SPARQL Query Formulation and Execution using FedViz

Syeda Sana e Zainab[1], Muhammad Saleem[2], Qaiser Mehmood[1], Durre Zehra[1], Stefan Decker[1], and Ali Hasnain[1]

[1] Insight Centre for Data Analytics, National University of Ireland, Galway
firstname.lastname@insight-centre.org
[2] Universität Leipzig, IFI/AKSW, PO 100920, D-04009 Leipzig
{lastname}@informatik.uni-leipzig.de

Abstract. Health care and life sciences research heavily relies on the ability to search, discover, formulate and correlate data from distinct sources. Although the Semantic Web and Linked Data technologies help in dealing with data integration problem, there remains a barrier adopting these for non-technical research audiences. In this paper we present FedViz, a visual interface for SPARQL query formulation and execution. FedViz facilitates both federated and non-federated SPARQL queries formulation through a visual interface and uses FedX for query execution and the results retrieval. We evaluate the usability of our system by using the Standard System Usability scale. Overall usability score of **74.16** suggests that FedViz interface is easy to learn, consistent, and adequate for frequent use.

1 Introduction

The researchers in health care, life sciences and biomedical also known as domain users, adopted Semantic Web and Linked Data technologies due to the data integration challenges faced as a result of excessive data produced [3,11]. The use of these technologies facilitated the domain users for issuing structured SPARQL queries over highly heterogeneous data spread over diverse data sources [2]. Such structured queries are important to drive meaningful biomedical correlations. Meaningful query such as *find out the drugs and their interactions with other drugs that help in curing lung cancer?* can only be answered by querying and aggregating data from more than a single data sources. Due to lack of Semantic Web expertise, writing the desired SPARQL queries to retrieve such data is a challenging task for most of the biologist or clinical researchers [6,7]. Even for the computer scientist, assembling a federated SPARQL query (i.e., the one which integrates data from multiple data sources) is a time-consuming and technical task as it heavily relies on the familiarity with the data representation and the granularity of each federated data source [9,5,8]. An alternative to this is an intuitive and interactive platform that can facilitate domain users to assemble complex meaningful SPARQL queries (can be SPARQL 1.0 federated queries as well) through visual interface. At present several visual query formulation systems including SPARQLViz [1], iSPARQL[1], ReVeaLD [9] and QueryVOWL[4] use specific graphs, available predicate links and user may need sufficient SPARQL knowledge interacting these. We introduce FedViz, an intuitive visual query interface for formulating and running complex SPARQL queries.

[1] http://oat.openlinksw.com/isparql/

FedViz allows the user to select concepts and properties from multiple datasets using nodes and edges and makes SPARQL query in a background. Furthermore, it also allow users to edit the resultant SPARQL query before sending it to the the query execution engine. The assembled SPARQL query is then executed through FedX [12], a well-known SPARQL query federation engine. Finally, the query execution results are displayed by FedViz. Note FedX can executes both federated (without using SPARQL `SERVICE` clause) and non-federated SPARQL queries. Current version of FedViz supports six real time Life Science data sources namely Kegg (http://www.genome.jp/kegg/), Drugbank (http://www.drugbank.ca/), DailyMed (http://dailymed.nlm.nih.gov/dailymed/index.cfm), Medicare (http://wifo5-03.informatik.uni-mannheim.de/medicare/), Sider (http://wifo5-03.informatik.uni-mannheim.de/sider/), and Diseasome (http://wifo5-03.informatik.uni-mannheim.de/diseasome/).

