# MEPDaW+LDQ Preface*

Jeremy Debattista[1], Javier D. Fernández[2], Magnus Knuth[3], Dimitris Kontokostas, Anisa Rula[5], Jürgen Umbrich[2], and Amrapali Zaveri[4]

[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] Hasso Plattner Institute, University of Potsdam, Germany
magnus.knuth@hpi.uni-potsdam.de
[4] Stanford Center for Biomedical Informatics Research, Stanford University, USA
amrapali@stanford.edu
[5] University of Milano-Bicocca, Milan, Italy
rula@disco.unimib.it

**Abstract.** This joint volume of proceedings gathers together papers from the 2nd Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) and the 3rd Workshop on Linked Data Quality (LDQ), held on the 30th of May of 2016 during the 13th ESWC conference in Anissaras, Crete, Greece.

## 1 Managing the Evolution and Preservation of the Data Web

This workshop targeted one of the emerging and fundamental problems in the Semantic Web, specifically the preservation of evolving linked datasets. 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 in the 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". The most noteworthy advantage of the Data Web is that, rather than documents, facts are recorded, which become the basis for discovering new knowledge that is not contained in any individual source, and solving problems that were not originally anticipated. In particular, Open Data published according to the Linked Data Paradigm are essentially transforming the Web into a vibrant information ecosystem.

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, would conflict with their evolution. There are a number of approaches and frameworks, such as the LOD2 stack, that manage a full life-cycle of the Data Web.

* MEPDaW+LDQ join proceedings are publicly available in [1].

More specifically, these techniques are expected to tackle major issues such as the synchronisation problem (how can we monitor changes), the curation problem (how can data imperfections be repaired), the appraisal problem (how can we assess the quality of a dataset), the citation problem (how can we cite a particular version of a linked dataset), the archiving problem (how can we retrieve the most recent or a particular version of a dataset), and the sustainability problem (how can we spread preservation ensuring long-term access).

Preserving linked open datasets poses a number of challenges, mainly related to the nature of the LOD principles and the RDF data model. In LOD, datasets representing real-world entities are structured; thus, when managing and representing facts we need to take into consideration possible constraints that may hold. Since resources might be interlinked, effective citation measures are required to be in place to enable, for example, the ranking of datasets according to their measured quality. 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 makes archiving a headache.

This workshop aimed at addressing the above mentioned challenges and issues by providing a forum for researchers and practitioners who apply linked data technologies to discuss, exchange and disseminate their work.

The workshop included an inspiring talk by Dr. Axel Polleres on *Archiving Linked and Open Data*, three research papers and one industry paper, and a plenary discussion. Based on the review scores, the best paper award has been given to Ruben Taelman, Ruben Verborgh, Pieter Colpaert, Erik Mannens and Rik Van de Walle for their work *"Continuously Updating Query Results over Real-Time Linked Data"*.

## 2 Linked Data Quality

The focus of this workshop was to reveal novel methodologies and frameworks in assessing, monitoring, maintaining, and improving the quality of Linked Data (LD) as well as introduce tools and user interfaces which can effectively assist in the assessment and repair. In addition, the workshop sought methodologies that help to identify the current impediments in building real-world Linked Data applications leveraging data quality. The benefits of addressing Linked Data quality issues would 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.

To guarantee the full exploitation of the published or consumed Linked Data, it is important to assure LD quality. In this way, it is possible to understand whether data is appropriate for the task at hand before using it. There are several issues in LD which hampers the use of datasets in building real-world LD-based applications and research solutions. One of issues is the method to find the most relevant LD data for a particular application. Also, generating meaningful associations between the LD datasets, at the ontology, data or property level is an important issue to be considered when building such applications. Essentially, the quality of LD is a deciding factor as to which datasets can be used in building such real-world applications. Currently, there is no full-proof method of performing this kind of quality assessment.

In general detecting the quality of datasets available and making this information explicit is a challenge. This entails the (semi-)automatic identification of existing problems, which is either insensitive to the use case or is limited in identifying only specific (objective) quality issues. In LD few efforts are currently available to standardize how data quality tracking and assurance should be implemented and it poses yet other challenges: (i) LD refers to a Web-scale knowledge base consisting of interlinked published data from a multitude of autonomous information providers (variety of data). The quality of provided information may depend on the intention of the information provider; (ii) the increasing diffusion of the LD paradigm as a standard way to share knowledge on the Web allows consumers to fully exploit vast amount of data that were not available in the past (high volume of data). We are likely to find more low quality in LD than in smaller data sets because in large data sets data are produced with automatic information processes which are often error prone; (iii) data sets in LD formats may often be used by third-party applications in ways not expected by the original creators of the data set; (iv) LD provides data integration through interlinking of data between heterogeneous data sources. The quality of integrated data will depend on the quality of original data sources; (v) last but not least relevant, Linked Data can be considered as a dynamic environment where information can change rapidly and cannot be assumed to be static (velocity of data). Changes in LD sources should reflect changes in the real world, otherwise data can soon become out-dated. Out-of-date information can reflect data inaccuracy problems and can deliver invalid information. For example, more up-to-date information should Data be preferred over less up-to-date information in data integration and fusion applications. Moreover, none of the current approaches use the assessment to ultimately improve the quality of the underlying dataset.

