Preface for the Knowledge Graph Building and Large Scale RDF Analytics Workshops

Pieter Heyvaert[1], David Chaves-Fraga[2], Freddy Priyatna[2], Anastasia Dimou[1], Juan Sequeda[3], Hajira Jabeen[4], Damien Graux[5], Gezim Sejdiu[4], Mohammed Saleem[6], and Jens Lehmann[4]

[1] IDLab, Dept of Electronics and Information Systems, Ghent University – imec
{pheyvaer.heyvaert,anastasia.dimou}@ugent.be
[2] Ontology Engineering Group, Universidad Politécnica de Madrid
{dchaves,fpriyatna}@fi.upm.es
[3] data.world
juan@data.world
[4] University of Bonn
{sejdiu,jens.lehmann}@cs.uni-bonn.de,jabeen@iai.uni-bonn.de
[5] Fraunhofer IAIS
damien.graux@iais.fraunhofer.de
[6] University of Leipzig
saleem@informatik.uni-leipzig.de

1 Introduction

More and more Knowledge Graphs are generated for private, e.g. Siri[7], Alexa[8], or public use, e.g. DBpedia[9], Wikidata[10]. While techniques to automatically generate Knowledge Graphs from existing Web objects exist (i.e. scraping Web tables), the majority is typically generated by transforming the content of existing datasets in different heterogeneous formats (e.g. RDB, CSV, XML, etc).

Initially, generating Knowledge Graphs from existing datasets was considered an engineering task. However, different scientific methods recently emerged. Lately, declarative methods (in the form of mapping languages) for describing rules to generate Knowledge Graphs and separate approaches and tools to execute those rules (so-called processors according to R2RML W3C recommendation) emerged. Addressing the challenges related to Knowledge Graphs generation requires well-funded research, including the investigation of concepts and development of tools and methods for their evaluation.

R2RML was recommended by W3C in 2012, and since then, different generalizations, extensions and alternatives were proposed, as well as processors for different languages' execution: RML [1], KR2RML [2], xR2RML [3], R2RML-F [4],

[7] https://www.apple.com/siri/
[8] https://developer.amazon.com/alexa
[9] http://dbpedia.org
[10] https://www.wikidata.org

and RMLC-iterator [5]. Certain approaches followed the *ETL-like paradigm*, e.g., R2RMLParser[11] [6], RMLMapper[12], RMLStreamer[13] [7] and CARML[14], while others the *query-answering paradigm*, e.g. Ultrawrap [8], Morph[15] [9], Sparqlify[16] [10], Ontop[17] [11], and morph-xR2RML [12]. Besides R2RML-based extensions, alternative approaches were proposed, e.g. SPARQL-Generate [13].

With the constant advancements in KG building, the size of Knowledge Graphs (KG) has reached a scale where centralized approaches for analytics are no longer feasible. Additionally, the ability to ingest heterogeneous data into KGs has opened novel challenges of scalable learning from this data. While the data within KGs can be transformed and preprocessed to be ingested by traditional learning algorithms, e.g. using Kernels or Propositionalization approaches, this requires additional computation and potentially loses the semantic information. It is, therefore, desirable to develop "scalable" approaches that exploit the semantic information contained in these KGs and present insightful analytical results. Recent technological advancements in distributed in-memory processing frameworks e.g. Apache Spark[18], Apache Flink[19] have made it easier to perform distributed computing using their specialised data structures. However, these, and many other such frameworks are not specialised to handle KGs and it remains challenging to perform "distributed analytics on semantic knowledge graphs". There is a strong need to bridge this gap and develop scalable and distributed analytics that make use of partial data, and at the same time exploit the semantic relationships to develop semantic-aware models for analysing KGs and data represented as RDF. The first Workshop on "Large Scale RDF Analytics (LASCAR)", has served as a platform to present and discuss the challenges and outcomes of distributed RDF processing and analytics.

2 The Knowledge Graph Building Workshop

The objective of organizing Knowledge Graph Building (KGB) was to provide a venue for scientific discourse, systematic analysis and rigorous evaluation of languages, techniques and tools for generating knowledge graphs, as well as practical and applied experiences and lessons-learnt from generating knowledge graphs in academia and industry. This workshop had special focus on Mapping Languages.

