chem-bla-ics

Scholia configurability

Egon Willighagen

Published August 23, 2024

Citation
Willighagen, E. (2024, August 23). Scholia configurability. *Chem-bla-ics*. https://doi.org/10.59350/epanj-4t315

Keywords
Scholia, Wikidata, Sparql

Abstract
Scholia is a visual layer on top of Wikidata providing a rich user experience for browing scholarly research related knowledge. I am using the combinatie for various things, including exploring new research topics (a method, compound, or protein I do not know so much about yet), indexing notable research output (including citations), progress of Citation Typing Ontology uptake, etc.

Scholia is a visual layer on top of Wikidata providing a rich user experience for browing scholarly research related knowledge. I am using the combinatie for various things, including exploring new research topics (a method, compound, or protein I do not know so much about yet), indexing notable research output (including citations), progress of Citation Typing Ontology uptake, etc. This weekend I hope to send around the final draft for the *Scholia Chemistry* paper.

Scholia has received a fair share of scholarly and social attention. The Scholia paper has been cited over 100 times and the websites received about 200 thousand page views each day (though we do not know how to get Toolforge to give us sufficient insight into the how and what of that count). There is a Wikipedia template to link to Scholia and some of projects I am involved in link Scholia for articles, such as WikiPathways.

With that, there is also interest in using it for other Wikibases and perhaps even random SPARQL endpoints. These things are not trivial, as Scholia uses complementary APIs, various URL patterns for some of the functionality, and generally, all SPARQL queries are tweaked to the Wikidata Blazegraph SPARQL endpoint to ensure results are returned in reasonable time. But that last requires use of Blazegraph extensions to the SPARQL standard.

All this requires Scholia to become more independent, in a better model-view-controller model. And that actually turns out very important at this moment. That is, Wikidata is not a RDF-first database, but a Wikibase-based store. Whenever an edit is made, RDF is generated and the SPARQL endpoint is updated. Now, the number of edits in Wikidata is enormous and the notion that the SPARQL endpoint is often minutes at most behind is a huge accomplishment. But the Blazegraph platform cannot keep up with Wikidata. Blazegraph is open source, but has been bought up and development stopped from one day to another.

Therefore, a split of the Wikidata SPARQL platform is planned. This split will put one part of the knowledge in on endpoint and the other half in the other. Any query that needs information from both graphs, will have to do a federated SPARQL query. Basically, there are very few Scholia queries that do not rewriting. My first rewrite actually failed, because the rewriting is not obvious and quickly times out. To some extend, this is because now lots of results of subqueries need to be send over the network from one endpoint to the other. When the combined query basically covers half of each endpoint, that's a lot of network traffic.

An immediate use case of the configuration is therefore running Scholia against the current three endpoints: the current official endpoint, and the two split endpoints under development. With a recent patch Finn and I worked on, this configuration looks like this (and saved as `scholia.ini`:

```
[query-server]
# Wikidata:
#sparql_endpoint = https://query.wikidata.org/sparql
#sparql_editurl = https://query.wikidata.org/#
#sparql_embedurl = https://query.wikidata.org/embed.html#
```

chem-bla-ics

```
# Wikidata Split Main
sparql_endpoint = https://query-main.wikidata.org/sparql
sparql_editurl = https://query-main.wikidata.org/#
sparql_embedurl = https://query-main.wikidata.org/embed.html#

# Wikidata Split Scholar
#sparql_endpoint = https://query-scholarly.wikidata.org/sparql
#sparql_editurl = https://query-scholarly.wikidata.org/#
#sparql_embedurl = https://query-scholarly.wikidata.org/embed.html#
```

So, right now, we can test the impact of the split with Scholia and this patch. We would fire up a local instances of Scholia, running against one of the split endpoints, and use the Toolforge instance as baseline.

Now, on my system I need to use Python virtualenv so, I first start a Scholia `venv`:

```
source ~/.venvs/scholia/bin/activate
```

After that, I can select an other endpoint, e.g. the `main` Wikidata split endpoint (`query-main-experimental.wikidata.org`) were it not they are currently offline as part of the transition and run Scholia on a unique port:

