Report on the 2nd Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW 2016)

Jeremy Debattista
University of Bonn and Fraunhofer IAIS, Bonn, Germany
debattis@cs.uni-bonn.de

Javier D. Fernández, Jürgen Umbrich
Vienna University of Economics and Business, Vienna, Austria
{*javier.fernandez,juergen.umbrich*}*@wu.ac.at*

Abstract

Research on preserving evolving linked datasets is gaining increasingly attention in the Semantic Web community. The 2nd workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW 2016) aimed at addressing the numerous and diverse emerging challenges, from change discovery and scalable archive representations/indexes/infrastructures to time-based query languages and evolution analysis. In this report, we motivate our workshop and outline the keynote by Axel Polleres and the papers (3 full papers and one industry paper) presented at MEPDaW 2016 co-located with the ESWC 2016 conference in Anissaras, Crete, Greece. Finally, we conclude with an outlook of future directions.

1 Introduction

There is a vast and rapidly increasing quantity of scientific, corporate, government and crowd-sourced data published on the emerging Data Web. In fact, the last decade has seen an impressive growth of the use of RDF (*Resource Description Framework*) [14] as the de-facto graph-based data model to express facts about any field of knowledge on the Web. The main driving force has been the *Linked Open Data* community, which promotes to publicly share such structured data on the Web and to connect it by re-using HTTP Unique Resource Identifiers (URIs) between different data sources [3]. Besides accessible identifiers, at the most basic level, data providers serve RDF dumps in a wide range of formats. Additionally, data providers expose data through public query APIs with different expression power. On the one hand, services such as Linked Data Fragments [20] and other RESTful APIs mainly focus on traditional simple lookups. While, SQL-like structured queries can be written in the SPARQL query language [12] and served via SPARQL endpoints.

All this offers a great potential for building innovative products and services that create new value from already collected data. Open Data published according to the Linked Data

Paradigm are transforming the Web into a vibrant information ecosystem. However, recent studies [13] show that these publication schemes are subject to unnoticed changes and service disruptions, following the same scale-free nature of the Web. In other words, the emerging huge, heterogeneous and interconnected cloud of data-to-data hyperlinks evolves widely; without notice, data providers often add or delete data, thus creating new versions of the served datasets, and migrate or abandon APIs to obsolescence. As an example of this latter, SPARQLES [4], a service that monitors 553 SPARQL endpoints, shows that more than 55% of the endpoints are currently unavailable (as of September 2016).

In this scenario, and lacking from a scalable and queryable archive, current applications and businesses on top of any of the aforementioned publication schemes have been left to bear all problems associated with version management and discontinued services. A traditional view of digitally preserving them by "pickling them and locking them away" for future use, would conflict with their evolution. There are a number of approaches and frameworks, such as the LOD2 stack [1], that manages a full life-cycle of the Data Web. More specifically, these techniques are expected to tackle major issues such as the synchronisation, curation, appraisal, citation, archiving, and sustainability problems.

The goal of this workshop was to provide a forum for researchers and practitioners who apply Linked Data technologies to discuss, exchange and disseminate their work on preserving Linked Data. More broadly, our aim was to enable communities interested in data, knowledge and ontology dynamics to network and cross-fertilise. The topics of interest included, but were not limited to the following themes related to the evolution and preservation of linked data:

- Change Discovery
  - Change detection and computation in data and/or vocabularies;
  - Change traceability;
  - Change notifications (e.g., PubSubHubPub, DSNotify, SPARQL Push);
  - Visualisation of evolution patterns for datasets and vocabularies;
  - Prediction of changes.
- Formal models and theory
  - Formal representation of changes and evolution;
  - Change/Dynamicity characteristics tailored to graph data;
  - Query language for archives;
  - Freshness guarantee for query results;
  - Freshness guarantee in databases.
- Data Archiving and preservation
  - Scalable versioning and archiving systems/frameworks;
  - Query processing/engines for archives;
  - Efficient representation of archives (compression);
  - Benchmarking archives and versioning strategies.

