Metrics-driven Framework for LOD Quality Assessment

Behshid Behkamal

Ferdowsi University of Mashhad, Mashhad, Iran
Behkamal@stu.um.ac.ir

Abstract. The main objective of the Linked Open Data paradigm is to crystallize knowledge through the interlinking of already existing but dispersed data. The usefulness of the developed knowledge depends strongly on the quality of the aggregated and published data. Researchers have observed many challenges with the quality of Linked Open Data; therefore, our main objective in this thesis is to propose a metric-driven framework for evaluating the inherent quality dimensions of datasets before they are published as a viable part of the linked open data cloud.

Keywords: #eswcphd2014Behkamal

1 Introduction

Linked Open Data (LOD) provides a distributed model for the semantic Web that allows any data provider to publish its publicly available data and meaningfully link them with other information sources over the Web. The main goal of the LOD initiative is to create knowledge by interlinking dispersed data. It is undeniable that the realization of this goal depends strongly on the quality of the published data. Therefore, quality evaluation is an important issue that must be addressed with the objective of helping data providers to evaluate their data before publishing as a dataset in the LOD cloud.

In the area of data quality assessment, researchers have developed several frameworks, metrics and tools to evaluate data quality in general. For example, [1] describes subjective and objective assessments of data quality and presents three functional forms for developing objective data quality metrics. In [2], the authors have proposed a methodology for the assessment of organizational Information Quality (IQ), which consists of a questionnaire to measure IQ. In the area of the methodologies for data quality assessment, [3] provides a comparative description of existing methodologies and provides a comprehensive comparison of these methodologies. Also, the database community has developed a number of approaches such as user experience, expert judgment, sampling, parsing and cleansing techniques [4, 5] for measuring and enhancing data quality.

While data quality is an important requirement for the successful organic growth of the LOD, only a very limited number of research initiatives exist, which focus on data quality for the Semantic Web and specifically for LOD. Based on our practical experience in publishing linked data [6], we have observed that many of the published

datasets suffer from quality issues such as syntax errors, redundant instances, and incorrect/incomplete attribute values. One of the better strategies to avoid such issues is to evaluate the quality of a dataset 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.

2 State of the art

Here we primarily focus on data quality with respect to Semantic Web and LOD, but also briefly touch upon quality assessment frameworks as relevant to our work.

2.1 Information Quality (IQ) frameworks and quality models

Many attempts have been made to compile and classify information quality criteria with different goals in mind. Naumann identifies three different kinds of classifications including goal-oriented models; semantic-oriented models; and processing-oriented models [4]. Since we are going to extract inherent quality dimensions and customize them for LOD, we have systematically reviewed the semantics-oriented quality models and frameworks focusing on those models proposing inherent or intrinsic quality characteristics [7-10]. Given the models presented in [7] and [10] are proposed specifically for databases and data warehousing, they cannot be applied directly to our work. Only [9] investigates the quality dimensions in the context of our work which classifies the quality dimensions and criteria proposed by other researches in the LOD domain. Also, ISO 25012[8] is a general data quality model which defines quality dimensions from inherent and system dependent viewpoints.

2.2 Data Quality in the context of Semantic Web and Linked Data

Despite its importance, data quality has only recently being receiving attention from the Semantic Web community. Most of related works in the context of quality assessment of LOD investigate the quality problems of the published datasets. For example, the authors of [11] have proposed a comprehensive approach that classifies quality problems of the published linked datasets and discuss common errors in RDF publishing, their consequences for applications, along with possible publisher-oriented approaches to improve the quality of machine-readable and open data on the Web. In another work, Furber and Hepp propose an approach to evaluate the quality of datasets using SPARQL queries in order to identify quality problems. Using this approach, the authors identify quality problems of already available datasets such as Geonames and DBPedia [12]. There are also a number of works focusing on the quality evaluation of ontologies [13] that we have not investigated them, because our aim is the quality evaluation of datasets.

