Preface of MEPDaW 2020: Managing the Evolution and Preservation of the Data Web

Fabrizio Orlandi[1] ⓘ, Damien Graux[1] ⓘ, Maria-Esther Vidal[2] ⓘ,
Javier D. Fernández[3] ⓘ, and Jeremy Debattista[4] ⓘ

[1] ADAPT SFI Centre, Trinity College Dublin, Ireland. `{orlandif,grauxd}@tcd.ie`
[2] Technische Informationsbibliothek (TIB), Germany. `mvidal@umiacs.umd.edu`
[3] F. Hoffmann-La Roche AG, Switzerland. `javier_d.fernandez@roche.com`
[4] TopQuadrant Inc. `jdebattista@topquadrant.com`

Abstract. The MEPDaW workshop series targets one of the emerging and fundamental problems of the Web, specifically the management and preservation of evolving knowledge graphs. During the past six years, the workshop series has been gathering a community of researchers and practitioners around these challenges. To date, the series has successfully published more than 30 articles allowing more than 50 individual authors to present and share their ideas.
This 6th edition, virtually co-located with the International Semantic Web Conference (ISWC 2020), gathered the community around nine research publications and one invited keynote presentation. The event took place online on the 1st of November, 2020.

Keywords: Web Data evolution · Data preservation, provenance and lineage · Temporal & Evolving Knowledge Graphs · RDF archiving and versioning

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 openly published on the Web. Open Data plays a catalyst role in the way structured information is exploited on a large scale. A traditional view of digitally preserving these datasets by "pickling and locking them away" for future use, like groceries, conflicts with their evolution. There are several approaches and frameworks (*e.g.* Linked Data Stack [10], PoolParty Suite[1], Metaphactory[2], etc.) targeted at managing the life-cycle of the Data Web. More specifically, these solutions are expected to tackle major issues such as the synchronisation problem (monitoring changes) [12,16], the curation problem (repairing data imperfections) [14], the appraisal problem (assessing the quality of a dataset) [11], the citation problem (how to cite a particular version of a dataset) [4], the archiving problem (retrieving a specific version of a

---

[1] `https://semantic-web.com/poolparty-semantic-suite/`

[2] `https://metaphacts.com/`

dataset) [13,15], and the sustainability problem (preserving at scale, ensuring long-term access) [4].

The **sixth** edition of this workshop was organised for the first time at the International Semantic Web Conference (ISWC) and followed the structure of the previous editions. We invited a number of experts in the field of Linked Data and Data Evolution & Preservation in order to suggest and advise on the different topics that our workshop covered this year. This year, at ISWC 2020, we successfully gathered more than 60 participants for our half-day event. In line with most academic events, this year MEPDaW was held as a virtual event and we had to re-think the interactions between participants.

MEPDaW Scientific programme

The workshop started with the keynote entitled "Sharing, Tracking, and Enhancing Highly Dynamic Knowledge Graphs" given by Prof. Philippe Cudré-Mauroux from eXascale Infolab[3] (University of Fribourg, Switzerland). He described how Knowledge Graphs are in practice highly dynamic and incomplete and the challenges that this entails. In particular, Prof. Cudré-Mauroux gave an overview of some of the recent techniques they developed in his lab to improve the automated processing of large-scale and evolving Knowledge Graphs. First, he described data-driven techniques to identify information gaps in Knowledge Graphs (*e.g.*, in terms of missing classes or properties). Then, he presented a series of methods to impute missing values from the graphs. He eventually gave insights on two of their funded projects and large-scale system deployments: one for Swiss open research data, and one for knowledge tracking on Microsoft Azure. Overall, this keynote gave the audience in-depth details on practical (and industrial) use cases backed by cutting-edge research techniques.

The first article presented dealt with an approach to detect and assess the semantic drift among timely-distinct versions of an ontology [1]. It was followed by [4] which proposes the implementation of a global, persistent identifier system built upon time-based immutable resource revisioning of generic HTTP resources, as identified by their URL and resolved via time-based HTTP content negotiation, building upon existing Web standards.

Following the keynote, three research efforts dedicated to specific use-cases and systems were presented. Both Jamal A. Nasir [7] and André Regino [8] focused on Linked Open Data related challenges. The former presented *i*LOD: a decentralised file system dedicated to the LOD cloud [7]. The latter described a new strategy to discover semantically broken links within LOD datasets [8]. In [9], the authors tackled the representation of scientific literature evolution across time. As new research is conducted, knowledge evolves, getting documented in dissertations, theses and articles. They presented new methods exploiting Temporal Knowledge Graphs (TKGs) to model knowledge evolution in corpora of unstructured texts.