The workshop included a keynote talk by Dr. Axel Polleres (Section 2), three research papers (Section 3) and one industry contribution (Section 4). The workshop was closed with a plenary discussion (Section 5) on the most relevant and open questions, collected for

future directions of the workshop (Section 6). Furthermore, 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"*. MEPDaW proceedings are publicly available in [6].

2 Keynote Talk

Dr. Axel Polleres gave an inspiring talk[1] on *Archiving Linked and Open Data*. Dr. Polleres is a full professor in the area of "Data and Knowledge Engineering" at the Institute of Information Business of Vienna University of Economics and Business (WU Wien). He has published more than 100 articles, most of them in top journals and conferences, and he actively contributed to World Wide Web Consortium (W3C) standardisation efforts, co-chairing the W3C SPARQL working group. He was invited as a renowned expert on querying and reasoning about ontologies, rules languages, Semantic Web technologies, Linked Open Data and knowledge management.

During his talk, Dr. Polleres first motivated the need of monitoring evolution and archiving as a necessary step to understand the evolution (e.g. climate change, social and topic trends, etc.), the validity of the information (e.g. continuous revisions in Wikipedia articles) and its consequences (e.g. predicting data in the Open City Data Pipeline [2]). He briefly reviewed existing preservation projects in the Web, such as the Common Crawl[2], Internet Memory[3], Internet Archive[4], and time-based access such as the HTTP memento protocol (RFC 7089). As stated by Dr. Polleres, most of current archiving approaches present a poor granularity (access to "some" snapshots), offer aggregated data instead of raw data, suffer from scalability problems and, in general, disregard a large-scale structured access (e.g. with temporal query languages). Similar problems arise in the particular field of Linked Data archiving, with even fewer and more limited approaches, such as the dynamic linked data observatory[5] [13]. Dr. Polleres established important open research challenges on evolving structured interlinked data, summarised as follows:

- How can we represent archives of continuously evolving linked datasets? (efficiency vs. compact representation).
- How can we improve completeness of archiving?
- How can emerging retrieval demands in archiving (e.g. time-traversing and traceability) be satisfied?
- 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?

Then, Dr. Polleres focused on three general archiving challenges: **the synchronisation problem**, i.e. how we can monitor changes, **the appraisal problem**, how we can assess the quality of a dataset (via archiving) and **the archiving and query problem**, i.e. how

[1]Slides of the talk: https://aic.ai.wu.ac.at/ polleres/presentations/20160530Keynote-MEPDaW2016.pdf
[2]http://commoncrawl.org/
[3]http://internetmemory.org/
[4]https://archive.org/index.php
[5]http://swse.deri.org/dyldo/

we can efficiently archive and perform time-based retrieval queries of a dataset. As for synchronisation, he distinguished between pull changes (crawl) vs. push changes (notify), and summarised one of his previous work on adaptive archiving [18] to accurately capture content change, and the DBpedia Wayback Machine [9] as an example of recreating the versions from the original sources (in this particular case, re-apply mappings on the Wikipedia revision history). Regarding the appraisal problem, he presented a recent work on archiving and analysing the history of datasets in open data portals [19]. Finally, Dr. Polleres addressed the archiving and query problem reviewing current RDF archiving techniques and structured query languages managing time [8], presenting an initial blueprint on benchmarking archives of semantic data [10]. To conclude, Dr. Polleres summarised some open questions, such as the compact representation of archives, the missing query language for archives and the need of a large-scale archiving indexing/infrastructure, and discuss with the audience if there is an actual and urgent need of archiving in the community.

3 Research Papers

The first paper presented in this workshop, *Continuously Updating Query Results over Real-Time Linked Data*[6], describes how Triple Pattern Fragments [20] can be extended in order to allow clients to execute real-time SPARQL queries whose results are "continuously" updated on the client-side. In their presentation, the authors discuss various experiments showing how the server's costs are lowered, however at the expense of execution time per query on the client. This paper was chosen by the program committee as the best paper and was extended in [16].

