CHIST-ERA Triple: improving data interoperability and federation across RDF knowledge graphs and Solid Pods

Jerven Bolleman[1], Elias Crum[2], Iulian Dragan[1], Jakub Galgonek[3], Mark Ibberson[1], Tarcisio Mendes de Farias[1], Marek Moos[3], Marco Pagni[1], Ruben Taelman[2], Jiří Vondrášek[3] and Ana Claudia Sima[1,*]

[1]*SIB Swiss Institute of Bioinformatics*
[2]*Ghent University*
[3]*IOCB Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences*

Abstract

The TRIPLE project, a collaborative effort between the SIB Swiss Institute of Bioinformatics, the University of Ghent and the IOCB Prague, aims to boost the (re)usability of existing knowledge graph resources and improve software tools for RDF data access, documentation and data model visualization. In addition, TRIPLE will increase interoperability between existing public SPARQL endpoints and private data stored in Solid Pods, thus creating an ecosystem of research data that can be seamlessly integrated through efficient and expressive federated SPARQL queries.

Keywords
RDF, federated SPARQL, Solid Pods, Open Research Data

The Resource Description Framework (RDF) provides a powerful way to structure data resources, coupled with the SPARQL query language, which can be used to interrogate these data, even when they are physically distributed, through the use of federated queries. Although powerful, these technologies are still currently underexploited due to significant technical and usability barriers limiting their use to a select few researchers. Knowledge graphs in RDF are generally not sufficiently described, with limited documentation, making it difficult to ascertain what to query and how to do so. In addition, execution times for complex queries are often very long and difficult to optimise. Finally, the query results are often difficult to integrate with
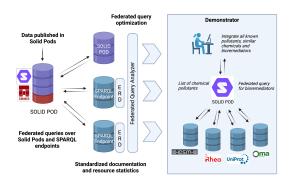
Figure 1: TRIPLE system architecture and data flow, interoperating Solid Pods and SPARQL endpoints

private data, such as preliminary, unpublished results. The TRIPLE project, a collaborative effort between the SIB Swiss Institute of Bioinformatics, the University of Ghent and the IOCB Prague, will address these challenges by developing innovative solutions on four fronts:

1. Storing private (unpublished) data in Solid [1] Pods, an emerging technology enabling decentralised data vaults to host private RDF endpoints, execute federated SPARQL queries and cache results.
2. Optimising federated queries spanning public and private SPARQL endpoints allowing users to query multiple resources from within their Solid Pod.
3. Adapting state-of-the-art RDF documentation tools and making them available for all SPARQL endpoints, including Solid Pods.
4. Developing data model visualisations, sets of standardised federated queries, and query analysis tools to help users understand the data and run efficient federated queries.

Finally, a demonstrator will show the impact of these advances when applied to a technically challenging use case of scientific relevance: the search for suitable organisms for bioremediation[1]. The components and data flow in the TRIPLE system architecture are illustrated in Figure 1. The TRIPLE project will improve the (re)usability of existing knowledge graph resources and software tools to access them. In doing so, TRIPLE will create conditions for reproducible research in any domain based on open or shared data and software. Furthermore, the interoperability with Solid Pods will enable researchers to integrate their data with knowledge graphs.

Acknowledgments

---

[1]https://en.wikipedia.org/wiki/Bioremediation

References

[1] A. V. Sambra, E. Mansour, S. Hawke, M. Zereba, N. Greco, A. Ghanem, D. Zagidulin, A. Aboulnaga, T. Berners-Lee, Solid: a platform for decentralized social applications based on linked data, MIT CSAIL & Qatar Computing Research Institute, Tech. Rep. (2016).