2 FedViz

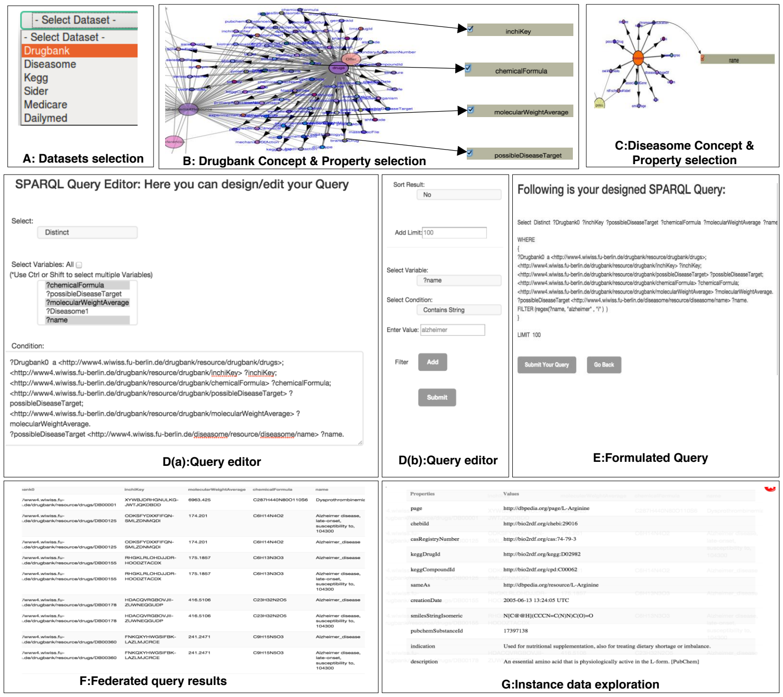
FedViz (Demo available at: http://srvgal86.deri.ie/FedViz/index.html) is a browser based application currently supported by *IE*, *Firefox* and *Chrome*, that provide users with a flexible visual interface to formulate and execute SPARQL queries. FedViz visualises six life sciences datasets and facilitate users to draw meaningful biological correlations including drug-drug interaction, drug-disease interaction, drug-side effect interaction and disease-gene interaction. FedViz supports full range of SPARQL1.1 constructs either selected through interface or can be directly added to the query in SPARQL Editor.

2.1 Query Formulation

We demonstrate our visual query formulation and execution process by using an example scenario, i.e., find the drug-disease interaction from the aforementioned data sources.
Scenario: Drug-Disease Interaction: *Drugs with their inchi key, molecular weight average and chemical structure for curing disease Alzheimer:*

This query requires data integration from Drugbank (containing drugs information) and Diseasome (containing disease information) and be formulated by using the following step-by-step approach (ref., Fig. 1):

1. The first step is to identify how Drugbank and Diseasome datasets are connected to each other? This connectivity (i.e., via classes `drugbank:drug` and `diseasome:disease`) can be found by using the Class visualisation View of FedViz.
2. User selects Drugbank from the Datasets Selection box (window A).
3. The visualisation for Drugbank dataset can be seen in Window B where user selects `drugbank:drug` class and its properties (i.e., `inchi key`, `molecular weight average` and `possible disease target`).
4. Step 2 and 3 are now followed for Diseasome dataset, i.e., select `diseasome:disease` class and it's `name` property (window C).
5. Next, FedViz SPARQL Query Editor allows user to add constraints to the formulated query such as select projection variables, apply SPARQL `LIMIT`, `FILTER`, `ORDERY BY` clauses, and can further edit the query according to his choice (window D(a) and D(b)).
6. The final query can be seen on submission (window E).



SPARQL Query Editor: Here you can design/edit your Query

Select: Distinct

Select Variables: All □
(*Use Ctrl or Shift to select multiple Variables)
?chemicalFormula
?possibleDiseaseTarget
?molecularWeightAverage
?Diseasome1
?name

Condition:

```
?Drugbank0  a <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/drugs>;
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/inchiKey> ?inchiKey;
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/chemicalFormula> ?chemicalFormula;
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/possibleDiseaseTarget> ?
possibleDiseaseTarget;
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/molecularWeightAverage> ?
molecularWeightAverage.
?possibleDiseaseTarget <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/name> ?name.
```

