Linked Data Quality: Identifying and Tackling the Key Challenges

Magnus Knuth
Hasso Plattner Institute,
University of Potsdam
Potsdam, Germany
magnus.knuth@hpi.de

Dimitris Kontokostas
AKSW, University of Leipzig
Leipzig, Germany
kontokostas@informatik.uni-leipzig.de

Harald Sack
Hasso Plattner Institute,
University of Potsdam
Potsdam, Germany
harald.sack@hpi.de

ABSTRACT

The awareness of quality issues in Linked Data is constantly rising as new datasets and applications that consume Linked Data are emerging. In this paper we summarize key problems of Linked Data quality that data consumers are facing and propose approaches to tackle these problems. The majority of challenges presented here have been collected in a Lightning Talk Session at the *First Workshop on Linked Data Quality (LDQ2014)*.

Keywords

linked data, data quality, RDF

1. INTRODUCTION

Since the start of the Linking Open Data initiative, we have seen an unprecedented volume of structured data published on the web, in most cases as RDF and Linked (Open) Data. The quality of these datasets varies a lot and can hardly be better than the original data source from which the data has been created. Datasets may originate from crowdsourcing projects like Wikipedia and OpenStreetMap as well as from highly curated sources e. g. the library domain.

A consumer's perception of data quality is highly individual and strongly depends on the field of application. Therefore, *data quality* is often regarded as *fitness for use*, e. g. the DBpedia dataset might be appropriate for a simple end-user application but should not be used in critical applications such as the medical domain for treatment decisions. However, quality is a key to the success of the data web and a major barrier for further industry adoption.

In this paper we want to highlight contemporary quality problems that occur in Linked Data and that already are or need to be addressed in future. Likewise, we want to suggest solutions that have been developed in order to tackle these difficulties.

LDQ 2014, 1st Workshop on Linked Data Quality Sept. 2, 2014, Leipzig, Germany.

2. LINKED DATA QUALITY

Although *quality* is a commonly used term in Linked Data, it's definition is far from straightforward. The reason is that Linked Data quality can have different meaning for different people and in different contexts. During the *First Workshop on Linked Data Quality (LDQ2014)*[1], a discussion session was held where people from different backgrounds raised their personal thoughts on *Linked Data Quality*[2]. It was surprising to notice the variety of definitions and concerns that among others included stalled data, version management, changeset updates, RDF typo identification, and proper ontology modeling.

RDF validation is a core part of Linked Data quality but validation alone cannot solve the quality problem. Quality is fitness for use, thus a general methodology [11] is required to assess the results of a validation. The validation results, along with other factors and based on the application context can only provide a meaningful quality overview. Such a quality assessment methodology should be an integral part of the Linked Data life-cycle.

RDF version management is an additional quality issue that is not natively covered in the Semantic Web technology stack and can facilitate error provenance and tracking. The non-deterministic statement order and blank nodes make graph comparison equivalent to the graph isomorphism problem and thus, beyond polynomial time computation complexity[3].

Reusing popular vocabularies or manually creating a correct ontology model can also be seen as a general data quality issue. General purpose vocabularies such as `foaf`[4], `skos`[5], `schema.org`[6] or `dbpedia` ontology[7] usually reflect a swallow depiction of the real world. For many people for example, the `dbo:Actor` class is not correct since a profession is a role in a person's life and a person can have many different roles at different stages of his life, e. g. `student` or `spouse`. In the

[1] `http://ldq.semanticmultimedia.org/` co-located with *10th SEMANTiCS conference* on September 2nd, 2014 in Leipzig, Germany
[2] `http://tinyurl.com/LDQ14LightningTalks`
[3] `http://mathworld.wolfram.com/GraphIsomorphism.html`
[4] `http://xmlns.com/foaf/0.1/`
[5] `http://www.w3.org/2004/02/skos/core#`
[6] `http://schema.org/`
[7] `http://dbpedia.org/ontology/`, prefixed `dbo:`

end it depends on the granularity level one wants to reflect in his data but granularity usually comes at the cost of data integration.

3. TACKLING THE PROBLEM

Specific solutions for tackling the problem of Linked Data Quality as a whole are currently far from reality. Nevertheless, in the following subsection we provide an overview of existing work and possible future directions to cope with Linked Data Quality.