Meimaris and Papastefanatos presented *The EvoGen Benchmark Suite for Evolving RDF Data* [15]. In this talk, the authors describe a benchmark framework that extends the Lehigh University Benchmark (LUBM) [11] in order to cover the need for creating synthetic evolving RDF data, that will then be used to benchmark versioning and change detection techniques in dynamic datasets. In this talk, the authors discuss the various parameters required to generate this kind of evolving RDF data.

The final research paper of this workshop was *Toward Semantic Sensor Data Archives on the Web*[7], authored by Calbimonte and Aberer [5]. In this presentation, the authors talk about the current barriers of archiving semantic sensor data and possible solutions. The authors presented their visionary semantic-based architecture using the well known DCAT[8] and SSN[9] ontologies, together with a use case based on their air quality monitoring deployment and preliminary results comparing them with an optimised ERI [7] interchange format for RDF.

[6]Slides are available at: `http://www.slideshare.net/RubenTaelman/continuously-updating-query-results-over-realtime-linked-data` by Taelman et al. [17]. Last Accessed on 01/09/2016

[7]Slides are available at: `http://www.slideshare.net/jpcik/toward-semantic-sensor-data-archives-on-the-web`. Last Accessed on 01/09/2016

[8]`http://www.w3.org/TR/vocab-dcat/`

[9]`http://www.w3.org/TR/vocab-ssn/`

4 Industry Paper

In this workshop we had one industry talk from the founders of Dydra (`http://dydra.com`). Dydra is a cloud graph database that can be accessed and updated via SPARQL. In their talk *Transaction-Time Queries in Dydra*, Anderson and Bendiken describe a taxonomy of transaction-time queries for enabling versioning of RDF graphs in Dydra. The authors present some examples showing how the Dydra architecture combined with extensions to the language supports stakeholders to work with versioned RDF data.

5 Closing Discussion

During the workshop, there was an active discussion between all presentations, discussing future directions, possible improvements, and raising controversial yet interesting questions to all our presenters. The plenary was closed by a round table discussion, with Dr. Pollares raising the following question:

> "Is there an actual and urgent need in the (Semantic Web) community for handling the dynamicity of the Data Web?"

This debatable questions led to many other insights towards the need of having techniques to handle the evolution and the subsequent preservation of such expansion. The audience asked and debated whether a *killer-app* is needed in this regard. Mixed reactions were given regarding the killer-app, where the creators of Dydra argued that whilst this evolution and preservation of the Data Web is of utmost importance for their clients, they just need an infrastructure on top of technologies such as SPARQL rather than an extra "complex" application.

6 Conclusion and Future Directions

The MEPDaW organisers were pleased of the quality of the papers and their presentation in the workshop, the lively discussing of the participants and the emerging involvement of industry partners. All this showed the timeliness of the topic and the lack of well-established techniques to cope with most of the challenges when managing preserving the evolving Data Web. In particular, we envision four important topics under discussion:

- Compact representation of RDF and Linked Data archives.
- Query languages for archives satisfying these requirements for evolving interlinked data.
- Index archives at large scale (and keeping up with evolution rate) to process time-based queries efficiently.
- Query optimization for archives, e.g. enabling query rewriting for querying archives of structured non-RDF sources.

The organisers are planning a follow-up workshop next year, expecting that the works can address this topics as well as other related areas such as evolution studies, adaptive change monitoring and practical archiving applications.

Acknowledgements

We would like to thank the authors for their high-quality contributions and their active participation in the workshop. We express our gratitude to all program committee members for reviewing the submissions for the workshop. We are also grateful to the organisers of the ESWC 2016 conference for their support, and our keynote speaker, "Axel Polleres" from the Vienna University of Economics and Business. This workshop was co-organised by members funded by the Austrian Science Fund (FWF): M1720-G11.

References

[1] S. Auer, L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes, B. V. Nuffelen, C. Stadler, S. Tramp, and H. Williams. Managing the life-cycle of linked data with the LOD2 stack. In *International Semantic Web Conference (2)*, volume 7650 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 2012.