Furthermore, some tools are developed for identifying common syntax errors in RDF documents in two groups of online validators, e.g. URIDebugger [14] and Va-

pour [15], and command line validatores, such as Jena Eyeball[16] and VRP [17]. Generally, all of these works primarily focus on data quality problems in published datasets, and none of them provides a solution for identifying the quality problems before the data is published. In this paper, we argue the importance of applying a quality model for assessing the quality of a given dataset before its publications a part of the linked open data cloud.

3 Problem Statement and Contribution

Although data quality is an important issue for the successful organic growth of the Web of Data, there are only a very limited number of research initiatives that focus on data quality for the Semantic Web and specifically for the Web of Data. Based on our practical experience in publishing linked data [6], we have recognized that many of the published datasets suffer from quality deficiencies, most of which are related to inherent quality aspects of a dataset and not the context of other datasets. Thus one of the better strategies to avoid quality issues of the published datasets is to assess the quality of a dataset before release. This will help publishers to filter out low-quality data based on quality assessment results, which in turn enables data consumers to make better and more informed decisions when using shared datasets.

Therefore, the objective of our work is to propose a metrics-driven framework that enables the automatic assessment of the quality of dataset before they are publicly published. For this purpose, we will explore the structural characteristics of data published in LOD cloud as well as the quality deficiencies of dataset itself. In other words, we will try to observe and clearly formulate a set of metrics that are quantitatively measurable for a given dataset. We will then try to find meaningful statistical correlations between the proposed metrics and inherent quality dimensions (that are not directly measurable) through empirical observational studies. Based on the correlations, we will create a framework that will be able to predict the inherent quality dimensions of datasets by only observing their measurable metrics.

For this purpose, we have identified the characteristics of data published on the LOD cloud to extract the inherent quality dimensions and propose a set of metrics, which are quantitatively measurable for a given dataset. This way, we are able to assess inherent quality characteristics of datasets before publishing the data by observing the measured values of the relevant metrics. The novel contributions of our work can be summarized as follows:

- We clearly identify a set of important inherent quality characteristics for LOD datasets based on existing standard quality models and frameworks, e.g. ISO-25012.
- We systematically propose and validate a set of metrics for measuring the quality characteristics of datasets before they are published to the LOD cloud.
- We propose a quality model for LOD that considers the inherent data quality indicators of such data.

- We introduce a novel approach for the assessment of the quality of datasets on LOD, which has its roots in measurement theory and software measurement techniques.

4 Research Methodology and Approach

Our approach for data quality assessment involves the measurement of quality dimensions focusing specifically on inherent quality aspects of linked open datasets. To achieve this goal, we have applied following approach:

1. Exploratory analysis of the previous and current well-known models and frameworks on data quality and comparing dimensions and indictors of data quality presented in these models;
2. Selecting the most appropriate quality model and extracting a subset of quality dimensions that could be applied to inherent quality characteristics of LOD datasets;
3. Devising a set of metrics for assessment of selected inherent quality dimensions;
4. Theoretical validation of proposed metrics;
5. Proposing a quality models consist of selected quality dimensions and proposed metrics;
6. Implementing an automated tool for measuring the proposed metrics;
7. Empirical evaluation of the quality model by measuring the quality metrics of various dataset;
8. Developing a questionnaire for subjectively evaluation of the datasets used in the previous step;
9. Developing predictive statistical (machine learning)-based techniques to find a correlation between proposed metrics with inherent quality dimensions;
10. Applying final framework o predict the inherent quality of datasets by measuring the metrics.

According to this approach, we have undertaken an exploratory analysis of the previous and current well-known models and frameworks on data quality focusing the models and Frameworks varying in their approach and application, but sharing a number of characteristics [7, 9, 10, 18-21]. We systematically review these data quality models and Frameworks focusing on those models proposing inherent or intrinsic quality [7-10]. Comparing existing dimensions and indictors of data quality presented in these models, we tried to identify the most appropriate quality dimensions that could be applied to inherent quality characteristics of LOD datasets. These inherent quality characteristics are namely completeness, semantic accuracy, syntactic accuracy, uniqueness, consistency and interlinking.

