Quality Metrics for Linked Open Data

Behshid Behkamal[1(✉)], Mohsen Kahani[1], and Ebrahim Bagheri[2]

[1] Computer Engineering Department, Ferdowsi University of Mashhad, Mashhad, Iran
{behkamal, kahani}@um.ac.ir
[2] Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada
bagheri@ryerson.ca

Abstract. The vision of the Linked Open Data (LOD) initiative is to provide a model for publishing data and meaningfully interlinking such dispersed but related data. Despite the importance of data quality for the successful growth of the LOD, only limited attention has been focused on quality of data prior to their publication on the LOD. This paper focuses on the systematic assessment of the quality of datasets prior to publication on the LOD cloud. To this end, we identify important quality deficiencies that need to be avoided and/or resolved prior to the publication of a dataset. We then propose a set of metrics to measure and identify these quality deficiencies in a dataset. This way, we enable the assessment and identification of undesirable quality characteristics of a dataset through our proposed metrics.

Keywords: Metrics · RDF datasets · Quality deficiencies · Linked open data

1 Introduction

The main goal of the Web of Data initiative is to create knowledge by interlinking dispersed but related data instead of linking related documents in the traditional Web. This massive amount of data on the LOD opens up significant challenges with regards to data quality. Since data is extracted via crowd sourcing of semi-structured sources, there are many challenges with the quality of published datasets. One of the better strategies to avoid such issues is to evaluate the quality of a dataset itself before it is published on the LOD cloud. This will help publishers to filter out low-quality data based on the quality assessment results, which in turn enables data consumers to make better and more informed decisions when using the shared datasets.

The rest of this paper is organized as follows: first, data quality research in the area of the LOD is reviewed in Sect. 2. In Sect. 3, our approach for proposing metrics starts by identifying significant quality issues in RDF datasets, followed by the development of suitable metrics to address the issues. Empirical evaluation of the developed metrics is provided in Sect. 4, and then some guidelines for quality improvement are presented in Sect. 5. Finally, the paper is concluded in Sect. 6.

Q. Chen et al. (Eds.): DEXA 2015, Part I, LNCS 9261, pp. 144–152, 2015.
DOI: 10.1007/978-3-319-22849-5_11

2 Related Works

In this section, we classify the related literature into two main groups: (i) quality problems of published data, and (ii) tools and applications for validation of RDF datasets.

The first group of related work investigates quality problems in published datasets. The most comprehensive work in this group discusses common errors in published datasets [1]. In another work, the quality problems of available datasets such as Geonames and DBPedia are identified using SPARQL queries [2].

The second group of works includes some tools for validating RDF datasets, each with its own error-checking functionalities. Some of them are syntax validators which accept an RDF/XML document as input and check whether the document is syntactically valid. Other kinds of online validators such as URIDebugger[1] and Vapour[2] check the dereferencability of a given URI. Also, there are some command line tools for identifying common errors in RDF documents, such as Jena Eyeball[3] and VRP[4]. Another group of tools are designed for checking common quality issues, such as Lodlaundromat[5] which accepts URLs of dirty datasets and removes syntax errors, duplicates, and blank nodes; and Luzzu[6] which is a framework allowing users to define some metrics and provides queryable quality metadata on the assessed datasets.

Generally, all of these works primarily focus on data quality problems in published datasets, and seldom provide a concrete solution for improving data quality, or attempt to identify the causes of the quality problems before the data is published. In this paper, we deliberate on the importance of filtering out poor quality data by assessing the quality of a given dataset before publishing it.

3 Our Proposed Approach for Metric Development

The objective of our work is to identify quality deficiencies of datasets and to suggest how they can be systematically evaluated before release. To this end, our approach is based on several significant quality issues identified in already published datasets on LOD. To the extent of our experience in publishing and interlinking academic data [3, 4], we found that many of the published datasets suffer from quality issues. We believe that most of these issues have roots in the deficiencies of the sources from where that data is extracted, and they can be avoided if they are identified in the initial stages of publishing. In our work when considering which quality deficiencies to consider, the main criteria for identifying and including a quality issue was based on one of the following criteria:

---

[1] http://linkeddata.informatik.hu-berlin.de/uridbg.

[2] http://validator.linkeddata.org/vapour.

[3] http://jena.sourceforge.net/Eyeball.

[4] http://139.91.183.30:9090/RDF/VRP.

[5] http://lodlaundromat.org/.

[6] http://eis-bonn.github.io/Luzzu/.