3.1 Linked Data validation and Quality assessment

Validation is a core part in the quality assessment of Linked Data. Although RDF exists already for many years there exists no official standard for Linked Data validation at the time of writing and a W3C working group has just been formed to define one[8]. Existing Linked Data application can either rely on ad-hoc options or use independently defined solutions such as: RDF Data Shapes [2], SPIN [6], SWRL [12], Dublin Core Profiles [10], RDFUnit [8] or `OWL` in `CWA` and a weaker form of `UNA`.

However, validation alone cannot be adequate. A general assessment methodology has to be built around validation that can interpret the validation results and assess the quality of the data. Rula et al. [11] propose a general three-phase and six-step methodology for assessing the quality of Linked Data involving manual, semi-automatic, and automated step in the process. On top of an assessment methodology, different applications can be built that automatically evaluate the quality of a dataset and provide automatic quality overviews or quality certifications [9].

3.2 Linked Data Cleansing

Data curation in general is a costly process for the publisher. The distributed nature of Linked (Open) Data may demand the involvement of multiple data providers to achieve satisfactory results. On the other hand those, who suffer from low data quality and most often identify these issues, are the data consumers. Unfortunately, only in some cases they provide feedback to create awareness of particular problems. A typical approach for Linked Data consumers is to duplicate a dataset and fix relevant problems within the local copy. These efforts are rarely communicated and hence not imitated on the original data so that other consumers could benefit equally. The *Patch Request vocabulary* [7] provides a standardized way to communicate change requests to data publishers and other consumers of a dataset. Additionally, Embury et al. [1] examine the feasibility of identifying data corrections in revisioned datasets that can than be applied to copies of that dataset.

The correction of errors in Linked Data should be distributed onto the shoulders of many and possibilities to distribute such changes should be researched.

3.3 Best Practices for Linked Data Creation and Reuse

Linked Data is primarily made for machine interpretation and therefore it needs to comply the technical standards. Typical RDF parser implementations do not cope with –even minor– syntactical errors. Many publishers create RDF data using scripts or perform changes manually. Such modalities raise the risk of introducing syntactical errors, which can be avoided by using RDF tools and programming libraries, such as the Redland RDF API[9] and Apache Jena[10]. Optionally, generated RDF data should be checked by subsequent syntax validation prior to publication with appropriate tools, such as the Raptor RDF parser utility[11] and Apache Jena CLI tools[12].

Hogan et al. [5] name prominent problems with publicly available RDF datasets and survey general RDF validation tools. Heath et al. [4] summarize best practices for publishing Linked Data.

Beyond validating the syntax of their RDF serialization, data providers should also keep an eye on the correct usage of vocabularies. *RDFUnit* [8] checks for proper vocabulary utilisation by creating tests in form of SPARQL queries from the vocabulary specification. These tests are created automatically and can be executed also on large scale datasets that provide a SPARQL endpoint, as DBpedia.

To ensure that entities within a dataset are described in a form that is required for usage by a particular application, such structures can be defined with *RDF Data Shapes* as it has been done to the WebIndex data portal [2]. *RDF Data Shapes* allow to express the expected structure of data, e. g. that a person entity has an `xsd:string` connected by the property `foaf:name`. Such shape templates can be used as a contract between data publisher and data consumer in order to guarantee that an application can digest the given data properly.

3.4 Versioning Linked Data

Continuous updates to and curation of datasets raises the aim for tracking changes in Linked Data. There is rare support in Linked Data publishing for versioning as well as for provenance of changes. *Apache Marmotta*[13] is one Linked Data publishing platform that supports versioning. *R43ples* [3] provides versioning for any triplestore implementation, it acts as a proxy SPARQL endpoint that allows to refer to prior revisions by extending the SPARQL query language while standard SPARQL queries always work transparently on the master revision.

4. CONCLUSION

In this paper we have discussed key issues of Linked Data quality aspects and possible ways to tackle them. We see quality as a core –and grey– component of the semantic web stack that, if addressed correctly and systematically, will enable further adoption.

[8] `http://www.w3.org/blog/data/2014/09/30/data-shapes-working-group-launched/`

[9] `http://librdf.org/`

[10] `https://jena.apache.org/`

[11] `http://librdf.org/raptor/rapper.html`

[12] `https://jena.apache.org/documentation/io/#command-line-tools`

[13] `http://marmotta.apache.org/`

Acknowledgements

We would like to thank all attendees of the *1st Linked Data Quality Workshop (LDQ2014)*, especially the people contributing to the Lightning Talks Session. These discussions delivered valuable input and witnessed that this topic is of high importance to researchers and practitioners working with Linked Data. In addition we would also like to thank all reviewers that have helped us and the authors with their valuable comments and feedback.