[2] S. Bischof, C. Martin, A. Polleres, and P. Schneider. Open city data pipeline-collecting, integrating, and predicting open city data. In *KNOW@ LOD*, 2015.

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. In *International Journal on Semantic Web and Information Systems*, volume 5, pages 1–22. Quarterly, 2009.

[4] C. Buil-Aranda, A. Hogan, J. Umbrich, and P.-Y. Vandenbussche. Sparql web-querying infrastructure: Ready for action? In *The Semantic Web–ISWC 2013*, pages 277–293. Springer, 2013.

[5] J.-P. Calbimonte and K. Aberer. Toward semantic sensor data archives on the web. In *Proceedings of the 2nd Workshop on Managing the Evolution and Preservation of the Data Web*, volume 1585 of *CEUR Workshop Proceedings*, pages 36–51, May 2016.

[6] 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.

[7] J. D. Fernández, A. Llaves, and O. Corcho. Efficient rdf interchange (eri) format for rdf data streams. In *Proceedings of the 13th International Semantic Web Conference - Part II*, ISWC '14, pages 244–259, New York, NY, USA, 2014. Springer-Verlag New York, Inc.

[8] J. D. Fernández, A. Polleres, and J. Umbrich. Towards efficient archiving of dynamic linked open data. pages 34–49, 2015.

[9] J. D. Fernández, P. Schneider, and J. Umbrich. The dbpedia wayback machine. In *Proceedings of the 11th International Conference on Semantic Systems*, pages 192–195. ACM, 2015.

[10] J. D. Fernandez Garcia, J. Umbrich, and A. Polleres. Bear: Benchmarking the efficiency of rdf archiving. 2015.

[11] Y. Guo, Z. Pan, and J. Heflin. Lubm: A benchmark for owl knowledge base systems. *Web Semant.*, 3(2-3):158–182, Oct. 2005.

[12] S. Harris and A. Seaborne. *SPARQL 1.1 Query Language*. W3C Recommendation. 2013.

[13] T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing Linked Data Dynamics. In *The Semantic Web: Semantics and Big Data*, volume LNCS 7882, pages 213–227. Springer Berlin Heidelberg, 2013.

[14] F. Manola, E. Miller, and B. McBride. Rdf 1.1 primer. *W3C Working Group Note, February*, 25, 2014.

[15] M. Meimaris and G. Papastefanatos. The evogen benchmark suite for evolving rdf data. In *Proceedings of the 2nd Workshop on Managing the Evolution and Preservation of the Data Web*, volume 1585 of *CEUR Workshop Proceedings*, pages 20–35, May 2016.

[16] R. Taelman. *Continuously Self-Updating Query Results over Dynamic Heterogeneous Linked Data*, pages 863–872. Springer International Publishing, Cham, 2016.

[17] R. Taelman, R. Verborgh, P. Colpaert, E. Mannens, and R. Van de Walle. Continuously updating query results over real-time Linked Data. In *Proceedings of the 2nd Workshop on Managing the Evolution and Preservation of the Data Web*, volume 1585 of *CEUR Workshop Proceedings*, pages 1–10, May 2016.

[18] J. Umbrich, N. Mrzelj, and A. Polleres. Towards capturing and preserving changes on the web of data. In *DIACHRON Workshop on Managing the Evolution and Preservation of the Data Web*, volume CEUR 1377, pages 50–65. 2015.

[19] J. Umbrich, S. Neumaier, and A. Polleres. Quality assessment and evolution of open data portals. In *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on*, pages 404–411. IEEE, 2015.

[20] R. Verborgh, O. Hartig, B. Meester, G. Haesendonck, L. Vocht, M. Vander Sande, R. Cyganiak, P. Colpaert, E. Mannens, and R. Walle. Querying datasets on the web with high availability. In *Proceedings of the 13th International Semantic Web Conference - Part I*, ISWC '14, pages 180–196, New York, NY, USA, 2014. Springer-Verlag New York, Inc.