The Knowledge graph Building workshop was a full-day workshop that took place on 3rd June 2019 in Portoroz, Slovenia. KGB was co-located with the 16th Extended Semantic Web Conference (ESWC2019).

[11] `https://github.com/nkons/r2rml-parser`
[12] `https://github.com/RMLio/rmlmapper-java`
[13] `https://github.com/RMLio/RMLStreamer`
[14] `https://github.com/carml/carml`
[15] `https://github.com/oeg-upm/morph-rdb`
[16] `http://aksw.org/Projects/Sparqlify.html`
[17] `https://ontop.inf.unibz.it`
[18] `https://spark.apache.org/`
[19] `https://flink.apache.org/`

Dr Mariano Rodriguez-Muro[20], Ontologist in the Knowledge Graph Schema team of Google, was the keyonote speaker. He delivered an inspiring talk on Knowledge Graphs, Information Extraction, Machine Learning, Logics etc.

The workshop followed an open review process. The papers were submitted to a dedicated page of Open Review which is available at `https://openreview.net/group?id=eswc-conferences.org/ESWC/2019/Workshop/KGB`. This way, not only the papers, but also the reviews and potential discussions are open.

In total, the workshop received **seven papers**, six of which were accepted for presentation and five to be included in the proceedings. The workshop, as it aimed, received papers both from industry and academia.

The workshop was organized in a series of **four sessions**. There were three sessions with paper presentations, each one followed by a discussion slot around the presented topics, while a session was dedicated to the keynote. The first session was dedicated on knowledge graphs generation and consisted of two in use and one research paper, the second session was dedicated to the keynote, while the third session on position papers. The fourth session was dedicated to implementations, applications and demos. It consisted of a paper presenting a new tool which was followed by spontaneous tools presentations.

The following papers were presented at the workshop:

- Building Knowledge Graphs from Survey Data: A Use Case in the Social Sciences [14]
- Building a Knowledge Graph for Products and Solutions in the Automation Industry [15]
- Leveraging Ontologies for Knowledge Graph Schemas [16]
- Mapping languages: analysis of comparative characteristics [17]
- RocketRML - A NodeJS implementation of a use-case specific RML mapper [18]

The workshop was accompanied with the launch of the new W3C community group on Mapping Languages and Knowledge Graphs generation. More information about the Knowledge Graph construction working group is available at `https://www.w3.org/community/kg-construct/`.

Organizing Committee

- David Chaves-Fraga, Universidad Politécnica de Madrid
- Pieter Heyvaert, Ghent University - imec
- Freddy Priyatna, Universidad Politécnica de Madrid
- Anastasia Dimou, Ghent University - imec
- Juan Sequeda, data.world

[20] `https://sites.google.com/site/marianomuro/`

Programme Committee

- Ahmet Soylu, SINTEF/NTNU
- Aidan Hogan, Universidad de Chile
- Amrapali Zaveri, Maastricht University
- Antoine Zimmermann, École des Mines de Saint-Étienne
- Ben De Meester, IDLab, Ghent University - imec
- Boris Villazón-Terrazas, Arvato
- Claus Stadler, University of Leipzig
- Craig Knoblock, University of Southern California
- Dumitru Roman, SINTEF/University of Oslo
- Emanuele Della Valle, Politecnico di Milano
- Frank Michael, Université Côte d'Azur, CNRS, Inria, I3S
- Manolis Koubarakis, National Kapodistrian University of Athens
- Oscar Corcho, Universidad Politécnica de Madrid
- Ruben Verborgh, IDLab, Ghent University - imec
- Soren Auer, Technische Informationsbibliothek (TIB)

