Federated Query Formulation and Processing through BioFed

Ali Hasnain[1], Syeda Sana e Zainab[1], Dure Zehra[1], Qaiser Mehmood[1], Muhammad Saleem[2], and Dietrich Rebholz-Schuhmann[1]

[1] Insight Center for Data Analytics, National University of Ireland, Galway
firstname.lastname@insight-centre.org
[2] Universität Leipzig, IFI/AKSW saleem@informatik.uni-leipzig.de

Abstract. A single interface for accessing life sciences (LS) data is a natural need to master the data deluge in this domain. The data in the LS requires integration and current integrative solutions increasingly rely on the federation of queries for distributed resources. This paper demonstrates BioFed, a federated SPARQL query processing system customised for LS-LOD. BioFed enables user to formulate as well as execute both federated and non-federated SPARQL queries over more than 130 public LS SPARQL endpoints.

Keywords: Linked Data (LD), Biomedical Data, Query Federation

1 Introduction

In bioinformatics research, it is often required to collect and correlate data from more than one data sources [7]. Collecting such distributed data is commonly carried out by using federated queries. The distributed and Linked architecture of Linking Open Data (LOD) has greatly attracted life sciences data providers to publish their data sources as SPARQL endpoints. This has motivated a considerable work on federated SPARQL query processing over LOD [13,2,1,11,10,8,12]. However, for a non-SPARQL expert, formulating meaningful federated SPARQL queries to collect required results from more than one SPARQL endpoints is still a challenging task [7].

BioFed [5] offers a single-point-of-access for distributed LS data, enabling scientists to access the data from LS sources without extensive expertise in SPARQL and Linked Data. It introduces a federated query processing system that is customised for Life Science Linked Open Data (LS-LOD) and able to execute SPARQL queries over more than 130 public LS SPARQL endpoints[1]. BioFed is an index-assisted [9] approach combined with SPARQL query re-writing to formulate and execute both federated and non-federated SPARQL queries. It performs a hybrid [9] source selection approach based on Autonomous Resource Discovery and Indexing (ARDI) [3,6] index and `SPARQL ASK` queries. The query

[1] The list of SPARQL endpoints was collected from publicly available Bio2RDF data sets and by searching for data sets in CKAN[2] tagged *"life science"* or *"healthcare"* and is available from `http://goo.gl/ZLbLzq`

re-writing component explicitly add the the `SPARQL SERVICE` clauses into the query. The resulting `SERVICES` annotated SPARQL 1.1 query is then executed on top of the Sesame framework. In particular, BioFed provides a graphical interface for formulating as well as executing federated SPARQL queries using drop-down menu. Further details and complete evaluation result can be found at [5]. In the next section, we demonstrate the BioFed's interface by using a running SPARQL query example.

2 BioFed Web Interface

The BioFed's (`http://vmurq09.deri.ie:8007/`) web interface provides the ability to directly enter a SPARQL query into an input box or to use the *Standard Query Builder*. The users are provided the option of viewing the results directly or downloading the results as a file in one of six (6) formats including Text, Comma Separated Values (CSV), Tab Separated Values (TSV), JavaScript Object Notation (JSON), Turtle and Extensible Markup Language (XML).

The default or standard query builder is an interface that provides a list of topics. When one topic is selected all the attributes of that topic are listed. This set of topics/concepts known as Query Elements (Qe), was defined in the context of Drug Discovery and Cancer Chemo prevention [14,4], come from list of dataset introduced in the previous section, and can be replaced by any other set of concepts define in other context *e.g.,* Protein-Protein Interaction. The user selects the attribute and enters the desired value either as a variable or a literal. The requisite lines are then added to the query input box. Multiple selections may be added to the query after which it can be edited.

After the query is executed, the user can choose to see the provenance data by clicking the "show" link on the query status. BioFed also keeps track of provenance information such as which endpoints were not available and how many triples were returned from each available endpoint.

SPARQL Query Formulation As mentioned before, BioFed is able to formulate and execute federated as well as non-federated SPARQL queries. In this example we show the formulation and execution of a non-federated SPARQL query. Suppose a biologist is interested to query *Molecular Formula, Average Molecular Weight, Cas Registry Number of class "Offer"*. In order to make such query the user first selects the concept Offer from the "Make Selection" drop box and is shown in Figure 1. It has various concepts such as Drugs, Molecules, Disease, Organism, Sugar *etc.* After selecting the concept *Offer*, he selects chemicalFormula, molecularWeightAverage, and casRegistryNumber properties from the *"Make Selection"* drop box that will now show various properties associated with the concept *Offer* and add them in query text area through "Add to Query" button (1).

After writing the query, the user can add *Filter*, *Limit* and *Order By* clause by himself. The "Output" drop box contains two options "Display the results" and "Download the results". After selecting "Display the results" option, the

