MEPDaW 2017 and LDQ 2017 Preface⋆

Jeremy Debattista[1], Javier D. Fernández[2], Jürgen Umbrich[2], Anisa Rula[3], Amrapali Zaveri[4], Anastasia Dimou[5], and Wouter Beek[6]

[1] University of Bonn and Fraunhofer IAIS, Bonn, Germany
`debattis@cs.uni-bonn.de`
[2] Vienna University of Economics and Business, Vienna, Austria
`{javier.fernandez,juergen.umbrich}@wu.ac.at`
[3] University of Milano-Bicocca, Milan, Italy
`rula@disco.unimib.it`
[4] Maastricht University, The Netherlands
`amrapali.zaveri@maastrichtuniversity.nl`
[5] Ghent University - imec, Belgium
`anastasia.dimou@ugent.be`
[6] VU Amsterdam, The Netherlands
`w.g.j.beek@vu.nl`

Abstract. This joint volume of proceedings gathers together papers from the 3rd Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW2017) and the 4th Workshop on Linked Data Quality (LDQ2017), held on the 28th and 29th of May of 2017 during the 14th ESWC conference in Portorož, Slovenia.

1 Managing the Evolution and Preservation of the Data Web

There is a vast and rapidly increasing quantity of scientific, corporate, government, and crowd-sourced data published on the emerging Data Web. Open Data are expected to play a catalyst role in the way structured information is exploited on a large scale. This offers a great potential for building innovative products and services that create new value from already collected data. It is expected to foster active citizenship (e.g., around the topics of journalism, greenhouse gas emissions, food supply-chains, smart mobility, etc.) and world-wide research according to the "fourth paradigm of science".

Published datasets are openly available on the Web. A traditional view of digitally preserving them by *pickling them and locking them away* for future use, like groceries, conflicts with their evolution. There are a number of approaches and frameworks, such as the Linked Data Stack, that manage a full life-cycle of the Data Web. More specifically, these techniques are expected to tackle major issues such as the synchronisation problem (how to monitor changes), the curation problem (how to repair data imperfections), the appraisal problem (how to assess the quality of a dataset), the citation problem (how to cite a particular version of a linked dataset), the archiving problem (how to retrieve the most recent or a particular version of a dataset), and the sustainability problem (how to support preservation at scale, ensuring long-term access).

⋆ Joint proceedings are publicly available in [3].

Preserving linked open datasets poses a number of challenges, mainly related to the nature of the Linked Data principles and the RDF data model. Since resources are globally interlinked, effective citation measures are required. Another challenge is to determine the consequences that changes to one LOD dataset may have to other datasets linked to it. The distributed nature of LOD datasets furthermore introduces additional complexity, since external sources that are being linked to may change or become unavailable. Finally, another challenge is to identify means to continuously assess the quality of dynamic datasets.

During last year's workshop [2], a number of open research questions were raised during the keynote and discussions:

1. How can we represent archives of continuously evolving linked datasets? (efficiency vs. compact representation)
2. How can we measure the performance of systems for archiving evolving datasets, in terms of representation, efficiency and compactness?
3. How can we improve completeness of archiving?
4. How can emerging retrieval demands in archiving (e.g. time-traversing and traceability) be satisfied? What type of data analytics can we perform on top of the archived Web of data?
5. How can certain time-specific queries over archives be answered? Can we re-use existing technologies (e.g. SPARQL or temporal extensions)? What is the right query language for such queries?
6. Is there an actual and urgent need in the community for handling the dynamicity of the Data Web?
7. Is there the need of a killer-app to kick start the management of the evolving Web of Data?

Last year's workshop discussions and papers were discussed in a SIGIR Forum report [1]. This year's workshop will showcase 6 papers, split into two main sessions: (1) Managing and Querying Evolving Data; and (2) Computing and Exploiting Changes in Evolving Data. These papers address most of the questions raised in last workshop. Furthermore, in this workshop, Prof. Dr. Maria-Esther Vidal keynote discusses challenges of Semantic data management in Big Data.