3 The Large Scale RDF Analytics Workshop

LASCAR, the workshop on Large Scale RDF Analytics was held as a part of ESWC -19. LASCAR invited papers covering the recent advancements to deal with the enormous growth of linked data. Olivier Curé from the Université Paris-Est Marne-la-vallée gave a keynote entitled "Analytical processing and reasoning in RDF stores". He explained why RDF database management is more an OLAP than an OLTP market. Three papers were accepted for the presentation in this half-day workshop. "Extending LiteMat toward RDFS++" [19] discussed an interesting encoding scheme for RDF data to support inferences based on RDFS and the `owl:sameAs` property, which is used in a distributed knowledge graph data management system. LiteMat proposes a simple dictionary look-up at query run-time. The details of the distributed implementation and efficiency of the encoding and query processing approaches over large synthetic datasets was discussed. The paper on "Enforceable Usage Policies for Industry 4.0" [20] discussed the use-control of business-critical in companies. It discussed that for an effective protection, both access and usage control enforcement is necessary for organizing Industry 4.0 collaboration networks. Formalized and machine-readable policies are a fundamental building block to achieve the needed trust level for real data-driven collaborations. Based on the experiences from the specification of the International Data Spaces Usage Control Language, the necessary implications and research gaps towards automatically monitored and enforced policies were outlined and necessary activities were presented. Sameh Mohamed presented "Unsupervised Hierarchical Grouping of Knowledge Graph Entities" [21] by describing a new unsupervised approach that learns to categorise entities into a hierarchy of named groups by effectively learning entity groups using a scalable procedure in noisy and sparse datasets. The authors have also published the collection of the group hierarchies.

The panel discussion in LASCAR was chaired by a group of experts including Prof. Dr. Olivier Curé from the University of Paris-Est Marne la Vallée (UPEM) , Prof. Dr. Jens Lehmann from the University of Bonn, and Dr. Maria Maleshkova from the University of Bonn. The interesting discussion covered the topics such as availability of large scale RDF data, challenges in RDF data distribution, and complexity of tasks like inference and analytics. The audience also participated in the discussions and asked questions to the panel members. LASCAR was successful in attracting approximately 30 participants.

Organizing Committee

- Hajira Jabeen, University of Bonn
- Damien Graux, Fraunhofer IAIS
- Gezim Sejdiu, University of Bonn
- Mohammed Saleem, University of Leipzig
- Jens Lehmann,University of Bonn

Programme Committee

- Afshin Sadeghi, University of Bonn, Germany
- Anisa Rula, University of Milano-Bicocca, Italy
- Claus Stadler, University of Leipzig, Germany
- Fabrizio Orlandi, Trinity College Dublin, Ireland
- Fathoni Musyafa, University of Bonn, Germany
- Gaurav Maheshwari, Fraunhofer IAIS, Germany
- Harsh Thakkar, University of Bonn, Germany
- Heba Mohamed, University of Bonn, Germany
- Mohamed. N. Mami, Fraunhofer IAIS, Germany
- Mohamed A. Sherif, University of Paderborn, Germany
- Patrick Westphal, University of Leipzig, Germany
- Priyansh Trivedi, Fraunhofer IAIS, Germany
- Rajjat Dadwal, Fraunhofer IAIS, Germany
- Shimaa Ibrahim, University of Bonn, Germany
- Simon Bin, University of Leipzig, Germany
- Nayef Roqaya, Coins Information system GmbH, Germany
- Tommaso Soru, Semantic Integration Ltd., London, United Kingdom

Acknowledgements

The described research activities were funded by Ghent University, imec, Flanders Innovation & Entrepreneurship (AIO), the Research Foundation – Flanders (FWO), and the European Union. The work presented in this paper is partially supported by the Spanish Ministerio de Economía, Industria y Competitividad and EU FEDER funds under the DATOS 4.0: RETOS Y SOLUCIONES - UPM Spanish national project (TIN2016-78011-C4-4-R) and by an FPI grant (BES-2017-082511).
LASCAR was partly supported by the following EU Horizon2020 projects Boost4.0 (GA no. 780732), QROWD (GA no. 723088), LAMBDA (GA no. 809965), SLIPO (GA no. 731581) and CLEOPATRA (GA no. 812997).

References

[1] Anastasia Dimou et al. "RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data". İn: *Proceedings of the 7th Workshop on Linked Data on the Web (LDOW)*. 2014.