Program Committee

- Maribel Acosta – Karlsruhe Institute of Technology, AIFB, Germany
- Volha Bryl – University of Mannheim, Germany
- Ioannis Chrysakis – ICS FORTH, Greece
- Stefan Dietze – L3S, Germany
- Marco Fossati – SpazioDati, Italy
- Fumihiro Kato – Kyushu University, Japan
- Christoph Lange – University of Bonn, Fraunhofer IAIS, Germany
- Maristella Matera – Politecnico di Milano, Italy
- Felix Naumann – Hasso Plattner Institute, Germany
- Matteo Palmonari – University of Milan-Bicocca, Italy
- Adrian Paschke – Free University of Berlin, Germany
- Heiko Paulheim – University of Mannheim, Germany
- Mariano Rico – Universidad Politécnica de Madrid, Spain
- Anisa Rula – Università di Milano-Bicocca, Italy
- Elena Simperl – University of Southampton, United Kingdom
- Patrick Westphal – AKSW, University of Leipzig, Germany
- Amrapali Zaveri – AKSW, University of Leipzig, Germany
- Jun Zhao – Lancaster University, United Kingdom
- Antoine Zimmermann – ISCOD / LSTI – École Nationale Supérieure des Mines de Saint-Étienne, France

Lightning Talk presenters

- Behshid Behkamal – Ferdowsi University of Mashhad, Iran
- Jeremy Debattista – University of Bonn, Germany
- Riccardo Del Gratta – Istituto di linguistica Computazionale CNR, Italy
- Nidhi Kushwaha – Ferdowsi University of Mashhad, Iran
- Gerard Kuys – Ordina NV, Netherlands
- Peter Vandenabeele – self-employed, Belgium
- Lieke Verhelst – Linked Data Factory, Netherlands
- Amrapali Zaveri – AKSW, University of Leipzig, Germany

5. REFERENCES

[1] S. Embury, B. Jin, S. Sampaio, and I. Eleftheriou. On the feasibility of crawling linked data sets for reusable defect corrections. In *Proceedings of the 1st Workshop on Linked Data Quality (LDQ2014)*, volume 1215. CEUR Workshop Proceedings, 2014.

[2] J. E. L. Gayo, E. Prud'Hommeaux, H. Solbrig, and J. M. A. Rodriguez. Validating and describing linked data portals using RDF shape expressions. In *Proceedings of the 1st Workshop on Linked Data Quality (LDQ2014)*, volume 1215. CEUR Workshop Proceedings, 2014.

[3] M. Graube, S. Hensel, and L. Urbas. R43ples: Revisions for triples - an approach for version control in the semantic web. In *Proceedings of the 1st Workshop on Linked Data Quality (LDQ2014)*, volume 1215. CEUR Workshop Proceedings, 2014.

[4] T. Heath and C. Bizer. *Linked Data: Evolving the web into a global data space*. Morgan & Claypool, 2011.

[5] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. In *Linked Data on the Web Workshop (LDOW 2010) at WWW 2010*, volume 628, pages 30–34. CEUR Workshop Proceedings, 2010.

[6] H. Knublauch, J. A. Hendler, and K. Idehen. SPIN - overview and motivation. W3C Member Submission, W3C, February 2011.

[7] M. Knuth, J. Hercher, and H. Sack. Collaboratively patching linked data. In *Proceedings of 2nd International Workshop on Usage Analysis and the Web of Data (USEWOD 2012), co-located with the 21st International World Wide Web Conference 2012 (WWW 2012)*, Lyon, France, April 2012.

[8] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. J. Zaveri. Test-driven evaluation of linked data quality. In *Proceedings of the 23rd International Conference on World Wide Web*, 2014.

[9] J. P. Mccrae, C. Wiljes, and P. Cimiano. Towards assured data quality and validation by data certification. In *Proceedings of the 1st Workshop on Linked Data Quality (LDQ2014)*, volume 1215. CEUR Workshop Proceedings, 2014.

[10] M. Nilsson. Description set profiles: a constraint language for dublin core application profiles, 2008.

[11] A. Rula and A. Zaveri. Methodology for assessment of linked data quality: A framework. In *Proceedings of the 1st Workshop on Linked Data Quality (LDQ2014)*, volume 1215. CEUR Workshop Proceedings, 2014.

[12] World Wide Web Consortium (W3C). SWRL: A semantic web rule language combining OWL and RuleML, 2004.