In order to make the characteristics quantifiable, we define a set of metrics to measure the above six inherent quality characteristics. The employed approach for metric definition is Goal-Question-Metric(GQM) [22]. In GQM, the goals are gradually refined into several questions and each question is then refined into metrics. Also, one metric can be used to answer multiple questions. Considering the fact that only

few studies have been conducted which define quality metrics for LOD [5, 9, 23], we had to define the required metrics from scratch and prior work could not be reused for our purpose. We propose 32 metrics as measurement references for the inherent quality of linked open dataset to address six inherent quality dimensions.

The main idea behind the design of these metrics has been comprehensiveness and simplicity. To achieve comprehensiveness, we have tried to cover as many quality deficiencies of a dataset as possible that can be identified at the time of publishing, which is the focus of our work. We have also considered as much structural characteristics of a dataset as possible. Therefore, the metrics are proposed in two main groups: quality-driven and structural. Quality-driven metric measures specific quality deficiency in a given dataset e.g. redundant instances in a dataset; while structural metrics represent a feature of any dataset presented in the RDF model, and is not related to the quality issues that might exist in those datasets, e.g. ratio of properties to classes. Furthermore, we have tried to define all of metrics in simple ratio scale. Taking into account of proposed metrics, it is understood that developing simple metrics is our secondary objective.

After defining metrics, a hierarchical data quality model focusing on the inherent viewpoint is developed as shown in Figure 1. At the first level, it consists of six inherent quality dimensions, namely interlinking, uniqueness, consistency, syntactic accuracy, semantic accuracy and completeness. The quality metrics for assessing these quality dimensions are proposed at the second level of the model, some of which are used for measuring two quality dimensions. There are 32 metrics in this model, each of which are assigned by a number and defined in Table 2.

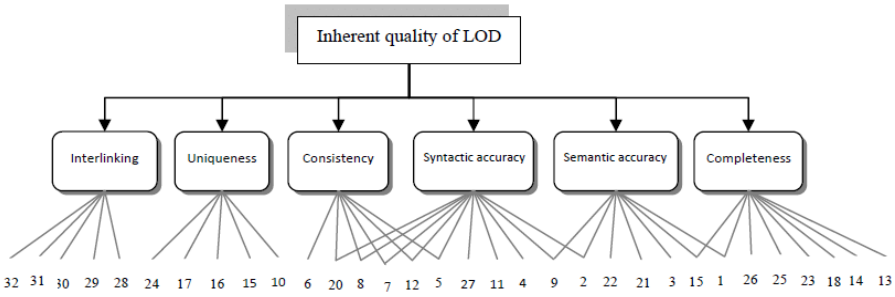

Figure 1- The structure of LODQM

5 Intermediate Results

An important outcome of this work is the evaluation of our assumption that a dataset within LOD can be automatically processed using metrics and evaluated based on statistical predictive models for measuring its quality before release. We expect that given the metrics values for a dataset, the developed predictive models are able to estimate/predict the values of inherent quality dimensions that are not directly measurable. Generally, the main outcomes that we have achieved are the following:

- Identifying the inherent quality dimensions of LOD that can be assessed before publishing;
- Formulating a set of measurement-theoretic metrics for assessing the inherent quality of LOD;
- Developing a quality model for LOD by customizing ISO-25012;
- Achieving an innovative solution for quality assessment in the context of linked data in the early stage of publishing.

In order to put proposed metrics into practice, we have implemented a tool that is able to automatically compute the metric values for any given input dataset. The code is implemented in the Java programming language (JDK 7 Update 25 x64) using Jena 2.6.3 semantic web library. For better observation of metric behavior, different datasets from a variety of LOD domains are selected. Our codebase and selected datasets for this study are publicly accessible at [24, 25], respectively. Here, we have reported the results of our experiments over three datasets. The results of our observations over all of the datasets are reported in [26]. Table 1, presents the details of the selected datasets; and Table 2 summarizes the results of our experiments over them.