[2] Jason Slepicka et al. "KR2RML: An Alternative Interpretation of R2RML for Heterogenous Sources." İn: *COLD*. 2015.

[3] Franck Michel et al. *xR2RML: Relational and Non-Relational Databases toRDF Mapping Language*. Tech. rep. 2017. URL: `https://hal.archives-ouvertes.fr/hal-01066663/document/`.

[4] Christophe Debruyne and Declan O'Sullivan. "R2RML-F: Towards Sharing and Executing Domain Logic in R2RML Mappings". İn: *LDOW*. 2016.

[5] David Chaves-Fraga et al. "Virtual Statistics Knowledge Graph Generation from CSV files". İn: *Emerging Topics in Semantic Technologies: ISWC 2018 Satellite Events*. Studies on the Semantic Web. IOS Press, 2018.

[6] Nikolaos Konstantinou et al. "An Approach for the Incremental Export of Relational Databases into RDF Graphs". İn: *International Journal on Artificial Intelligence Tools* (2015).

[7] Gerald Haesendonck et al. "Parallel RDF Generation from Heterogeneous Big Data". İn: *Proceedings of the International Workshop on Semantic Big Data*. 2019.

[8] Juan F. Sequeda and Daniel P. Miranker. "Ultrawrap: SPARQL Execution on Relational Data". İn: *Web Semantics* (2013).

[9] Freddy Priyatna, Oscar Corcho, and Juan Sequeda. "Formalisation and Experiences of R2RML-based SPARQL to SQL Query Translation Using Morph". İn: *Proceedings of the 23rd International Conference on World Wide Web*. 2014.

[10] Claus Stadler et al. "Simplified RDB2RDF Mapping". İn: *Proceedings of the 8th Workshop on Linked Data on the Web (LDOW2015)*. 2015.

[11] Diego Calvanese et al. "Ontop: Answering SPARQL Queries over Relational Databases". İn: *Semantic Web Journal* (2017).

[12] Franck Michel et al. "Translation of Relational and Non-relational Databases into RDF with xR2RML". İn: *WEBIST*. 2015.

[13] Maxime Lefrançois, Antoine Zimmermann, and Noorani Bakerally. "A SPARQL Extension for Generating RDF from Heterogeneous Formats". İn: *The Semantic Web: 14th International Conference*. 2017.

[14] Lars Heling et al. "Building Knowledge Graphs from Survey Data: A Use Case in the Social Sciences". İn: *Proceedings of the 1st Workshop on Knowledge Graph Building*. 2019.

[15] Thorsten Liebig et al. "Building a Knowledge Graph for Products and Solutions in the Automation Industry". İn: *Proceedings of the 1st Workshop on Knowledge Graph Building*. 2019.

[16] Daniela Oliveira, Ratnesh Sahay, and Mathieu d'Aquin. "Leveraging Ontologies for Knowledge Graph Schemas". İn: *Proceedings of the 1st Workshop on Knowledge Graph Building*. 2019.

[17] Ben De Meester et al. “Mapping Languages: Analysis of Comparative Characteristics”. İn: *Proceedings of the 1st Workshop on Knowledge Graph Building*. 2019.

[18] Umutcan Şimşek, Elias Kärle, and Dieter Fensel. “RocketRML - A NodeJS implementation of a use-case specific RML mapper”. İn: *Proceedings of the 1st Workshop on Knowledge Graph Building*. 2019.

[19] Olivier Curé et al. “Extending LiteMat toward RDFS++”. İn: *Proceedings of the 1st Workshop on Large Scale RDF Analytics*. 2019.

[20] Sebastian Bader and Maria Maleshkova. “Towards Enforceable Usage Policies for Industry 4.0”. İn: *Proceedings of the 1st Workshop on Large Scale RDF Analytics*. 2019.

[21] Sameh Mohamed. “Unsupervised Hierarchical Grouping of Knowledge Graph Entities”. İn: *Proceedings of the 1st Workshop on Large Scale RDF Analytics*. 2019.