2   Linked Data Quality

The 4th Linked Data Quality Workshop[7] focuses on novel methodologies and frameworks for assessing, monitoring, maintaining, and improving the quality of Linked Data as well as to highlight tools and user interfaces which can effectively assist in its assessment and repair. In addition, the workshop seeks methodologies that help to identify the current impediments in building real-world Linked Data applications leveraging data and ontology quality, and use cases that reveal success stories or aspects that have been neglected so far. The benefits of addressing Linked Data quality issues will not only help in detecting inherent data quality problems currently plaguing Linked Data, but also provide the means to fix these problems and maintain the quality in the long run.

---

[7] `ldq.semanticmultimedia.org`

In this year's contributions we see a focus on quality assessment and validation *services*, rather than client-side solutions. Since Software-as-a-Service (SaaS) has known benefits, such as reduced consumer cost and increased ease of installation and use, the rise of Quality-Assessment-as-a-Service is promising.

Linked Data validation has been difficult so far because Linked Data schemas are not used as constraints, but for deriving new facts (i.e., entailment). The ongoing standardization of Linked Data validation languages such as SHACL[8] and ShEX[9] provides new opportunities for automating Linked Data quality assessment. It is promising to see that these standardization efforts have already resulted in novel Linked Data Quality approaches. Finally, the recent publication of a Linked Data Quality vocabulary by W3C[10] makes it possible to represent and disseminate the results of quality assessment as Linked Open Data, which opens up new approaches as well.

This year we accepted three papers and invited a keynote speaker, which we describe in brief. Mihindukulasooriya et al. [6] present *Loupe*, a data profiling service for Linked Data. Data profiling is a common approach for assessing Data Quality in relational databases, but has not yet been applied to Linked Data. Loupe builds on recent standardization efforts for Linked Data validation such as SHACL and ShEX.

Mc Gurk et al. [5] presents a systematic overview of existing ontology and Linked Data quality metrics, by categorizing them according to data quality standards established by ISO. Building on the quality assessment framework *Luzzu* and the ontology visualization library *VOWL*, they present a new approach for visualizing the extent to which the identified quality metrics apply to a given ontology.

Hashimoto et al. [4] focuses on the use of Linked Data ontologies in order to automatically detect and resolve conflicts in a manufacturing design process. This poses challenges for the data, which must be of sufficient quality in order to reliably model the design process, but also provides opportunities when conflicts can be detected and mitigated at an early stage.

Péter Király's keynote discusses how metadata quality was performed in the Europeana use case. His talk shows the process of metadata quality assurance in big digital libraries, such as Europeana, the findings of the functional requirement analyses of Europeana records, the data quality analyzing framework built, as well as the general and specific metrics considered and the scalability issues raised.

Acknowledgments

---

[8] https://www.w3.org/TR/shacl/
[9] http://shex.io/
[10] https://www.w3.org/TR/vocab-dqv/

References

1. J. Debattista, J. D. Fernández, and J. Umbrich. Report on the 2nd workshop on managing the evolution and preservation of the data web (mepdaw 2016). *SIGIR Forum*, 50(2):82–88, Feb. 2017.

2. J. Debattista, J. D. F. García, M. Knuth, D. Kontokostas, A. Rula, J. Umbrich, and A. Zaveri, editors. *Joint proceedings of the 2nd Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW 2016) and the 3rd Workshop on Linked Data Quality (LDQ 2016)*, number 1585 in CEUR Workshop Proceedings, Aachen, May 2016.

3. J. Debattista, J. D. F. García, J. Umbrich, A. Rula, A. Zaveri, A. Dimou, and W. Beek, editors. *Joint proceedings of the 3rd Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW 2017) and the 4th Workshop on Linked Data Quality (LDQ 2017)*, number 1824 in CEUR Workshop Proceedings, Aachen, May 2017.

4. K. Hashimoto, Y. Yamane, S. Suzuki, M. Takaai, M. Watanabe, and H. Umemoto. Towards Ontology Quality Assessment. In *4th Workshop on Linked Data Quality (LDQ2017)*.

5. S. Mc Gurk, C. Abela, and J. Debattista. Towards Ontology Quality Assessment. In *4th Workshop on Linked Data Quality (LDQ2017)*, 2017.

6. N. Mihindukulasooriya, R. García-Castro, F. Priyatna, E. Ruckhaus, and N. Saturno. A Linked Data Profiling Service for Quality Assessment. In *4th Workshop on Linked Data Quality (LDQ2017)*, 2017.