Table 1- The details of the datasets used in our experiments

Datasets	No. of triples	No. of instances	No. of classes	No. of properties
DS1- FAO Water Areas	5,365	293	7	19
DS2- Water Economic Zones	25,959	693	22	127
DS3-Large Marine Ecosystems	6,006	358	9	31

Table 2 - The results of our experiments

No.	**Metrics**	DS1	DS2	DS3
1	Missing properties values (Miss_Prp_Vlu)	0.67	0.26	0.44
2	Out-of-range properties values (Out_Prp_Vlu)	0.84	0.81	0.78
3	Misspelled property values (Msspl_Prp_Vlu)	0.84	1.00	0.85
4	Undefined classes (Und_Cls)	1.00	1.00	1.00
5	Membership of disjoint classes (Dsj_Cls)	1.00	1.00	1.00
6	Inconsistent properties values (Inc_Prp_Vlu)	0.80	0.81	0.81
7	Functional properties with inconsistent values (FP)	1.00	1.00	1.00
8	Invalid usage of inverse-functional properties (IFP)	1.00	1.00	1.00
9	Improper data types for the literals (Im_DT)	1.00	1.00	1.00
10	Similar classes (Sml_Cls)	0.71	0.23	0.89
11	Undefined properties (Und_Prp)	1.00	0.72	1.00
12	Using disjoint properties (Dsj_Prp)	1.00	1.00	1.00
13	Unused classes (Unusd_Cls)	0.86	0.50	0.89
14	Unused properties (Unusd_Prp)	0.74	0.76	0.74
15	Similar properties (Sml_Prp)	1.00	0.95	1.00

No.	Metrics	DS1	DS2	DS3
16	Using similar properties (Usg_Sml_Prp)	1.00	0.95	1.00
17	Redundant triples (Rdn_Trp)	0.90	1.00	0.91
18	Heterogeneity of data types (DT)	3.00	3.00	3.00
19	Average missing properties (Avg_Miss_Prp_Vlu)	0.67	0.24	0.44
20	Misusage of properties (Msusg_Prp)	1.00	1.00	1.00
21	Misspelled classes (Msspl_Cls)	0.58	0.45	0.55
22	Misspelled properties (Msspl_Prp)	0.71	0.55	0.67
23	Ratio of properties to classes (Prp_Cls)	2.71	5.77	3.44
24	Redundant instances (Rdn_Ins)	0.00	0.00	0.00
25	Ratio of instances to classes (Ins_Cls)	41.86	31.50	39.78
26	User-defined properties (User_Def_Prp)	0.00	0.86	0.00
27	Misplaced classes/properties (Misplc_Cls_Prp)	1.00	1.00	1.00
28	Object properties (Obj_Prp)	0.13	0.21	0.13
29	Imported triples (Imp_Trp)	0.00	0.88	0.89
30	External linking (Ext_Lnk)	1.00	0.09	1.00
31	Connectivity of RDF graph (Gr_Cn)	0.02	0.01	0.02
32	Intra-linking (Int_Lnk)	0.96	0.98	0.96

In this step of our research, we are not able to address all of the research questions presented in Section 3. We can only answer RQ1 and RQ2. In response to RQ1, we have identified six inherent quality dimensions of linked open datasets that are presented in the first level of LODQM as shown in Figure 1. Regarding the second question (RQ2), a set of automatic measurable metrics is defined to measure six quality dimensions, as presented in Table 2. Currently, we are collecting the experts' subjective perception about inherent quality dimensions to find relations between the measured values of the metrics and perceived quality. Thus, the other research questions, RQ3 and RQ4, can be addressed after completion of the work.

