Web Data Quality:
Current State and New Challenges

Amrapali Zaveri[1], Andrea Maurino[2] and Laure-Berti Equille[3]

1 Universität Leipzig, Institüt für Informatik, AKSW, Postfach 100920, D-04109, Leipzig, Germany
2 University of Milano Bicocca, Department of Computer Science Viale Sarca 336, Milano, Italy
3 Qatar Computing Research Institute, Tornado Tower 18th, P.O. Box 5825 Doha, Qatar

Abstract

The standardization and adoption of Semantic Web technologies has resulted in an unprecedented volume of data being published as Linked Data (LD). However, the "publish first, refine later" philosophy leads to various quality problems arising in the underlying data such as incompleteness, inconsistency and semantic ambiguities. In this article, we describe the current state of Data Quality in the Web of Data along with details of the three papers accepted for the International Journal on Semantic Web and Information Systems' (IJSWIS) Special Issue on Web Data Quality. Additionally, we identify new challenges that are specific to the Web of Data and provide insights into the current progress and future directions for each of those challenges.

Web Data Quality

The standardization and adoption of Semantic Web technologies has resulted in an unprecedented volume of data being published as Linked Data (LD)[1]. However, the "publish first, refine later" philosophy leads to various quality problems arising in the underlying data such as incompleteness, inconsistency and semantic ambiguities. These problems affect every application domain, be it scientific (e.g., life science, environment), governmental or industrial.

Traditional efforts focused on data quality assessment include the definition and modeling of several data quality dimensions (or criteria), such as data completeness, accuracy, timeliness, consistency or detection of duplicates. However, in the Web of Data, many other dimensions are particularly relevant considering the uniform structure of the data, such as representational conciseness, interoperability as well as interpretability of the data. Logical or formal consistency, trustworthiness and relevancy are yet another set of dimensions that are important for the quality of the Web of Data. Even though these dimensions have been identified, the means to measure them, that is, their corresponding metrics, are still needed and thus Web Data Quality opens a new line of research.

The quality in LD is advocated as an essential concept; however, few efforts are currently in place to standardize how Web Data Quality monitoring, assessment and improvement should be implemented. LD, in particular, presents new challenges that were not handled before in other research areas. Thus, adopting existing approaches for assessing the

[1] *http://lod-cloud.net/state/*

quality in LD is not a straightforward solution. These challenges are related to the openness of the LD, the diversity of the information and the unbounded, dynamic set of autonomous data sources and publishers. Providing semantic links is another new aspect that requires an initial exploration and understanding of the data. Additionally, detecting the quality of data sets available and making the information explicit is yet another challenge. Moreover, the current approaches do not use the assessment to ultimately improve the quality of the underlying data set.

Moreover, there are several issues in LOD which hampers the use of data sets in building real-world LOD-based applications and research solutions. One of the challenges is related to resource discovery to find the most relevant LOD data sets for a particular application. Also, generating meaningful associations between the data sets, at the ontology, data (instance) or property level is an important issue to be considered when building such applications.

Besides data sets, the quality of ontologies is yet another important aspect on the Web of Data. The vocabulary, syntax, structure, semantics, representation, consistency and context of an ontology are various means by which it can be evaluated. There exist a number of metrics to evaluate the accuracy, adaptability, completeness, computational efficiency, consistency and organizational fitness of ontologies [21]. However, there still exists a lack in domain and task-specific solutions for evaluation as well as experimental verification and improvement of the quality of ontologies. The current special issue intends to address some of these challenging questions as well as to inspire new research contributions in the field of Web data quality.

Summary of Accepted Papers

The special issue received 24 papers that were reviewed by at least three anonymous and independent reviewers. After the first round of reviews, only five papers were invited to submit an improved version for further evaluation and, of those, only three papers (about 8% of the total number of submitted papers) were finally published.

The first article "*Improving Curated Web-Data Quality with Structured Harvesting and Assessment*" by Kevin Chekov Feeney, Declan O'Sullivan, Wei Tai and Rob Brennan from the Trinity College Dublin in Ireland, describes a methodological framework, namely DaCura, and a set of tools for harvesting, assessing, improving and maintaining high-quality Linked Data. The DaCura framework is focused to support LD that evolves and thus maintains this data by preserving the quality over time. The topic of the paper is important from both research and practitioners' point-of-view; in fact, the "publish first, refine later" philosophy had the great merit to increase the volume of LD available on the Web, but this was realized by providing less attention to data quality requirements. The DaCura framework thus plays a significant role here as it can help the data provider improve the quality of published data. A working implementation has been produced and applied to the publication of a data set in Social Sciences. Experimental results quantify the promising results of the DaCura process and tools on data quality.