Sort Result: No

Add Limit: 100

Select Variable: ?name

Select Condition: Contains String

Enter Value: alzheimer

Filter

Following is your designed SPARQL Query:

```
Select  Distinct  ?Drugbank0  ?inchiKey  ?possibleDiseaseTarget  ?chemicalFormula  ?molecularWeightAverage  ?name

WHERE
{
?Drugbank0  a <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/drugs>;
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/inchiKey> ?inchiKey;
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/possibleDiseaseTarget> ?possibleDiseaseTarget;
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/chemicalFormula> ?chemicalFormula;
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/molecularWeightAverage> ?molecularWeightAverage.
?possibleDiseaseTarget <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/name> ?name.
FILTER (regex(?name, "alzheimer" , "i" ) )
}

LIMIT  100
```

ank0	inchiKey	molecularWeightAverage	chemicalFormula	name
/www4.wiwiss.fu-.de/drugbank/resource/drugs/DB00001	XYWBJDRHGNULKG-JWTJQKDBDD	6963.425	C287H440N80O110S6	Dysprothrombinemia
/www4.wiwiss.fu-.de/drugbank/resource/drugs/DB00125	ODKSFYDXXFIFQN-SMLZDNMQDI	174.201	C6H14N4O2	Alzheimer disease, late-onset, susceptibility to, 104300
/www4.wiwiss.fu-.de/drugbank/resource/drugs/DB00125	ODKSFYDXXFIFQN-SMLZDNMQDI	174.201	C6H14N4O2	Alzheimer_disease
/www4.wiwiss.fu-.de/drugbank/resource/drugs/DB00155	RHGKLRLOHDJJDR-HOODZTACDX	175.1857	C6H13N3O3	Alzheimer_disease
/www4.wiwiss.fu-.de/drugbank/resource/drugs/DB00155	RHGKLRLOHDJJDR-HOODZTACDX	175.1857	C6H13N3O3	Alzheimer disease, late-onset, susceptibility to, 104300
/www4.wiwiss.fu-.de/drugbank/resource/drugs/DB00178	HDACQVRGBOVJII-ZUWNEQGUDP	416.5106	C23H32N2O5	Alzheimer_disease
/www4.wiwiss.fu-.de/drugbank/resource/drugs/DB00178	HDACQVRGBOVJII-ZUWNEQGUDP	416.5106	C23H32N2O5	Alzheimer disease, susceptibility to, 104300
/www4.wiwiss.fu-.de/drugbank/resource/drugs/DB00360	FNKQXYHWGSIFBK-LAZLMJCRCE	241.2471	C9H15N5O3	Alzheimer_disease
/www4.wiwiss.fu-.de/drugbank/resource/drugs/DB00360	FNKQXYHWGSIFBK-LAZLMJCRCE	241.2471	C9H15N5O3	Alzheimer disease, late-onset, susceptibility to, 104300

Properties	Values
page	http://dbpedia.org/page/L-Arginine
chebiId	http://bio2rdf.org/chebi:29016
casRegistryNumber	http://bio2rdf.org/cas:74-79-3
keggDrugId	http://bio2rdf.org/kegg:D02982
keggCompoundId	http://bio2rdf.org/cpd:C00062
sameAs	http://dbpedia.org/resource/L-Arginine
creationDate	2005-06-13 13:24:05 UTC
smilesStringIsomeric	N[C@@H](CCCN=C(N)N)C(O)=O
pubchemSubstanceId	17397138
indication	Used for nutritional supplementation, also for treating dietary shortage or imbalance.
description	An essential amino acid that is physiologically active in the L-form. [PubChem]

Fig. 1: Federated Query Execution Step by Step Approach.

7. Query is executed over FedX and the retrieve results are displayed by FedViz (Result window F).
8. Finally, by selecting any URI from the retrieved result, FedViz provided detailed information regarding that entity (window G).

3 Evaluation

To assess the usability of FedViz, we used the standardised, *System Usability Scale* (SUS) [10] questionnaire[2]. The SUS is a low-cost, reliable usability scale that can be used for global assessments of systems usability[10]. The survey was filled by 15 users[3] including researchers and engineers in Semantic Web (by 10th July 2015). We achieved a mean usability score of **74.16** indicating a high level of usability according to the SUS score. The average scores (out of 5) for each question along with standard deviation is shown in Figure 2. The responses to question 1 suggests that FedViz is adequate for frequent use (average score to question 1 = 3.8 ± 0.86) by users all of type. The responses to question 7 (average score 4.06 ± 0.96) suggests that most people would learn to use this system very quickly and the responses to question 3 indicates that FedViz is easy to use (average score 4 ± 0.84).