- The quality problems should have been spotted within published data and well documented in the literature;
- The quality issues should be detectable and hence avoidable in the preliminary stages of data publication, i.e., prior to their publication and release to the LOD;
- The quality deficiencies are not related to the other datasets, e.g. inconsistency with other published datasets.

Therefore, our approach for metric development starts by identifying quality deficiencies of existing datasets, specifically those that can be avoided or fixed before publishing. We will then propose a set of metrics to address the identified issues, and subsequently the proposed metrics are placed under empirical evaluation.

3.1 Identifying Quality Deficiencies

In this section, we present the quality deficiencies of data within the published datasets focusing on those which are related to the dataset itself and also detectable in the initial phase of publication. To this end, we have classified the quality issues into schema level and instance level as presented in Table 1.

Table 1. Classification of quality deficiencies

Quality deficiency	Issues	Level	Resolution method	Ref
Improper usage of vocabularies	- Not using appropriate existing vocabularies to describe the resources	Schema	Domain Expert	[1, 5]
Redefining existing classes/properties	- Redefining the classes/properties in the ontology that already exist in the vocabularies	Schema	Domain Expert	[1, 5]
Improper definition of classes/properties	- Classes with different name, but the same relations	Schema	Semi-Automated	[3]
	- Properties with different name, but the same meaning	Schema	Ontologist	[6]
	- Inadequate number of classes/properties used to describe the resources	Schema	Domain Expert	[3]
Misuse of data type	- Not using appropriate data types for the literals	Schema	Automated	[1, 2]
Errors in property values	- Missing values	Instance	Automated	[2, 5–9]
	- Out-of-range values	Instance	Automated	
	- Misspelling	Instance	Semi-Automated	
	- Inconsistent values	Instance	Automated	

(*Continued*)

Table 1. (*Continued*)

Quality deficiency	Issues	Level	Resolution method	Ref
Miss-match with the real-world	- Resources without correspondence in real-world	Instance	Domain Expert	[6, 7, 10]
Syntax errors	- Triples containing syntax errors	Instance	Validator	[1]
Misuse of data type / object property	- Improper assignment of object property to the data type property or vice versa	Instance	Validator	[1]
Improper usage of classes/properties	- Using undefined classes/properties	Instance	Semi-Automated	[1]
	- Membership of disjoint classes	Instance	Automated	
	- Misplaced classes/properties	Instance	Validator	
Redundant/similar individuals	- Individuals with similar property values, but different names	Instance	Ontologist	[3]
Invalid usage of Inverse-functional properties	- Inverse-functional properties with void values	Instance	Automated	[2]

In summary, we have identified eleven quality deficiencies characterizing eighteen quality issues at both schema and instance levels. Among all, six quality issues, which their resolution methods are domain expert or ontologist, cannot be detected by any kind of automated methods and needs the intervention of human experts. It is clear that all of these metrics are very subjective and it is hard, if not impossible, to asses them automatically. Among the remaining quality issues, only three issues can be detected and resolved by validators. To the extent of our knowledge, there is no validator to cover all of the remaining issues, particularly the issues relating to incompatibility of schema, naming, and inconsistent data. Thus, we propose a set of metrics to address the remaining quality issues. We note that the identified issues and the following proposed metrics are not meant to be comprehensive and are only limited to the issues reported in the literature.

3.2 Proposed Metrics

In this section, a set of metrics are proposed to address the quality issues extracted from Table 1 that can be resolved in an automated or semi-automated way. To achieve this, metrics, proposed in the areas of the Linked Data, relational databases, and data quality

Table 2. Proposed metrics

Name	Description	Related quality deficiencies
Miss_Vlu	The ratio of the properties defined in the schema, but not presented in dataset	Errors in property values
Out_Vlu	The ratio of the triples of dataset which contain properties with out of range values	Errors in property values
Msspl_Prp_Vlu	The ratio of the properties of dataset which contain misspelled values	Errors in property values
Und_Cls_Prp	The ratio of the triples of dataset using classes or properties without any formal definition	Improper usage of classes/properties
Dsj_Cls	The ratio of the instances of dataset being members of disjoint classes	Improper usage of classes/properties
Inc_Prp_Vlu	The ratio of the triples of dataset in which the values of properties are inconsistent	Errors in property values
FP	The ratio of the number of triples of dataset with functional properties which contain inconsistent values	Errors in property values
IFP	The ratio of the number of triples of dataset which contain invalid usage of inverse-functional properties	Invalid usage of Inverse-functional properties
Im_DT	The ratio of the number of triples of dataset which contain data type properties with inappropriate data types	Not using appropriate data types for the literals
Sml_Cls	The ratio of the classes of dataset with different names, but the same instances	Improper definition of classes

models have been considered [5, 7, 9, 11, 12]; the results of which were taken into account as guidelines for designing a useful set of metrics for our purpose. The proposed metrics are presented in Table 2.

