# Faceted Search over OWL 2 Life Science Datasets and Ontologies with SemFacet*

Bernardo Cuenca Grau, Evgeny Kharlamov, Šarūnas Marciuška,
Dmitriy Zheleznyakov, and Yujiao Zhou

Department of Computer Science, University of Oxford
first.middle.lastname@cs.ox.ac.uk

**Abstract.** Faceted search is the de facto query paradigm in e-commerce applications. Recently this approach was adapted to the context of the Semantic Web. In this demonstration we present our faceted search system SemFacet and show how it can enhance access to RDF and OWL 2 datasets and OWL 2 ontologies in the domain of life sciences. SemFacet combines keyword and faceted search and it is based on a solid theory, in particular it employs novel ontology projection techniques to enable faceted navigation for OWL 2. SemFacet relies on PAGOdA and HermiT for logical reasoning and on JRDFox and Sesame for storing and querying RDF triples.

## 1 Introduction

In the last decade numerous RDF datasets and OWL ontologies in the life sciences domain have become available [1–3]. Accessing the required information, however, remains a challenging task for end users and often requires proficiency in SPARQL. In order to make data and ontological knowledge more human accessible numerous query formulation, data exploration, and browsing tools have been developed. Many such interfaces have beed tailored for specific life science datasets [3, 4]. More generic systems typically rely on controlled natural language, [5, 6] diagrammatic query constructors [7, 8], or exploratory search [9].

Faceted search is the de facto query paradigm in e-commerce applications [10]. A facet typically consists of a property (e.g., 'gender' or 'occupation' when querying documents about people) and a set of possible string values (e.g., 'female' or 'research'), and documents in the collection are annotated with property-value pairs. During faceted search, users iteratively select facet values and the documents annotated according to the selection are returned as the search result.

Several authors have proposed faceted search for querying RDF, and a number of systems have been developed [11–15]. Existing systems, however, have been designed for plain RDF data, and do not take into account ontological axioms other than subsumption statements between atomic classes and properties [16, 17], with reasoning playing little or no role. In stark constrast to other domains, life sciences applications tend to require a great deal of the expressive power available in OWL 2; in particular,

* Work supported by the Royal Society, the EPSRC projects Score!, Exoda, and MaSI$^3$, and the FP7 project OPTIQUE under the grant agreement 318338.

data often involves complex class or property assertions (e.g., see FlyBase [3]) and ontologies largely consist of complex axioms which encapsulate highly valuable information for faceted search. As a result, existing faceted search systems are not well-suited for typical life sciences applications.

In [18] we developed a faceted search approach for RDF data enhanced with OWL 2 ontologies. Our solution is based on a solid theoretical framework and it addresses many of the limitations of existing techniques. To put our ideas into practice we developed SemFacet [18, 19]: a faceted search system that relies on state-of-the-art triple stores and OWL 2 reasoners to generate and update faceted query interfaces, as well as for computing search results. For demonstration purposes our platform integrates JRDFox [20] and Sesame [21] as RDF triple stores, as well as PAGOdA [22] and HermiT [23] as fully-fledged OWL 2 reasoners. Our system is fully generic and can be used to query arbitrary data and ontologies. In this demonstration we will show how SemFacet can be used to access several datasets and ontologies from the domain of life sciences and illustrate the main advantages of our approach over existing techniques designed for plain RDF.

## 2 The SemFacet System

SemFacet [24] combines keyword search and faceted navigation to query arbitrary ontology-enhanced RDF datasets. Our system offers the following main functionality.