[2] Our survey can found at: http://goo.gl/forms/bhReuNgd6O

[3] Users from AKSW, University of Leipzig and INSIGHT Centre, National University of Ireland, Galway. Summary of the responses can be found at: https://goo.gl/ZOrJx9

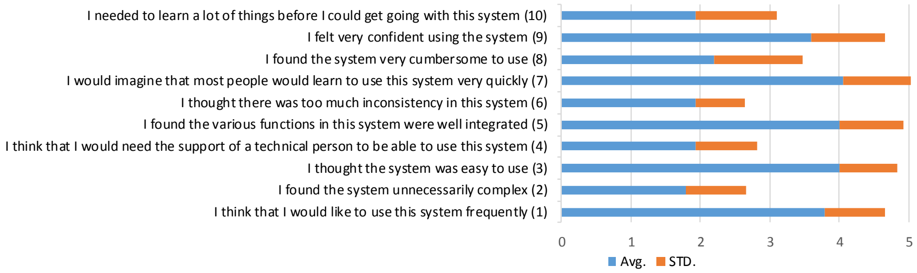


Fig. 2: Result of usability evaluation using SUS questionnaire.

4 Acknowledgement

The work presented in this DEMO paper has been partly funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

1. J. Borsje and H. Embregts. Graphical query composition and natural language processing in an rdf visualization interface. *Erasmus School of Economics and Business Economics, Vol. Bachelor. Erasmus University, Rotterdam*, 2006.
2. B. Chen, D. J. Wild, Q. Zhu, Y. Ding, X. Dong, M. Sankaranarayanan, H. Wang, and Y. Sun. Chem2bio2rdf: A linked open data portal for chemical biology. *arXiv:1012.4759*, 2010.
3. H. Chen, T. Yu, and J. Y. Chen. Semantic web meets integrative biology: a survey. *Briefings in bioinformatics*, 14(1):109–125, 2013.
4. F. Haag, S. Lohmann, S. Siek, and T. Ertl. Visual querying of linked data with QueryVOWL. In *Joint Proceedings of SumPre 2015 and HSWI 2014-15*. CEUR-WS, to appear.
5. A. Hasnain, R. Fox, S. Decker, and H. F. Deus. Cataloguing and linking life sciences LOD Cloud. In *1st International Workshop on Ontology Engineering in a Data-driven World collocated with EKAW12*, 2012.
6. A. Hasnain, M. R. Kamdar, P. Hasapis, D. Zeginis, C. N. Warren Jr, et al. Linked Biomedical Dataspace: Lessons Learned integrating Data for Drug Discovery. In *International Semantic Web Conference (In-Use Track), October 2014*, 2014.
7. A. Hasnain, Q. Mehmood, S. S. e Zainab, and S. Decker. A provenance assisted roadmap for life sciences linked open data cloud. 2015.
8. A. Hasnain, S. S. E. Zainab, M. R. Kamdar, Q. Mehmood, C. Warren Jr, et al. A roadmap for navigating the life scinces linked open data cloud. In *International Semantic Technology (JIST2014) conference*, 2014.
9. M. R. Kamdar, D. Zeginis, A. Hasnain, S. Decker, and H. F. Deus. Reveald: A user-driven domain-specific interactive search platform for biomedical research. *Journal of biomedical informatics*, 47:112–130, 2014.
10. J. R. Lewis and J. Sauro. The factor structure of the system usability scale. In *HCD*. 2009.
11. M. Saleem, S. S. Padmanabhuni, A.-C. N. Ngomo, A. Iqbal, J. S. Almeida, S. Decker, and H. F. Deus. Topfed: Tcga tailored federated query processing and linking to lod. *Journal of Biomedical Semantics*, 2014.
12. A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt. Fedx: Optimization techniques for federated query processing on linked data. In *ISWC*. 2011.