According to the definitions presented for the metrics, it is clear that all of the metrics are defined to measure the quality problems in the scope of the RDF dataset itself, not in the context of other datasets. From the level of quality deficiency point of view, the last two metrics (Im_DT and Sml_Cls) are defined to address intrinsic quality issues at the schema level, while the others are related to the intrinsic quality problems at the instance level. According to [13], the preferred way for metric definition is to calculate the number of the undesirable outcomes divided by that of the total outcomes. Thus, all of the formulas presented for computation of quality deficiencies illustrate the undesirable outcomes using the ratio scale. The proposed metrics are theoretically validated in our recently published work [14].

4 Empirical Evaluation

The main purpose of our work is to propose a set of appropriate metrics to address the quality issues of RDF datasets before their publication. For this purpose, it is necessary to place them under empirical evaluation to observe their behavior and show their applicability in practice. Hence, we first calculated the values of the metrics for eight datasets in order to show the metric behavior over datasets of different domains and sizes. We have selected eight datasets from the EU FP6 Networked Ontology (NeOn) project.[7] We have implemented an automated tool that is able to automatically compute the metric values for any given input dataset. The code of the implemented tool and the employed datasets are publicly available[8].

Next, we manipulated the quality of these datasets by applying some heuristics and then recalculated the metric values to observe the behavior of the metrics over these changes. The aim of our dataset manipulation work was to investigate the trends of metrics over real datasets and to compare the results of applying metrics on good and poor quality data. To this end, fourteen heuristics are introduced to be used in the dataset contamination process. Some of these quality issues such as misspelling errors were made using an ontology editor, i.e. Protégé. For those quality issues such as invalid usage of inverse functional properties, errors were introduced manually. We have randomly applied the heuristics to the different datasets. The rationale for this was to measure the values for our metrics both before and after the quality issues were injected. The aim of the second experiment is to show the trends of the proposed metrics by recalculating the values of the metrics over manipulated datasets. As a result, we would ideally expect to observe meaningful changes in the values of the metrics according to the heuristics used to manipulate the datasets. In light of the results, it is observed that most of the outcomes are desirable; however, some of the values need more discussions which are presented as follows.

Some of the heuristics were not independent and as a result, the order of applying these heuristics affected the results of measuring the quality problems by the metrics. This occurs because introducing some errors into a given dataset can have a number of side effects on other metrics. Thus, the change in one quality issue can implicitly impact other quality issues and therefore, their corresponding metrics. For better investigation of metrics behavior, it is better not to concurrently apply these heuristics on the same dataset.

Another factor affecting our result was related to the ratio of heuristics done over the size of datasets. Based on this experiment, whenever the number of changes is less that 10 % of the triples, the changes of metric values cannot be properly reported.

Although in our scenario, no radical shift in the metric values was observed, but we are not going to generalize our finding about the trends of metrics, because of the limited number of datasets that we have used in this experiment. As a result, we believe more experiments need to be done to reach a valid conclusion about the reaction of metrics to the changes in the measurement subject.

[7] http://www.neon-project.org.

[8] https://bitbucket.org/behkamal/new-metrics-codes/src.

5 Guidelines for Quality Improvement

Based on our discussions about quality deficiencies and based on our experiments in manipulating the datasets, we propose a set of guidelines for data publishers as shown in Table 3. These solutions can be used as guidance for data publishers to improve such quality deficiencies in their datasets before publication.

As shown in Table 3, most of the quality problems measured by our metrics can be easily fixed by the publishers once they are made aware of the issues. Also, it is

