Drug Discovery and Big Linked Data

Ronald Siebes[1], Victor de Boer[1], Bryn Williams-Jones[2], and Stian Soiland-Reyes[3]

[1] VU University Amsterdam, the Netherlands,
rm.siebes@few.vu.nl, v.de.boer@vu.nl
[2] Open PHACTS Foundation, Cambridge, United Kingdom
bryn@openphactsfoundation.org
[3] School of Computer Science, University of Manchester, United Kingdom
soiland-reyes@cs.manchester.ac.uk

1 The Open PHACTS Drug Discovery Platform

A large part of the daily practice of a researcher doing in vitro Drug Discovery is comparing and manually matching high-quality information from multiple disciplines in the Life and Biomedical Sciences. The Open PHACTS Discovery Platform[4] is an initiative to integrate publicly available data relevant for both academia and the pharmaceutical industry. It integrates numerous datasets including for example ChEBI, ChemSpider, DrugBank and the GeneOntology. The platform provides an easy interface that allows researchers to consult the database without being confronted with the complexity of defining efficient Linked Data queries. A set of services are accessible via a RESTful interface.

The Open PHACTS Discovery Platform provides an interpretation of biomedical research activities (identified by domain experts) as *workflows* that are authored using visual tools. Workflows retrieve data via API calls. The platform executes the resulting instantiated queries at an endpoint that serves relevant data.Currently, the infrastructure uses commercial software to reason over the vast amount of RDF data and the Big Data Europe (BDE) project took up the challenge to get the same functionality but via open source Big Data technology.

2 The Big Data Europe infrastructure

The BDE project[5] is developing a re-usable Big Data infrastructure (BDI) needed by data-intensive science practitioners tackling a wide range of societal challenges. The infrastructure is designed to cover aspects of publishing and consuming semantically interoperable, large-scale, multi-lingual data assets and knowledge. This BDE infrastructure is designed to minimize the disruption of current workflows, and maximizes the opportunities by taking advantage of the latest European RTD developments, including multilingual data harvesting, data analytics, and data visualization. To test the effectiveness of the platform,

[4] http://www.openphactsfoundation.org
[5] https://www.big-data-europe.eu

multiple pilot implementations are developed in the various domains. The first of these pilots is the Drug Discovery Pilot implementation, which replicates much of the functionalities of the Open PHACTS platform. The infrastructure relies heavily on the Docker containers[6] and configuration via Docker Compose where generic 3rd party Docker containers (e.g. MemCached, MySQL, SPARK, HDFS) are combined with custom made pilot specific Docker containers.

3 Drug Discovery Pilot implementation

In the pilot we propose to demonstrate[7], the Open PHACTS functionality is implemented on the BDI. One goal of this pilot is to investigate dealing with the significant diversity of the entity name space in the bio-medical domain and exploring how this issue affects a generic big data infrastructure. Mapping this vast amount of entities leads to a significant increase of triples. A second goal is covering data and query security and privacy requirements and exploring how the methods used to handle this in the current implementation of the Open PHACTS Discovery Platform can be used to guide development of the generic BDE platform. An important challenge for this pilot is to replace the commercial cluster version RDF store, with an open source variant version: 4Store. To this end, we are implementing a 4Store BDE docker component and improving it in such a way that it can serve as a generic component on the BDE infrastructure[8].

The pilot integrates multiple datasets, available in RDF. The mappings between the identifiers used in the various datasets are freely available as RDF linksets[9]. Most datasets have a metadata description published in VoID. The functionality of the Open PHACTS services is described in SWAGGER. The following processing is carried out:

- Real time processing: Using an external service (such as the Scientific Lenses keyword expansion service) to process a query and then to execute the processed query on the data stored in the infrastructure.
- Batch processing: Data transformations that align and link datasets at ingestion time. The datasets above are regularly updated and must be periodically re-ingested.

The pilot implementation exposes a querying endpoint as well as a data ingestion endpoint for visualization or further processing.

The pilot itself is available in its entirety as Open Source software[10]. Both BDI and the pilot-specific components are implemented as Docker components.

Acknowledgements. This work is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 644564 www.big-data-europe.eu. We thank our BDE collaborators for their support.

[6] https://www.docker.com/what-docker
[7] https://github.com/big-data-europe/pilot-sc1-cycle1
[8] https://github.com/big-data-europe/docker-4store
[9] https://www.openphacts.org/2/sci/data.html
[10] Download and instructions at https://github.com/big-data-europe/pilot-sc1-cycle1