The second article "*Improving the Quality of Linked Data Using Statistical Distributions*" by Heiko Paulheim and Christian Bizer from the University of Mannheim in Germany, is a good example on how the use of statistical tools can significantly improve the quality of LD. Linked Data is often generated from existing data and, as a consequence, if original data is dirty and noisy, the generated LD will be messy as well. In this paper, the authors present two algorithms that exploit statistical distributions of data properties and data types for enhancing the quality of incomplete and noisy Linked Data sets: SDType adds missing type statements and SDValidate identifies faulty statements. An interesting concept introduced is to improve semantic data without using any external knowledge, i.e., the proposed algorithm exploits solely the data set itself.

The third article "*OOPS! (Ontology Pitfall Scanner!): An On-line Tool for Ontology Evaluation*" by María Poveda-Villalón, Asunción Gómez-Pérez and Mari Carmen Suárez-Figueroa from the Universidad Politecnica de Madrid in Spain, introduces the quality issues in the modeling phase of ontologies which represents an important component in the Web of Data since a large volume of data, annotated by means of ontologies, is shared on the Web.

It is worth noting that the accepted papers are mainly related to LD and in fact about 40% of all the submitted papers discussed quality issues in LD, while other discussed quality problems in the Web of Data including Open Data, Semantic Web services and Web pages.

In [4], the authors underlined that the marriage of Semantic Web and Web engineering would have brought old and new quality issues (with, amongst others, trust, linkage and privacy). The selected papers introduced new issues (including the curation of LD and ontology design) and the approaches showed that the problem of quality in Web of Data is far from being solved.

Research Challenges of Web Data Quality

As demonstrated by this special issue on Web Data Quality, we observe that this topic is emerging and is becoming a major concern from both research and industry. From a historical point-of-view, after the first phase of the Web of Data where the most important problem was to publish data, now we are entering in a second phase where both data providers and data consumers demand for more quality in published data. This new era is represented by the growing number of different quality dimensions, related metrics and tools proposed (see [25] for a survey) and also the increasing number of empirical evaluations of quality in published LD. Based on the experience from other research (and industrial) fields, we identify the following non-exhaustive set of challenges:

- New quality dimensions
- Data profiling for the Web of (big) Data
- Data Quality Assessment Tools

- Maintenance of published Data
- Quality-based search engine/query engine

New Quality Dimensions

The new format of data, in this case RDF, entails new and specific data quality dimensions and related metrics. By considering the originality in the definition of very large sets of interlinked data, it is clear that new quality requirements are needed in order to fully exploit them. There are a number of studies, which have defined and/or grouped data quality dimensions into different classifications [3,7,13,16,18,22,23]. Recently, there are a number of data quality dimensions that have been identified relevant to LOD, namely, accuracy, timeliness, completeness, relevancy, conciseness, consistency [5]. Further quality criteria such as uniformity, versatility, comprehensibility, amount of data, validity, licensing, accessibility and performance are also introduced as means of assessing the quality of LD [8]. In [25], the authors unify and formalize 18 data quality dimensions and metrics from several existing studies to provide a standardized classification along with a definition for each data quality dimension.

The most important and still unexplored data quality requirements are:

- Interlinking
- Believability
- Licensing

Interlinking refers to the problem of connecting new data sources with existing data on the Web. From a data quality perspective, there is the need to understand, discover and link the most appropriate existing data with the new one. This is a complex problem not only in terms of the size of the existing Web of Data, but also includes philosophic problems in terms of the choice of the correct link type, for example, the correct use of the *owl:sameAs* property (see [11] for further analysis).

Believability is defined in [17] as "the extent to which data are accepted or regarded as true, real and credible". This dimension can be considered as the result of the compositions of three quality dimensions namely trustworthiness, reliability of data and temporality of data. While interlinking is probably the most important quality problem from a data producers view point, believability is the most crucial quality issue from a data consumer; and it is more and more complex due to the nature of the Web of data where everyone can publish and link data. It is worth nothing that according to the provided definition of believability, it includes other relevant quality requirements such as time related dimensions [20], provenance and trust of data [12].

