Managing the Evolution and Preservation of the Data Web

Jeremy Debattista[1] and Javier D. Fernández[2,3] and Maria-Esther Vidal[4,5,6] and Jürgen Umbrich[7]

[1] ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[2] Vienna University of Economics and Business, Austria
[3] Complexity Science Hub Vienna, Austria
[4] Technische Informationsbibliothek-TIB, Germany
[5] L3S Research Center, Leibniz University of Hannover, Germany
[6] Universidad Simón Bolívar, Venezuela
[7] Onlim GmbH, Telfs, Austria
debattij@tcd.ie

1 Introduction

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).

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 implications to other datasets linked to it. The distributed, dynamic 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 afford on-going access to continuously assess the quality of such dynamic datasets.

The aim of this special issue of the Journal of Web Semantics is to present latest advances in the area and further attract attention to these issues from interested communities. Providing new techniques and innovative solutions to address challenges portrayed by the ever-growing Web of Data, will foster wider adoption of Semantic technologies within different domains and scenarios that would increase the consumption of Linked Data. We received 15 submissions for the special issue. The papers went through a rigorous procedure of review involving at least three expert reviewers for each paper. After three rounds of review, we selected five high-quality research papers that made contributions to both research and practice.

2 Data Versioning

Triple Storage for Random-Access Versioned Querying of RDF Archives, by Ruben Taelman, Miel Vander Sande, Joachim Van Herwegen, Erik Mannens and Ruben Verborgh, tackles the inherent scalability issues when representing and querying versioned RDF data. In this work, the authors propose a novel index for RDF archives that follows a hybrid multiversion storing strategy, i.e., it combines storing full materialise versions (snapshots) followed by delta chains. Then, the authors present how compressed indexing of such snapshots and delta chains, together with additional metadata can highly reduce storage needs and lookup times, at the cost of additional ingestion time. The novel RDF archiving solution is materialised in OSTRICH, an open-source implementation with offset support and result streaming, hence it is compatible with the Web-friendly Triple Pattern Fragments interface.

Decentralized Collaborative Knowledge Management using Git, by Natanael Arndt, Patrick Naumann, Norman Radtke, Michael Martin and Edgard Marx. The author focuses on supporting distributed collaboration on RDF datasets, building upon the idea of a Git system. In this article, the authors propose a formal model to express changes and represent the (decentralised) evolution of RDF datasets. Based on this model, different git-based operations are supported, such as tracking, reverting, branching, and merging changes. Regarding this latter, the authors inspect different implementations of the merge operation, including unsupervised and supervised approaches. The full list of operations is practically integrated in QuitStore, a suite to support distributed version control for RDF datasets.

3 Quality Issues in Evolving Knowledge Graphs

Completeness and Consistency Analysis for Evolving Knowledge Bases, by Mohammad Rashid, Giuseppe Rizzo, Marco Torchiano, Nandana Mihindukulasooriya, Oscar Corcho, Raúl García-Castro. The authors tackle the problem of identifying completeness and consistency issues using, respectively, evolution analysis and integrity constraints. They propose an approach that relies on information from data profiling to generate integrity constraints in the form of SHACL RDF shapes, to validate completeness and consistency of a knowledge base. The work is guided by two research questions that state the prominent role of knowledge evolution in the identification of quality issues in knowledge bases. First, the authors resort to data profiling to determining the completeness of the entities of the relevant classes of a KG during the evolution of the KG. Second, leaning models are utilised in the prediction of consistency using integrity constraints. Specifically, three types of constraints are evaluated: minimum cardinality, maximum cardinality, and range constraint. The paper also contributes with the results of two quantitative and qualitative experimental studies over two real-world knowledge bases: 3cixty Nice[1] and DBpedia[2]. The reported results indicate both the benefits of using dynamic features in the detection of completeness and the power of learning models during consistency analysis. These technical results are put in perspective and provide evidence of the benefits of tracking knowledge evolution in the prediction of data quality.

TISCO: Temporal Scoping of Facts, by Anisa Rula, Matteo Palmonari, Simone Rubinacci, Axel-Cyrille Ngonga Ngomo, Jens Lehmann, Andrea Maurino, and Diego Esteves, tackles the problem of curating time intervals associated with facts. The authors propose TISCO, a three-step solution based on combining contextual knowledge from both the Web of Documents and Linked Data resources, more specifically DBpedia, in order to try to detect the temporal scope of a dynamic triple. The proposed solution is implemented in a framework with a number of functions and configuration parameters providing efficient matching and selection of time intervals. This framework is included in a prototype which allows users to explore facts with temporal scopes and enables the testing of new algorithms for matching and selection functions.

