Preface of SeWeBMeDA 2018: Semantic Web solutions for large-scale biomedical data analytics*

Ali Hasnain[1], Oya Beyan[2], Stefan Decker[2], and Dietrich Rebholz-Schuhmann[1]

[1] Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland
{ali.hasnain,rebholz}@insight-centre.org
[2] Fraunhofer FIT, RWTH Aachen University
{beyan,decker}@dbis.rwth-aachen.de

The second edition of SeWeBMeDA-2018 workshop invited papers for life sciences and biomedical data processing, as well as the amalgamation with Linked Data and Semantic Web technologies for better data analytics, knowledge discovery and user-targeted applications.

This workshop at the Extended Semantic Web Conference (ESWC) targeted original contributions describing theoretical and practical methods and techniques that present the anatomy of large scale linked data infrastructure, which covers: the distributed infrastructure to consume, store and query large volumes of heterogeneous linked data; using indexes and graph aggregation to better understand large linked data graphs, query federation to mix internal and external data-sources, and linked data visualisation tools for health care and life sciences. It will further cover topics around data integration, data profiling, data curation, querying, knowledge discovery, ontology mapping / matching / reconciliation and data / ontology visualisation, applications / tools / technologies / techniques for life sciences and biomedical domain. SeWeBMeDA aims to provide researchers in biomedical and life science, an insight and awareness about large scale data technologies for linked data, which are becoming increasingly important for knowledge discovery in the life sciences domain.

This year, we accepted three papers, we invited a keynote speaker, organised a short hackathon and also discussed on current issues along with future steps for large scale data in biomedical domain.

Keynote talk was given by Maria-Esther Vidal who is the head of the Scientific Data Management group at TIB Leibniz Information Centre for Science and Technology, Germany and a full professor (on-leave) at Universidad Simón Bolívar (USB) Venezuela. Her interests include Big data and knowledge management, knowledge representation, and semantic web with more than 130 peer-reviewed papers in Semantic Web, Databases, Bioinformatics, and Artificial Intelligence. The title of her talk was "Synthesizing Big Data into Actionable Knowledge", where she discussed the role of Big data in promoting emerging scientific and interdisciplinary research by enabling decision-making. She described that knowledge-driven approach is capable to ingest Big data sources and integrate them into a knowledge graph that represents not only the meaning of the entities published by these data sources, but also that provides the basis for the discovery of unknown patterns and associations between these entities. The features of this knowledge-driven framework are shown in the context of

* Joint proceedings are publicly available in [1].

the EU funded project iASiS (http://project-iasis.eu/) , where it is used to pave the way for personalized diagnosis and treatments. The presentation slides are available at: (https://goo.gl/aH92pM).

As mentioned we had three paper presentations:

Gleim et al [3], proposes an automated schema extraction approach compatible with existing Semantic Web-based technologies. The extracted schema enables ad-hoc query formulation against privacy sensitive data sources without requiring data access, and successive execution of that request in a secure enclave under the data provider's control. The developed approach permit user to extract structural information from non-uniformed resources and merge it into a single schema to preserve the privacy of each data source. Initial experiments show that this approach overcomes the reliance of previous approaches on agreeing upon shared schema and encoding a priori in favor of more exible schema extraction and introspection.

Hasnain et al [2], assess the FAIR principles against the LOD principles to determine, to which degree, the FAIR principles reuse LOD principles, and to which degree they extend the LOD principles. This assessment helps to clarify the relationship between both schemes and gives a better understanding, what extension FAIR represents in comparison to LOD. This publication concluds, that LOD gives a clear mandate to the openness of data, whereas FAIR asks for a stated license for access and thus includes the concept of reusability under consideration of the license agreement. Furthermore, FAIR makes strong reference to the contextual information required to improve reuse of the data, e.g., provenance information. According to the LOD principles, such meta-data would be considered interoperable data as well, however, the requirement of extending of data with meta-data does indicate that FAIR is an extension of the LOD (in contrast to the inverse).

Nayak et al [4], propose that the use of topic modeling, specifically non-negative matrix factorization (NMF), as a first step towards dimensionality reduction when dealing with large amounts of data. In this position paper, as a use case, author applied NMF to the BioSamples metadata and present preliminary results.

At the end of the workshop we organised a short Hackathon title "Privacy-Preserving Information Extraction with Bloom Filters". At the beginning of the hackathon, we provided a short introduction to the prerequisites, such as bloom filters, general privacy issues and frameworks that can be used (Python or KNIME). Then, each team involved in the hackathon was given a unique Knowledge Graph onto which they could apply information retrieval techniques to build up some experience with the given framework. Next, the the Bloom Filters were applied and discussed the suitable metrics for valuing an unseen knowledge graph based on a query response that may contain false positives. Finally, each team formulated queries for estimating the worth of an unseen Knowledge Graph and ultimately made a decision about which other teams Knowledge Graph complements their own Knowledge Graph the best.

Acknowledgments

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 2018 conference for their support, and our keynote speaker, Maria-Esther Vidal who is the head of the Scientific Data Management group at TIB Leibniz Information Centre for Science and Technology, Germany and a full professor (on-leave) at Universidad Simón Bolívar (USB) Venezuela.

SeWeBMeDA-2018 workshop was co-organised by Insight Centre for Data Analytics NUI Galway and Fraunhofer FIT, RWTH Aachen University. This workshop has been supported in part by Science Foundation Ireland under Grant Number SFI/12/RC/2289.

References

1. O. Beyan, J. Debattista, S. Decker, J. D. Fernández, A. Hasnain, M. I. Ali, P. Patel, D. Rebholz-Schuhmann, D. T. Amit Sheth, J. Umbrich, and M.-E. Vidal, editors. *Joint proceedings of the 4th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW), the 2nd Workshop on Semantic Web solutions for large-scale biomedical data analytics (SeWeBMeDA), and the Workshop on Semantic Web of Things for Industry 4.0 (SWeTI)*, number 2112 in CEUR Workshop Proceedings, Aachen, 2018.
2. A. Hasnain and D. Rebholz-Schuhmann. Assessing fair data principles against the 5-star open data principle. In *2nd workshop on Semantic Web solutions for large-scale biomedical data analytics (SeWeBMeDA)*, 2018.
3. L. Z. O. K. H. S. S. D. Lars C. Gleim, Md. Rezaul Karim and O. Beyan. Using schema extraction for query design without data access to enable privacy maintaining processing of sensitive data. In *2nd workshop on Semantic Web solutions for large-scale biomedical data analytics (SeWeBMeDA)*, 2018.
4. A. Z. Stuti Nayak and M. Dumontier. Quality assessment of biomedical metadata using topic modeling. In *2nd workshop on Semantic Web solutions for large-scale biomedical data analytics (SeWeBMeDA)*, 2018.