- *Keyword search*. Seach in SemFacet typically starts with a set of keywords, which are matched against the annotations in the ontology and data.
- *Faceted interface generation and update*. SemFacet implements dedicated infrastructure for automatically generating a faceted interface from the result of a keyword search as well as for updating an interface in response to users' actions. A distinguishing aspect of our algorithms for interface generation and update is that they are 'guided' by both explicit and implicit information in the ontology and data (see [18] for details).
- *Query answering*. User selections of facet values in an interface are compiled into SPARQL queries, which are then evaluated against the ontology and data using a reasoner. Our system allows for both disjunctive facets (i.e., those where multiple value selections are interpreted disjunctively) and conjunctive facets. Thus, the SPARQL graph patters relevant to our approach can be captured by the AND-UNION fragment of SPARQL 1.1. The current version of SemFacet integrates the following reasoners: Sesame [21] (a widely used system for RDF(S) reasoning), JRDFox [20] (a parallel in-memory RDF triple store supporting sound and complete reasoning for OWL 2 RL), HermiT [23] (a standard fully-fledged OWL 2 reasoner), and PAGOdA [22] (a pay-as-you-go reasoner for OWL 2 that combines JRDFox and HermiT for increased efficiency).
- *Refocusing*. SemFacet provides functionality for changing the focus of the search from one type of object to another. For instance, if the system is displaying as search results neurons that develop from cells, where "develops from" is a facet name and "cell" is a facet value, we can refocus the search and display as search results the particular cells that are related to the selected neurons.

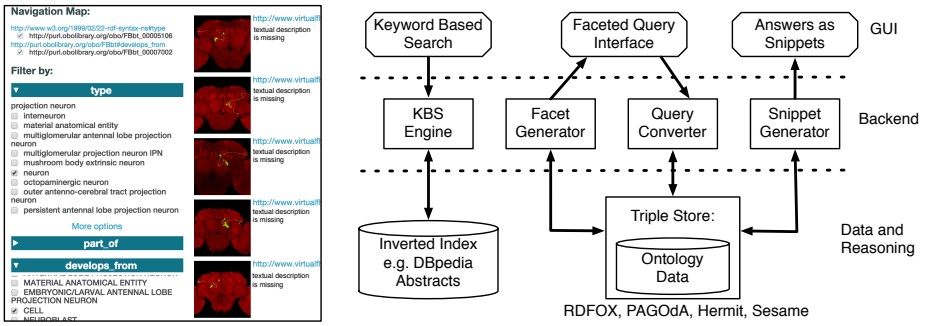


**Fig. 1.** Left: screenshot of SemFacet over FlyBase OWL 2 data, Right: architecture of SemFacet

– *Customisation*. Our system is generic and highly customisable for different datasets and applications. Users can upload arbitrary ontologies and datasets, select the reasoner to be exploited for faceted navigation and query answering, customise the kinds of annotations relevant for keyword search, select which facets should be interpreted disjunctively or conjunctively as well as which facets should be excluded from the search process, or select what properties are relevant for image thumbnails and snippets (if any).

On the left-hand-side of Figure 1 we can see a screenshot of SemFacet with a search over the Adult Brain Anatomy dataset [1]. The *navigation map* in the interface enables refocusing, the *filter by* section displays the relevant facet names and values, and search results (i.e., query answers) are displayed on the rightmost part of the interface. The general architecture of SemFacet including its main software components is summarised on the right-hand-side of Figure 1.

## 3 Demonstration Scenarios

During the demonstration we will show how to explore and query OWL 2 life science datasets and ontologies with SemFacet. To this end, we will preconfigure the system for several test cases, including fragments of FlyBase [3], SNOMED CT [2], as well as a selection of Bio2RDF [1] datasets. In all cases the input for the search will be a dataset and an ontology. We will demonstrate the following variants of our algorithms for interface generation and update.

– *Data driven*, where only the data is exploited for interface generation and update. This configuration simulates existing approaches to faceted search over RDF.
– *Ontology driven*, where only the axioms in the ontology are considered. In this configuration, facet names and values in an interface reflect semantic relationships between entities in the input ontology.
– *Both data and ontology driven*, where both the data and ontology are exploited in interface generation and update. This is the default configuation of SemFacet, and the aim here is to show how reasoning and ontologies can improve data driven faceted interfaces and allow for enhanced data exploration.