Table 3. Guidelines for data publisher

No	Metrics	Related quality issues	Resolution strategy
1	Miss_Vlu	Missing property values	Checking the usage of each property defined in the schema, whenever it is used in the dataset
2	Out_Vlu	Out-of-range property values	For triples with data type properties, checking the defined range for the literal, and for object properties, checking the range of class used as predicate of triple. In both cases, ranges of values should be correct
3	Msspl_Prp_Vlu	Misspelled property values	Using a dictionary to check all of the terms used as literals in a dataset
4	Und_Cls	Using undefined classes	Where classes have been created, we suggest that the term be added to the schema or defined in a separate namespace to enable the reuse of the defined terms
5	Dsj_Cls	Membership of disjoint classes	Checking for this violation of disjoint classes can be done with a reasoner or with an appropriate query
6	Inc_Prp_Vlu	Inconsistent values of properties	Selecting correct values for properties according to the data source and removing triples contain inconsistent values for those properties
7	FP	Functional Properties with Inconsistent Values	Validating user input and checking the uniqueness and validity of functional property values
8	IFP	Invalid usage of inverse-functional properties	Validating user input and checking the uniqueness and validity of inverse-functional values
9	Im_DT	Not using appropriate data types for the literals	Syntactic fixes to the publishing framework and removing or changing the data types of the literals
10	Sml_Cls	Classes with different name, but the same instances	Checking the reported classes by an ontologist for removing similar classes if needed

understood that many of the above deficiencies are a result of not validating user input or due to not using appropriate syntactic validator for the content. Thus, a recommended approach for avoiding these quality issues is using trusted APIs to produce content as well as syntax validator to check the syntactic correctness of datasets.

6 Conclusion and Future Works

In this paper, a set of metrics has been proposed for the assessment of a dataset before its release as a part of the LOD cloud. We have shown how concrete valid metrics can be developed for RDF datasets by implementing such metrics. The proposed metrics have been validated through empirical evaluations.

We are currently focusing on the extension of our work in two main directions *(i)* we are also considering to develop statistical models for predicting the quality dimensions of a dataset using the values of the related metrics. We have undertaken similar studies for building predictive models of quality from metrics in our prior research [15]; and *(ii)* while in this paper, we have focused on quality issues of datasets which can be avoided before release, the quality issues after interlinking into the LOD remain to be further explored.

References

1. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: 3rd International Workshop on Linked Data on the Web (2010)
2. Fürber, C., Hepp, M.: Using semantic web resources for data quality management. In: Cimiano, P., Pinto, H. (eds.) EKAW 2010. LNCS, vol. 6317, pp. 211–225. Springer, Heidelberg (2010)
3. Behkamal, B., Kahani, M., Paydar, S., Dadkhah, M., Sekhavaty, E.: Publishing Persian linked data; challenges and lessons learned. In: 5th International Symposium on Telecommunications (IST), pp. 732–737. IEEE (2010)
4. Paydar, S., Kahani, M., Behkamal, B.: Publishing data of ferdowsi university of mashhad as linked data. In: Computational Intelligence and Software Engineering (2010)
5. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality Assessment for Linked Data: A Survey. Accepted in Semantic Web Journal (2014). http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey
6. Lei, Y., Nikolov, A., Uren, V., Motta, E.: Detecting quality problems in semantic metadata without the presence of a gold standard. In: 5th International EON Workshop at International Semantic Web Conference, pp. 51–60 (2007)
7. Bizer, C., Cyganiak, R.: Quality-driven information filtering using the WIQA policy framework. Web Semant.: Sci., Serv. Agents World Wide Web **7**, 1–10 (2009)
8. Brüggemann, S., Grüning, F.: Using ontologies providing domain knowledge for data quality management. In: Pellegrini, T., Auer, S., Tochtermann, K., Schaffert, S. (eds.) Networked Knowledge - Networked Media. SCI, vol. 221, pp. 187–203. Springer, Heidelberg (2009)
9. Naumann, F., Leser, U., Freytag, J.C.: Quality-driven integration of heterogeneous information systems. In: 25th International Conference on Very Large Data Bases (VLDB 1999), Edinburgh, Scotland, UK, pp. 447–458 (1999)

10. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Commun. ACM **45**, 211–218 (2002)
11. ISO: ISO/IEC 25012- Software engineering - Software product Quality Requirements and Evaluation (SQuaRE). Data quality model (2008)
12. Peralta, V.: Data freshness and data accuracy: A state of the art. Instituto de Computacion, Facultad de Ingenieria, Universidad de la Republica (2006)
13. Eppler, M.J., Wittig, D.: Conceptualizing information quality: A review of information quality frameworks from the last ten years. In: 5th International Conference on Information Quality, pp. 83–96 (2000)
14. Behkamal, B., Kahani, M., Bagheri, E., Jeremic, Z.: A Metrics-Driven approach for quality Assessment of Linked open Data. J. Theoritical Appl. Electron. Commer. Res. **9**, 64–79 (2014)
15. Bagheri, E., Gasevic, D.: Assessing the maintainability of software product line feature models using structural metrics. Softw. Qual. J. **19**, 579–612 (2011)