Licensing is a cross-cutting quality requirement due to the fact that it affects not only the legal framework but also raises new challenges in computer science. In fact, the absence of clarity about the licensing terms under which the data is released prevents data re-use, and thus data publication and interlinking at the expenses of the Web of Data itself. In [19], the authors propose deontic logic semantics, which is able to formally define the deontic components of the licenses, i.e., permissions, obligations, and prohibitions. With these components, one is able to reason over the licenses, verify the compatibility of the elements composing the single licenses, return those elements which can be included into the composite license and provide a formal account of the heuristics proposed to guide the composition.

Data Profiling for the Web of (big) Data

The definition of a large set of quality dimensions and corresponding metrics is crucial for designing useful data profiling tools. Recently, an approach for profiling LOD has been proposed [1], which allows a user to perform real-time profiling on the data sets and thus perform quality assessment on them. The authors propose a tool, which allows clustering of the data, which helps to identify the domain(s) of the data and their representative schema. Moreover, the user can interactively learn more about the data during subsequent profiling steps and perform on-the-fly profiling.

Due to the lack of apriori known schema describing a single data set in the Web of Data, the interlinking dimension mentioned earlier requires to extract and summarize existing data. Moreover, in the case of LD, due to the lack of strong typing systems, the approach proposed in [14] could be considered as a type of data profile activity.

Data Quality Assessment Tools

There exist several approaches towards developing frameworks for assessing the data quality of LD. These frameworks can be broadly classified into: (i) automated (e.g., [10]), (ii) semi-automated (e.g., [8]) or (iii) manual (e.g., [6,15]) methodologies. These approaches are useful at the process level wherein they introduce a methodology to assess the quality of a data set. However, their drawbacks include considerable efforts required from the user, inability to produce interpretable results and that they do not allow the user to choose the input data set.

From the above mentioned and related areas of research, a new generation of assessment tools are emerging which enable the integration of automatic tasks with human tasks in a crowdsourcing approach LD quality assessment. Crowdsourcing is highly appropriate for any assignment involving large to huge numbers of small tasks that require human judgment. In terms of LD, crowdsourcing quality assessment may involve, for example, verifying the completeness or correctness of a fact wrt. the original data set. Such a task does not require underlying knowledge about the structure of the data and can be done fairly quickly, without bias and cost effectively [24]. In a recent study [2], a comparison between assessments done by LD experts (those conversant with RDF) and Amazon Mechanical Turk[2] workers on the DBpedia data set was undertaken. The study showed promising results, in terms of time as well as cost efficiency, for using a combination of crowdsourcing as well as manual (by experts) methodologies for quality assessment.

[2] *https://www.mturk.com/mturk/welcome*

Maintenance of published Data

As shown in many data quality methodologies [3], after the assessment phase, it is possible to improve existing data by taking into account the results from the assessment phase. This maintenance phase becomes challenging in the Web of Data mainly because Web Data is generated from existing data and thus its correction can be even more difficult and time consuming when provenance meta-information are not available anymore. Additionally, even a relatively small problem in a data set can ultimately affect the quality of multiple interlinked data sets, much like the butterfly effect [9]. Moreover, there are some methodological new questions related to the maintenance of existing Web of data that are open: How to advertise (in the Web of Data) the modification of one data source?; How can one evaluate if existing links among data still hold after a modification?; How can one prevent the propagation of errors in an interlinked Web of Data? What happens to old data - should they be removed from the Web of Data?; or is it better to not remove them for preserving existing links and in such cases what about the believability of old data?

Quality-based search engines

Finally, when existing Linked Data sets are exposed or tagged with some quality and provenance metadata, it can be possible to define a new generation of quality-based search engines or query engines, which rely on this information to deliver useful and relevant results [16]. In order to provide the answer to user queries in a meaningful way, it is necessary to define what should be in the result, how it can be obtained and how one should represent the query result [16]. In this case, the data model issues such as representation of the data play an important role. On the other hand, completeness, consistency (logical/formal), timeliness etc. of the data affects the results considerably. For example, querying an integrated data set for a particular flight time, the time from the source with the higher update frequency can be chosen. Thus, query answering can be increased in effectiveness and efficiency using data quality criteria as a leverage to filter the more relevant results.

Acknowledgement

We thank the authors, whose submissions made this special issue possible and showed how active and exciting this field is; the reviewers who made the effort to read and offer constructive comments, enabling authors to see their work from a different perspective; and the IJSWIS editorial board for their guidance in the overall process.