---

[3] `https://exascale.info/`

In addition, MEPDaW 2020 gathered three publications addressing challenges related to the RDF standard. First, Cuevas & Hogan [2] explored solutions for querying archives of versioned RDF data using SPARQL and off-the-shelf engines; in particular they considered multiple representations of RDF archives, and described how input queries can be automatically rewritten to return solutions for a particular version (or solutions that change between versions). Second, in [6], the authors presented a framework for an automatic adaptation of RDF-based semantic annotations when RDF graphs are modified. Third, Gleim *et al.* [5] focused on provenance tracking and presented a concrete alignment of all roles and relations in the FactDAG model to the W3C PROV provenance standard, allowing future software implementations to directly produce standard-compliant provenance information.

Finally, wrapping up the article sessions and starting the open-discussion, Lars Gleim and Stefan Decker [3] presented a review of open challenges for the management and preservation of evolving data on the Web.

Organizing Committee

– Fabrizio Orlandi, ADAPT SFI Centre, Trinity College Dublin, Ireland
– Damien Graux, ADAPT SFI Centre, Trinity College Dublin, Ireland
– Maria-Esther Vidal, TIB, Hannover, Germany
– Javier D. Fernández, F. Hoffmann-La Roche AG, Switzerland
– Jeremy Debattista, TopQuadrant Inc.

Advisory Board

– Laure Berti-Equille, IRD Marseille, France
– Declan O'Sullivan, ADAPT Centre, Trinity College Dublin, Ireland
– James Anderson, Dydra - Datagraph, USA
– Axel Polleres, Vienna University of Economics and Business, Austria

Programme Committee

– Natanael Arndt, Leipzig University, Germany
– Ioannis Chrysakis, FORTH-ICS, Greece & Ghent Uni. (imec), Belgium
– Diego Collarana, Fraunhofer IAIS, Germany
– Pieter Colpaert, Ghent University, Belgium
– Christophe Debruyne, Trinity College Dublin, Ireland
– Luis Ibanez-Gonzalez, University of Southampton, England
– Harshvardhan J. Pandit, ADAPT Centre - Trinity College Dublin, Ireland
– George Papastefanatos, IMIS / RC "Athena", Greece
– Giuseppe Pirrò, Sapienza University of Rome, Italy
– Julio Cesar dos Reis, University of Campinas, Brazil
– Ruben Taelman, Ghent University – imec, Belgium
– Brecht Van de Vyvere, Ghent University, Belgium

Acknowledgements

We would like to thank all the authors, reviewers, committee members and the invited speaker for their contributions, support and commitment during this particularly challenging year.

Articles presented at MEPDaW 2020

1. Capobianco, G., Cavaliere, D., Senatore, S.: Ontodrift: a semantic drift gauge for ontology evolution monitoring. In: Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2020)
2. Cuevas, I., Hogan, A.: Versioned Queries over RDF Archives: All You Need is SPARQL? In: Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2020)
3. Gleim, L., Decker, S.: Open Challenges for the management and preservation of evolving data on the Web. In: Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2020)
4. Gleim, L., Decker, S.: Timestamped URLs as persistent identifiers. In: Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2020)
5. Gleim, L., Tirpitz, L., Pennekamp, J., Decker, S.: Expressing FactDAG provenance with PROV-O. In: Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2020)
6. de Jesus Pontes Monteiro, E., dos Reis, J.C.: A framework for the automatic adaptation of RDF-based semantic annotations. In: Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2020)
7. Nasir, J.A., McCrae, J.P.: iLOD: Interplanetary file system based linked OpenData cloud. In: Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2020)
8. Regino, A.G., dos Reis, J.C.: Discovering semantically broken links in LOD datasets. In: Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2020)
9. Rossanez, A., dos Reis, J.C., da Silva Torres, R.: Representing scientific literature evolution via temporal knowledge graphs. In: Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2020)

References

10. Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., Lehmann, J., Martin, M., Mendes, P.N., Van Nuffelen, B., et al.: Managing the life-cycle of linked data with the LOD2 stack. In: International semantic Web conference. pp. 1–16. Springer (2012)
11. Debattista, J., Auer, S., Lange, C.: Luzzu—a methodology and framework for linked data quality assessment. J. Data and Information Quality **8**(1) (Oct 2016)
12. Endris, K.M., Faisal, S., Orlandi, F., Auer, S., Scerri, S.: Interest-based RDF update propagation. In: Proceedings of the 14th International Conference on The Semantic Web - ISWC 2015 - Volume 9366. p. 513–529. Springer-Verlag, Berlin, Heidelberg (2015)
13. Fernández, J.D., Polleres, A., Umbrich, J.: Towards efficient archiving of dynamic linked open data. In: MEPDaW workshop at ESWC'15 (2015)
14. Freitas, A., Curry, E.: Big data curation. In: New Horizons for a Data-Driven Economy (2016)
15. Pelgrin, O., Galárraga, L.: Towards fully-fledged archiving for RDF datasets. (Accepted with Minor Revisions), Semantic Web Journal (2020)
16. Tasnim, M., Collarana, D., Graux, D., Orlandi, F., Vidal, M.E.: Summarizing entity temporal evolution in knowledge graphs. In: Companion Proceedings of The 2019 World Wide Web Conference. p. 961–965. WWW '19, Association for Computing Machinery, New York, NY, USA (2019)