4 Ontology Evolution and Concept Drift

SemaDrift: A hybrid method and visual tools to measure semantic drift in ontologies by Thanos G. Stavropoulos, Stelios Andreadis, Efstratios Kontopoulos, Ioannis Kompatsiaris, applies to idea of concept drift into ontologies in order to measure changes in ontologies across different version in time. In this article, the authors argue that widely applied drift measuring methods and tools are not directly applicable to data models described using semantic formalisms. To address this challenge, the authors propose a morphing-based hybrid approach to identify concepts in ontology versions across time. This also enabled the identification of concept identities within an ontology without any human intervention. These methods are embedded into the SemaDrift application suite, which also enables non-experts to visualise drifts within a versioned ontology

[1] https://www.3cixty.com
[2] http://wiki.dbpedia.org

5 Opportunities and Future Directions in Managing the Evolution and Preservation of the Data Web

Based on the open research avenues of the papers in this issue, and the research questions raised during the keynotes and discussions in previous editions of the MEPDaW workshop on managing the evolution and preservation of the data web[3], we identify the following opportunities and future directions:

Practical guidelines. While there are general Linked Data guidelines and best practices regarding publishing data on the web\footnote{\url{https://www.w3.org/TR/dwbp/}}, there is a lack of practical recipes for publishing and managing evolving data and knowledge. The heterogeneity of design models and non-standard strategies hinder the reusability of the data and prevent further adoption by researchers and practitioners. They expect that the emerging concept of Knowledge Graphs could serve as a catalyst to tackle these issues.

Query planners and optimisers. In the last decade, the traditionally static query planners an optimisers for semantic data have given way to dynamic strategies that adapt to different conditions, such as network delays, server overwork, presence of replicas or peers, etc. However, little attention has been paid to optimisations for evolving data. We expect a novel generation of stores to consider dynamic changes as a core component for large-scale and complex querying of temporal evolving data and knowledge.

Data integration. Linked Data and the more recent concept of Knowledge Graphs excel in integrating data from different domains and heterogeneous sources of data. However, data integration tasks present even more challenges for data or knowledge (e.g. schema) that evolve, as it is well known that data and knowledge are subject to unnoticed changes, given the free nature of the Web. Novel integration algorithms and tools need to consider this evolving nature.

Benchmarks. Although, in recent years, there were several works on benchmarking archives of RDF data, these are mainly based on evaluating the performance of the indexes on versioned RDF data, in space and the performance to resolve time-based queries. In contrast, there is a need of full set of benchmarks for multiple tasks for evolving data and knowledge, such as efficient update mechanisms, reasoning and prediction, exploration, quality assessment and validation and, in general, managing, predicting, and curating evolution.

Reasoning and prediction. Finally, we expect to witness an increasing attention to techniques for extracting and predicting evolving patterns. For instance, novel trends on machine learning and deep learning could be exploited to “understand” how concepts and instances are evolving, or to identify quality issues in evolving graphs. Related topics, such as trend analysis of evolving knowledge graphs and concept drift detection and prediction are also in scope for the development of knowledge graphs.

6 Acknowledgements

We would like to thank all the people involved in this special issue as well as the funds who support the main work of the editors, the EU’s Horizon 2020 research and innovation programme SPECIAL [grant number: 731601], the Austrian Research Promotion Agency's (FFG) program “ICT of the Future” [grant number: 861213], the Irish Research Council Government of Ireland Postdoctoral Fellowship [grant number: GOIPD/2017/1204] and Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology [grant number: 13/RC/2106].

Shepherds. Some of the papers in this special issue have their origins in the 3rd edition of the MEPDaW workshop on managing the evolution and preservation of the data web. We would like to particularly thank the mentors who provided assistance and feedback to improve the quality of the workshop papers:

- Christoph Lange, Fraunhofer IAIS Sankt Augustin, Germany and University of Bonn, Germany
- Giorgos Flouris, FORTH-ICS, Greece
- Claudio Gutiérrez, Universidad de Chile, Chile

[3] https://mepdaw2018.ai.wu.ac.at

- Olaf Hartig, Linköping University, Sweden
- Axel Polleres, Vienna University of Economics and Business, Austria

Reviewers. Finally, we would like to thank all reviewers of the special issue for their selfless help, comments and suggestions.

- Aiden Hogan
- Albin Ahmeti
- Aldo Gangemi
- Alessandro Adamou
- Anastasia Dimou
- Anisa Rula
- Antoine Zimmermann
- Atul Nautiyal
- Bernhard Haslhofer
- Christoph Debruyne
- Danh Le Phuoc
- Diego Collarana
- Elmar Kiesling
- Fabrizio Orlandi
- Fajar J. Ekaputra
- Giuseppe Pirrò
- Hajira Jabeen
- Harald Sack
- Harshvardhan Pandit
- Ioannis Chrysakis
- Jose Emilio Lara Gayo
- Juan Sequeda
- Kostas Stefanidis
- Laura Hollink
- Magnus Knuth
- Marios Meimaris
- Mauro Dragoni
- Miguel A. Martínez-Prieto
- Nadeschda Nikitina
- Natasha Noy
- Rafael S. Gonvalves
- Rob Brennan
- Ruben Taelman
- Ruben Verborgh
- Sebastian Neumaier
- Simon Scerri
- Tobias Kafer
- Vadim Savenkov
- Valeria Fionda
- Vit Novacek
- Wouter Beek
- Yaron Kanza