The workshop included a keynote by Christian Dierschl on *Data quality assurance in data-intensive systems*, three paper presentations and a lightning talk session with following discussions. Based on the review scores, the best paper award has been given to Tomáš Knap for his work on *Increasing Quality of Austrian Open Data by Linking them to Linked Data Sources: Lessons Learned* [2].

## 3 Organisation

### 3.1 Organising Committees

- **MEPDaW**
  - Jeremy Debattista, Enterprise Information Systems, University of Bonn, Germany / Organized Knowledge, Fraunhofer IAIS, Germany
  - Jürgen Umbrich, Vienna University of Economics and Business, Austria
  - Javier D. Fernández, Vienna University of Economics and Business, Austria
- **LDQ**
  - Anisa Rula – University of Milano-Bicocca, Italy
  - Amrapali Zaveri – Stanford University, United States
  - Magnus Knuth – Hasso Plattner Institute, University of Potsdam, Germany
  - Dimitris Kontokostas – AKSW, University of Leipzig, Germany

### 3.2 Program Committees

- **MEPDaW**
  - Judie Attard, University of Bonn/Fraunhofer IAIS, Germany
  - Ioannis Chrysakis, FORTH-ICS, Greece
  - Keith Cortis, University of Passau, Germany
  - Giorgos Flouris, FORTH-ICS, Greece
  - Marios Meimaris, ATHENA R.C., Greece
  - Fabrizio Orlandi, University of Bonn/Fraunhofer IAIS, Germany
  - Fouad Zablith, American University of Beirut, Lebanon
  - Magnus Knuth, Hasso Plattner Institute – University of Potsdam, Germany
  - Anisa Rula, University of Milano-Bicocca, Italy
  - Wouter Beek, VU University Amsterdam, Netherlands
  - Yannis Stavrakas, ATHENA R.C., Greece
  - Amrapali J. Zaveri, Dumontier Lab - Stanford University, USA
  - Mathieu d'Aquin, The Open University, United Kingdom
  - Yannis Roussakis, FORTH-ICS, Greece
  - Kemele M. Endris, University of Bonn
  - Charlie Abela, University of Malta, Msida, Malta
  - George Papastefanatos, ATHENA R.C., Greece
  - Nandana Mihindukulasooriya, Universidad Politécnica de Madrid (UPM), Spain
  - Niklas Petersen, University of Bonn/Fraunhofer IAIS, Germany
  - Joseph Bonello, University of Malta, Msida, Malta
- **LDQ**
  - Maribel Acosta, Karlsruhe Institute of Technology – AIFB, Germany
  - James Anderson, Datagraph, United States
  - Volha Bryl, Springer Science+Business Media, Germany
  - Ioannis Chrysakis, ICS FORTH, Greece
  - Mathieu d'Aquin, Knowledge Media Institute, The Open University, United Kingdom
  - Jeremy Debattista, University of Bonn, Fraunhofer IAIS, Germany
  - Anastasia Dimou, MultimediaLab, Ghent University – iMinds, Belgium
  - Suzanne Embury – University of Manchester, United Kingdom
  - Christian Fürber, Information Quality Institute GmbH, Germany
  - Jose Emilio Labra Gayo, University of Oviedo, Spain
  - Markus Graube, Technische Universität Dresden, Germany
  - Tom Heath, The Open Data Institute, United Kingdom
  - Tomáš Knap, Semantic Web Company, AT, and Charles University in Prague, Czech Republic
  - Maristella Matera, Politecnico di Milano, Italy
  - John McCrae, CITEC, University of Bielefeld, Germany
  - Matteo Palmonari, University of Milan-Bicocca, Italy
  - Heiko Paulheim, University of Mannheim, Germany
  - Mariano Rico, Universidad Politécnica de Madrid, Spain
  - Patrick Westphal, AKSW, University of Leipzig, Germany
  - Antoine Zimmermann, École Nationale Supérieure des Mines de Saint-Étienne, France

## Acknowledgements

We would like to thank the authors for their contribution and active participation in the workshops, and all the program committee members for reviewing the submissions and provide valuable feedback. We are also grateful to the organisers of the ESWC 2016 conference for their support, and our keynote speakers, Axel Polleres from the Vienna University of Economics and Business (Austria), and Christian Dirschl from Wolters Kluwer (Germany).

The MEPDaW workshop was co-organised by members funded by the Austrian Science Fund (FWF): M1720-G11.

The LDQ workshop was co-organised by members funded by the German Government, Federal Ministry of Education and Research under the project: 03WKCJ4D.

## References

1. 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.
2. T. Knap. Increasing quality of austrian open data by linking them to linked data sources: Lessons learned. In Debattista et al. [1].