References

1. Z. Abedjan, T. Grütze, A. Jentzsch, & F. Naumann (2014). *Profiling and mining RDF data with ProLOD++*. In IEEE 30th International Conference on Data Engineering (ICDE), (pp 1198–1201).

2. M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, S. Auer, and J. Lehmann (2013). *Crowdsourcing linked data quality assessment*. In H. Alani et al. (Eds.) 12th International Semantic Web Conference (ISWC), 21-25 October 2013, Sydney, Australia, (pp 260–276). Springer.

3. C. Batini & M. Scannapieco (2006). *Data Quality: Concepts, Methodologies and Techniques*. New York, Inc., Springer-Verlag.

4. E. Bertino, A. Maurino, & M. Scannapieco (2010). *Guest Editors' Introduction: Data quality in the Internet era*. In IEEE Internet Computing, 14(4) (pp11–13) 2010.

5. C. Bizer (2007). *Quality-Driven Information Filtering in the Context of Web-Based Information* Systems. PhD thesis, Freie Universität Berlin.

6. C. Bizer and R. Cyganiak (2009). Quality-driven information filtering using the WIQA policy framewor*k. Web Semantics*, 7(1), 1 – 10.

7. M. Bovee, R. P. Srivastava, & B. Mak (2003). A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems*, 18(1), 51–74.

8. A. Flemming (2010). *Quality characteristics of linked data publishing data sources*. Master's thesis, Humboldt-Universität of Berlin.

9. J. Gleick (1988). *Chaos: Making a New Science*. New York: Penguin Books.

10. C. Guéret, P. T. Groth, C. Stadler & J. Lehmann (2012). *Assessing linked data mappings using network measures*. In E. Simperl, P. Cimiano, A. Polleres, O. Corcho, V. Presutti (Eds.) ESWC, Vol. 7295 (pp 87–102). Springer.

11. H. Halpin, I. Herman, & J. Hayes (2010). *When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web*. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, B. Glimm (Eds.) ESWC Vol. 18(1) (pp 51–74). Springer.

12. O. Hartig & J. Zhao (2010). P*ublishing and consuming provenance metadata on the web of Linked Data*. In D. L. McGuinness, J. R. Michaelis, L. Moreau (Eds.) Third International Provenance and Annotation Workshop, 6378, (pp 78–9). Springer.

13. M. Jarke, M. Lenzerini, Y. Vassiliou, & P. Vassiliadis (2010). *Fundamentals of Data Warehouses*. Springer Publishing Company, 2nd edition.

14. S. Khatchadourian & M. P. Consens (2010). *ExpLOD: Summary-Based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud.* In The Semantic Web: Research and Applications, 6089, ESWC, (pp 272–287).

15. C. Bizer, P.N. Mendes & H. Mühleisen (2012). *Sieve: Linked data quality assessment and fusion*. In LWDM at EDBT (pp 116 – 123). ACM, New York.

16. F. Naumann (2002). *Quality-Driven Query Answering for Integrated Information Systems*. LNCS. Springer.

17. N. Prat & S. E. Madnick (2008). *Measuring data believability: A provenance approach.* Hawaii International Conference on System Sciences 0, 393.

18. T. C. Redman. *Data Quality for the Information Age*. Artech House, 1st edition, 1997.

19. A. Rotolo, S. Villata & F. Gandon (2013). *A deontic logic semantics for licenses composition in the web of data*. In ICAIL, (pp 111–120). ACM.

20. A. Rula, M. Palmonari, A. Harth, S. Stadtmüller & A. Maurino (2012). On the diversity and availability of temporal information in linked open data. In Cudré-mauroux et al. (Eds.) ISWC, 7649, (pp 492–507). Springer.

21. D. Vrandecic (2010). *Ontology Evaluation.* PhD thesis, Karlsruher Institute für Technologie (KIT).

22. Y. Wand & R. Y. Wang (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11) (pp86–95).

23. R. Y. Wang & D. M. Strong (1996(. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, 12(4) (pp 5–33).

24. A. Zaveri, D. Kontokostas, M. A. Sherif, L. Bühmann, M. Morsey, S. Auer & J. Lehmann (2013). *User-driven quality evaluation of DBpedia*. In M. Sabou et al. (Eds.) 9th International Conference on Semantic Systems, I-SEMANTICS (pp 97–104). ACM.

25. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer (2014). Quality assessment methodologies for Linked Data: A Survey. *Semantic Web Journal* (Under Review). http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-data-survey.