6 Evaluation Strategy

In previous section, the results of empirical evaluation of the proposed metrics are presented. Here, we theoretically support our claim by validation of the metrics and evaluation of the quality model. Initially, the proposed metrics are validated from a measurement-theoretic perspective, and subsequently, the suitability of the proposed quality model will be discussed. Furthermore, we are going to subjectively evaluate our proposed model using expert' opinion as will be explained in the conclusion.

6.1 Theoretical Validation

Generally, any kind of measure is a homomorphism from an empirical relational system to a numerical relational system[27]; therefore, it is imperative that measures be theoretically analyzed within the framework of measurement theory. There are two main groups of frameworks for the theoretical validation of metrics in the literature. The first group consists of frameworks directly based on measurement theory principles [28]; while the second group expresses the desirable properties of the numerical relational system that need to be satisfied by the metrics [29]. In this work, we have examined the properties of our metrics according to one of the most well-known frameworks in the latter group, namely Property-based measurement framework [29]. This framework provides five types of metrics including size, length, complexity, coupling and cohesion and offers a set of desirable properties for each of these types.

Since, all of the proposed metrics are of the size type and according to [29], they are expected to exhibit three main properties, namely, non-negativity, null value and additivity. In other words, size cannot be negative (non-negativity), and it is expected to be null when a system does not contain any elements (null-value). Also, when modules of a system do not have any elements in common, we expect size to be additive (additivity).We have analyzed these three important properties for our proposed metrics and recognized that all of the metrics respect the properties required by the property-based measurement framework to form valid metric space.

6.2 Criteria Based Evaluation

In this section, we evaluate our proposed quality model, LODQM, according to criteria in two dimensions of analytical criteria and practical criteria. These meta-criteria are presented in [30] to analyzes seven well-known conceptual frameworks on information quality. The analytical criteria require clear definitions of the terms used in a framework, a positioning of the framework within existing literature, and a consistent and systematic structure. The practical dimension consists of criteria which make the framework applicable, namely conciseness and the inclusion of tools that are based on the framework. To better evaluation of the our model based on these meta-criteria, we have answer the questions corresponding to meta-criteria which is proposed in [30].

- Definitions: The exact definitions for all of the quality metrics and quality dimensions are presented in LODQM.
- Positioning: LODQM is clearly positioned within existing information quality literature in the context of LOD.
- Consistency: LODQM is divided into systematic dimensions that are collectively exhaustive. Since, our model has used GQM approach for metric development; there are common metrics for different quality dimensions.
- Conciseness: The quality model has six quality dimensions with 5-7 metrics for each of which.
- Tools: An automated tool is developed to measure the values of the proposed metrics for any input dataset in RDF format. Also, a questionnaire is developed

and will be applied to capture the experts' opinion in for subjectively evaluation of our model.

According to above discussions, it is clear that the proposed quality model is a practical model for any datasets of LOD.

7 Conclusion

The goal of this research is proposing a metrics-driven framework for predicting the quality of linked open datasets from an inherent point of view. To achieve this goal, we have followed an approach which is started by analysis of the well-known IQ frameworks, resulting in selection of six quality characteristics. Then, a set of metrics for assessing each of six quality dimensions are developed including quality-driven and structural. To put the proposed metrics into practice, we have implemented an automated tool and computed the metric values for various datasets from different domains of LOD. Finally, the suitability of the LODQM metrics is discussed.

In the next phase of our work, we are going to investigate and analyze whether the proposed metrics can be good early indicators of inherent dimensions. Following our approach, we use questionnaire to receive experts' subjective perception regarding inherent quality dimensions for all of the datasets used in this experiment to find relations between the measured values for the metrics and perceived quality by collecting the opinions of the experts in LOD domain. If the proposed metrics are shown to have meaningful correlation with the quality dimensions, then we are able to predict the inherent quality dimensions of any dataset once it is integrated into the LOD, by only observing the values of proposed metrics. The results will help publishers to filter out low-quality data, which in turn enables data consumers to make better and more informed decisions when using the shared datasets.