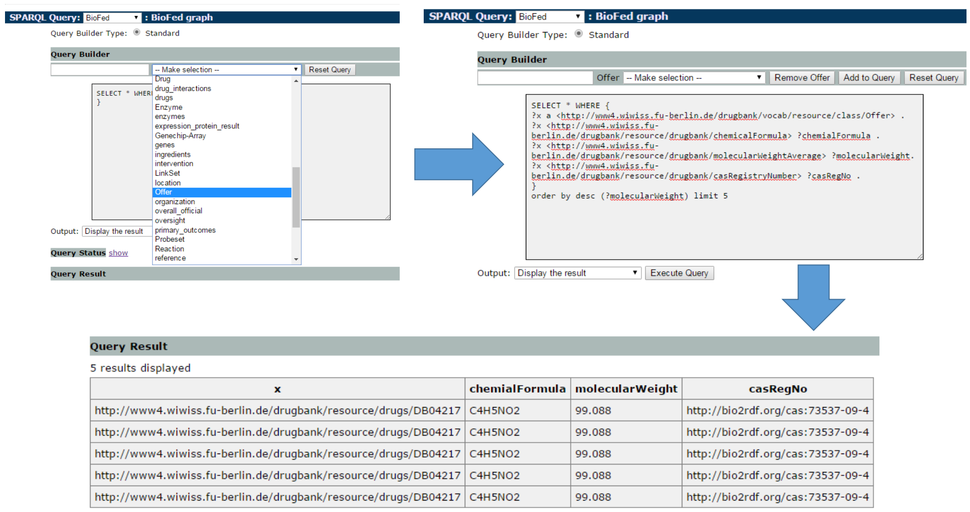

Fig. 1: BioFed Interface: Basic Query Formulation Execution and Displaying Results

user will click "Execute Query" button to display the results. The user can also download the query results in various formats such as CSV, JSON, XML, Text, TSV and Turtle. Instead of formulating the query from drop down menu if a SPARQL-wised user wants to execute any other query, he can write queries directly in SPARQL.

3 Conclusions

We believe that the proposed system can greatly help researchers in the biomedical domain to carry out their research by effectively retrieving relevant life science data. As the amount and diversity of biomedical data exceeds the ability of local resources to handle its retrieval and parsing, BioFed, facilitates federation over diverse resources. BioFed is a user friendly system for federated SPARQL query processes based on real biological data addressing meaningful biological queries. The web based interface provided by BioFed facilitates query generation which may pose difficulties for biological scientists. Currently the interface supports only basic query building using Qe defined in the context of drug discovery and in future we aim to provide an interface that can be able to formulate complex SPARQL queries.

Acknowledgements

The work presented in this paper has been partly funded by EU FP7 GRANATUM project (project number 270139) and Science Foundation Ireland under Grant No. SFI/12/RC/2289.

References

1. Acosta, M., Vidal, M.E., Lampo, T., Castillo, J., Ruckhaus, E.: Anapsid: an adaptive query processing engine for sparql endpoints. In: Proceedings of the 10th international conference on The semantic web - Volume Part I. pp. 18–34. ISWC'11 (2011)
2. Görlitz, O., Staab, S.: Splendid: Sparql endpoint federation exploiting void descriptions. In: Proceedings of the 2nd International Workshop on Consuming Linked Data, Bonn, Germany (2011)
3. Hasnain, A., Fox, R., Decker, S., Deus, H.F.: Cataloguing and linking life sciences LOD Cloud. In: 1st International Workshop on Ontology Engineering in a Data-driven World collocated with EKAW12 (2012)
4. Hasnain, A., Kamdar, M.R., Hasapis, P., Zeginis, D., Warren Jr, C.N., Deus, H.F., Ntalaperas, D., Tarabanis, K., Mehdi, M., Decker, S.: Linked biomedical dataspace: lessons learned integrating data for drug discovery. In: The Semantic Web–ISWC 2014, pp. 114–130. Springer (2014)
5. Hasnain, A., Mehmood, Q., Sana e Zainab, S., Saleem, M., Warren, C., Zehra, D., Decker, S., Rebholz-Schuhmann, D.: Biofed: federated query processing over life sciences linked open data. Journal of Biomedical Semantics 8(1), 13 (2017), http://dx.doi.org/10.1186/s13326-017-0118-0
6. Hasnain, A., Zainab, S.S.E., Kamdar, M.R., Mehmood, Q., Warren Jr, C., et al.: A roadmap for navigating the life scinces linked open data cloud. In: International Semantic Technology (JIST2014) conference (2014)
7. Kamdar, M.R., Zeginis, D., Hasnain, A., Decker, S., Deus, H.F.: ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. Journal of Biomedical Informatics 47(0), 112 – 130 (2014)
8. Khan, Y., Saleem, M., Mehdi, M., Hogan, A., Mehmood, Q., Rebholz-Schuhmann, D., Sahay, R.: Safe: Sparql federation over rdf data cubes with access control. Journal of Biomedical Semantics 8(1), 5 (2017), http://dx.doi.org/10.1186/s13326-017-0112-6
9. Saleem, M., Khan, Y., Hasnain, A., Ermilov, I., Ngomo, A.C.N.: A fine-grained evaluation of sparql endpoint federation systems. Semantic Web Journal (2014)
10. Saleem, M., Ngonga Ngomo, A.C.: HiBISCuS: Hypergraph-Based Source Selection for SPARQL Endpoint Federation, pp. 176–191. Springer International Publishing, Cham (2014), http://dx.doi.org/10.1007/978-3-319-07443-6_13
11. Saleem, M., Ngonga Ngomo, A.C., Xavier Parreira, J., Deus, H.F., Hauswirth, M.: DAW: Duplicate-AWare Federated Query Processing over the Web of Data, pp. 574–590. Springer Berlin Heidelberg, Berlin, Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-41335-3_36
12. Saleem, M., Padmanabhuni, S.S., Ngomo, A.C.N., Iqbal, A., Almeida, J.S., Decker, S., Deus, H.F.: Topfed: Tcga tailored federated query processing and linking to lod. Journal of Biomedical Semantics 5(1), 47 (2014), http://dx.doi.org/10.1186/2041-1480-5-47
13. Schwarte, A., Haase, P., Hose, K., Schenkel, R., Schmidt, M.: Fedx: a federation layer for distributed query processing on linked open data. In: The Semanic Web: Research and Applications, pp. 481–486. Springer (2011)
14. Zeginis, D., et al.: A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. Semantic Web (2013)