Towards an infrastructure for sourced, reproducible and verifiable knowledge graphs

Sylvie Cazalens - Philippe Lamarre
LIRIS - Database team
Lyon - Campus de la Doua
France

8 octobre 2020

Keywords : Knowledge Graphs, provenance, reproducibility, verification, workflows, RDF.

1 Context

Knowledge Graphs (KGs) penetrate our everyday life [3, 7], telling us what to buy, what to eat, what to watch, what to learn, where and when to travel, etc. Major companies maintain knowledge graphs to power personal voice assistants and search engines. No doubt Google KG, the Facebook Entity Graph and the Microsoft Satori are useful. However, the access to these KGs is restricted and the way they are built and maintained is not transparent. Meanwhile, let it be for scientific results or for information used in companies, in journalism or politics, requirements about information quality are deeply increasing : in physics, chemistry, biology, computer science or any other field, scientific experiments must be reproducible in order to be credible. Within companies, the provenance of the data to be used is crucial for their credibility (how have they been produced, from which data sets, with which algorithms, who certifies them, in which context are they used...).

In order to meet all these expectations, the ANR project DeKaloG (2021-2024) aims at a general framework to build community, decentralized knowledge graphs according to principles of accessibility and transparency. It gathers the efforts of three teams : GDD (LS2N, Nantes), Wimmics (Inria, Sophia Antipolis and Université Côte d'Azur) and BD (LIRIS, Lyon).

2 Objectives

The general aim of this thesis is to contribute to the DeKaloG framework, with a focus on transparency, and more particularly on reproducibility(e.g. [4]) : knowledge within the gragh (or the graph itself) should be reproducible and verifiable. For facts deduced within the graph, one may rely on works about provenance [8, 6]. For fact obtained using external tools and directly introduced into the graph, questions about provenance, reproducibility and verifiability have to be addressed. The following objectives should be targeted :

— *Requirements for an extensible model of transparency, up to reproducibility.* A first step consists in drawing a whole picture of the needs in the context of knowledge graphs, leveraging results in other different related domains such as linked data and semantic web, but also some achievements in other scientific domains (medicine, biology...). A second step consists in designing an extensible model of different levels of transparency, that can be queried, consistent with the current semantic web standards.

— *Use of the proposed model to enable more transparency in knowledge graphs.* This requires to inject more metadata into knowledge graphs, which raises problems of data volume and thus performances. This is a major hindrance for scalability. Recent approaches to this problem (e.g. [5]) provide a starting point. Additionally, linked data and workflows ([2, 1]) can be intertwined to push transparency up to reproducibility.

— *Estimating/verifying the transparency degree of a KG.* One should be able to obtain information qualifying and quantifying the transparency degree of a knowledge graph she wants to use. This is also very important when building an index of knowledge graphs.

Hence, this thesis should result in an infrastructure enabling to link a knowledge graph with external solutions, accessed through services, for KGs and facts to be reproducible and verifiable by anyone.

3 Team and contact information

The thesis will take place within the database team (DB) of the LIRIS laboratory (Campus de la Doua, Lyon-Villeurbanne, liris.cnrs.fr), in collaboration with the other teams of project DeKaloG.

Contact information :

Philippe Lamarre : Philippe.Lamarre@insa-lyon.fr
Sylvie Cazalens : Sylvie.Cazalens@insa-lyon.fr

Références

[1] About openwdl. `http://www.openwdl.org/`. Accessed : 2020-09-22.

[2] Common workflow language. `https://www.commonwl.org/`. Accessed : 2020-09-22.

[3] Introducing the knowledge graph : things, not strings. `https://www.blog.google/products/search/introducing-knowledge-graph-things-not/`. Last Access on : 2020-09-22.

[4] Taylor J ; Galaxy Team Goecks J, Nekrutenko A. Galaxy : a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8), 2010.

[5] Olaf Hartig. Foundations of rdf⋆ and sparql⋆ (an alternative approach to statement-level metadata in RDF). In *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web, Montevideo, Uruguay, June 7-9, 2017*, 2017.

[6] Luc Moreau, Paul T. Groth, James Cheney, Timothy Lebo, and Simon Miles. The rationale of PROV. *J. Web Semant.*, 35 :235–257, 2015.

[7] Heiko Paulheim. Knowledge graph refinement : A survey of approaches and evaluation methods. *Semantic Web*, 8(3) :489–508, 2017.

[8] Marcin Wylot, Philippe Cudré-Mauroux, Manfred Hauswirth, and Paul T. Groth. Storing, tracking, and querying provenance in linked data. *IEEE Trans. Knowl. Data Eng.*, 29(8) :1751–1764, 2017.