Besides querying preconfigured scenarios, the demo attendees will be able to try SemFacet end-to-end. This would require to load a data set and ontology, to customise the system parameters, and to query the uploaded ontology and data with the selected parameters. For the end-to-end test of SemFacet the demo attendees will be able to use datasets and ontologies either from the preconfigured scenarios or the ones they provide (of reasonable size), e.g., by downloading them from the Web.

## 4 References


[1] F. Belleau, M. Nolin, N. Tourigny, et al. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. In: *Journal of Biomedical Informatics* 41.5 (2008).

[2] *SNOMED CT*. `www.ihtsdo.org/snomed-ct`.

[3] *FLyBase*. `http://flybase.org/`.

[4] N. Milyaev, D. Osumi-Sutherland, S. Reeve, et al. The Virtual Fly Brain browser and query interface. In: *Bioinformatics* 28.3 (2012).

[5] E. Franconi, P. Guagliardo, M. Trevisan, and S. Tessaris. Quelo: an Ontology-Driven Query Interface. In: *DL*. 2011.

[6] A. Bernstein, E. Kaufmann, A. Göhring, and C. Kiefer. Querying Ontologies: A Controlled English Interface for End-Users. In: *ISWC*. 2005.

[7] D. Calvanese, C. M. Keet, W. Nutt, et al. Web-based graphical querying of databases through an ontology: the Wonder system. In: *SAC*. 2010.

[8] A. Soylu, M. G. Skjæveland, M. Giese, et al. A Preliminary Approach on Ontology-Based Visual Query Formulation for Big Data. In: *MTSR*. 2013.

[9] S. Ferré and A. Hermann. Semantic Search: Reconciling Expressive Querying and Exploratory Search. In: *ISWC*. 2011.

[10] D. Tunkelang. *Faceted Search*. Morgan & Claypool Publishers, 2009.

[11] P. Fafalios and Y. Tzitzikas. X-ENS: Semantic Enrichment of Web Search Results at Real-Time. In: *SIGIR*. 2013.

[12] R. Hahn, C. Bizer, C. Sahnwaldt, et al. Faceted Wikipedia Search. In: *BIS*. 2010.

[13] D. F. Huynh and D. R. Karger. Parallax and Companion: Set-based Browsing for the Data Web. 2013.

[14] P. Heim, J. Ziegler, and S. Lohmann. gFacet: A Browser for the Web of Data. In: *IMC-SSW*. 2008.

[15] G. Kobilarov and I. Dickinson. Humboldt: Exploring Linked Data. In: *LDOW*. 2008.

[16] M. Hildebrand, J. van Ossenbruggen, and L. Hardman. /facet: A Browser for Heterogeneous Semantic Web Repositories. In: *ISWC*. 2006.

[17] E. Oren, R. Delbru, and S. Decker. Extending Faceted Navigation for RDF Data. In: *ISWC*. 2006.

[18] M. Arenas, B. C. Grau, E. Kharlamov, et al. Faceted Search over Ontology-Enhanced RDF Data. In: *CIKM*. 2014.

[19] M. Arenas, B. C. Grau, E. Kharlamov, et al. SemFacet: semantic faceted search over yago. In: *WWW, Companion Volume*. 2014.

[20] *RDFox*. `www.cs.ox.ac.uk/isg/tools/RDFox/`.

[21] *Sesame*. `http://www.openrdf.org/`.

[22] Y. Zhou, Y. Nenov, B. C. Grau, and I. Horrocks. Complete Query Answering over Horn Ontologies Using a Triple Store. In: *ISWC*. 2013.

[23] B. Glimm, I. Horrocks, B. Motik, et al. HermiT: An OWL 2 Reasoner. In: *Journal of Automated Reasoning* 53.3 (2014).

[24] *SemFacet*. `http://www.cs.ox.ac.uk/isg/tools/SemFacet/`.