol. 2065, pp. 58–63. *CEUR-WS* (2017)
42. 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)
43. Sejdiu, G., Rula, A., Lehmann, J., Jabeen, H.: A scalable framework for quality assessment of rdf datasets. *Lecture Notes in Computer Science (including subseries*

- Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11779 LNCS**, 261–276 (2019)
44. Sessions, V., Valtorta, M.: The effects of data quality on machine learning algorithms. *ICIQ* **6**, 485–498 (2006)
 45. 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)
 46. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. *Communications of the ACM* **40**(5), 103–110 (May 1997)
 47. Trouillon, T., Dance, C., Gaussier, E., Welbl, J., Riedel, S., Bouchard, G.: Knowledge graph completion via complex tensor factorization. *Journal of Machine Learning Research* **18**, 4735–4772 (2017)
 48. 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
 49. Weiskopf, N., Weng, C.: Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association* **20**(1), 144–151 (2013)
 50. Wienand, D., Paulheim, H.: Detecting incorrect numerical data in dbpedia. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **8465 LNCS**, 504–518 (2014)
 51. Yoo, S., Jeong, O.: Automating the expansion of a knowledge graph. *Expert Systems with Applications* **141** (2020)
 52. Zaveri, A., Kontokostas, D., Sherif, M., Bühmann, L., Morsey, M., Auer, S., Lehmann, J.: User-driven quality evaluation of dbpedia. In: Proceedings of the 9th International Conference on Semantic Systems. pp. 97–104 (2013)
 53. 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)