Acknowledgement

I would like to gratefully thank Prof. Mohsen Kahani for his brilliant guidance and Dr. Ebrahim Bagheri for his great advice during my work.

References

1. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Communications of the ACM 45, 211-218 (2002)

2. Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, R.Y.: AIMQ: a methodology for information quality assessment. Information & management 40, 133-146 (2002)

3. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR), vol. 41, pp. 16 (2009)

4. Naumann, F., Rolker, C.: Assessment methods for information quality criteria. In: 5'th Conference on Information Quality pp. 148-162. (2000)

5. Batini, C., Scannapieca, M.: Data quality: concepts, methodologies and techniques. Springer (2006)
6. 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)
7. Helfert, M.: Managing and measuring data quality in data warehousing. In: World Multiconference on Systemics, Cybernetics and Informatics, pp. 55-65. (2001)
8. ISO: ISO/IEC 25012- Software engineering - Software product Quality Requirements and Evaluation (SQuaRE). Data quality model, (2008)
9. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., Hitzler, P.: Quality Assessment Methodologies for Linked Open Data. Submitted to Semantic Web Journal (2013)
10. Wang, R.Y., Strong, D.M., Guarascio, L.M.: Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems 12, 5-33 (1996)
11. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: 3rd International Workshop on Linked Data on the Web (LDOW2010). (2010)
12. Fürber, C., Hepp, M.: Using semantic web resources for data quality management. Knowledge Engineering and Management by the Masses, pp. 211-225. Springer (2010)
13. Brank, J., Grobelnik, M., Mladenić, D.: A survey of ontology evaluation techniques. (2005)
14. URI Debugger, http://linkeddata.informatik.hu-berlin.de/uridbg
15. Vapour online validator, http://validator.linkeddata.org/vapour
16. Jena Eyeball: Command line validator, http://jena.sourceforge.net/Eyeball
17. VRP: Command line validator, http://139.91.183.30:9090/RDF/VRP
18. Dedeke, A.: A Conceptual Framework for Developing Quality Measures for Information Systems. In: 5th International Conference on Information Quality, pp. 126-128. (2000)
19. Naumann, F., Rolker, C.: Do Metadata Models meet IQ Requirements? In: Iternational Conference on Information Quality (IQ), pp. 99-114. (1999)
20. Su, Y., Jin, Z.: A Methodology for Information Quality Assessment in Data Warehousing. In: Communications, 2008. ICC'08. IEEE International Conference on, pp. 5521-5525. IEEE, (2008)
21. Moraga, C., Moraga, M., Caro, A., Calero, C.: Defining the intrinsic quality of web portal data. In: 8th International Conference on Web Information Systems and Technologies (WEBIST), pp. 374-379. (2012)
22. Caldiera, V.R.B.G., Rombach, H.D.: The goal question metric approach. Encyclopedia of software engineering 2, 528-532 (1994)
23. Hartig, O.: Trustworthiness of data on the web. In: Proceedings of the STI Berlin & CSW PhD Workshop. Citeseer, (2008)
24. The code of metrics calculation tool https://bitbucket.org/behkamal/new-metrics-codes/src
25. Networked Ontology (NeOn) project, http://www.neon-project.org
26. Behkamal, B., Kahani, M., Bagheri, E., Jeremic, Z.: A Metrics-Driven Approach for Quality Assessment of Linked Open Data. Accepted in Journal of Theoretical and Applied Electronic Commerce Research (2013)
27. Fenton, N.E., Pfleeger, S.L.: Software metrics: a rigorous and practical approach. PWS Publishing Co. (1998)
28. Poels, G., Dedene, G.: Distance-based software measurement: necessary and sufficient properties for software measures. Information and Software Technology 42, 35-46 (2000)

29. Briand, L.C., Morasca, S., Basili, V.R.: Property-based software engineering measurement. Software Engineering, IEEE Transactions on 22, 68-86 (1996)
30. 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)