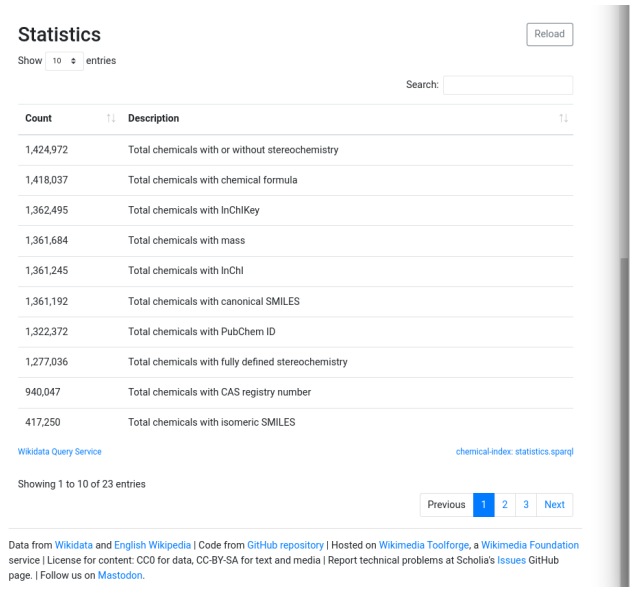
```
scholia run
```

Then I can have two browser windows along side and compare Scholia pages againt the current Scholia instance and when running against another SPARQL endpoint. For now, I can test how well Scholia runs on the QLever instance of Wikidata (superfast and updated data once a week). Here the configuration I have is not entirely complete, and many SPARQL queries do not work against QLever, including anything with graphical depiction. But that said, I can use this configuration:

```
[query-server]
# QLever
#sparql_endpoint = https://qlever.cs.uni-freiburg.de/api/wikidata
#sparql_editurl = https://qlever.cs.uni-freiburg.de/wikidata/?query=
#sparql_embedurl =
```

Then, I can compare, for example, the chemicals statistics the main Scholia with one running against QLever:

chem-bla-ics



Statistics

Reload

Show 10 entries

Search:

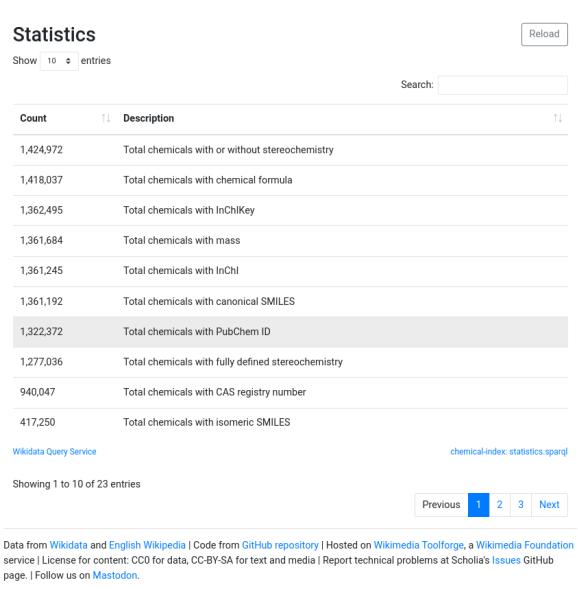
Count	Description
1,424,972	Total chemicals with or without stereochemistry
1,418,037	Total chemicals with chemical formula
1,362,495	Total chemicals with InChIKey
1,361,684	Total chemicals with mass
1,361,245	Total chemicals with InChI
1,361,192	Total chemicals with canonical SMILES
1,322,372	Total chemicals with PubChem ID
1,277,036	Total chemicals with fully defined stereochemistry
940,047	Total chemicals with CAS registry number
417,250	Total chemicals with isomeric SMILES

Wikidata Query Service chemical-index: statistics.sparql

Showing 1 to 10 of 23 entries

Previous 1 2 3 Next

Data from Wikidata and English Wikipedia | Code from GitHub repository | Hosted on Wikimedia Toolforge, a Wikimedia Foundation service | License for content: CC0 for data, CC-BY-SA for text and media | Report technical problems at Scholia's Issues GitHub page. | Follow us on Mastodon.



Statistics

Reload

Show 10 entries

Search:

Count	Description
1,424,972	Total chemicals with or without stereochemistry
1,418,037	Total chemicals with chemical formula
1,362,495	Total chemicals with InChIKey
1,361,684	Total chemicals with mass
1,361,245	Total chemicals with InChI
1,361,192	Total chemicals with canonical SMILES
1,322,372	Total chemicals with PubChem ID
1,277,036	Total chemicals with fully defined stereochemistry
940,047	Total chemicals with CAS registry number
417,250	Total chemicals with isomeric SMILES

Wikidata Query Service chemical-index: statistics.sparql

Showing 1 to 10 of 23 entries

Previous 1 2 3 Next

Data from Wikidata and English Wikipedia | Code from GitHub repository | Hosted on Wikimedia Toolforge, a Wikimedia Foundation service | License for content: CC0 for data, CC-BY-SA for text and media | Report technical problems at Scholia's Issues GitHub page. | Follow us on Mastodon.

This query ran without modification. For other queries rewriting is needed, but with this setup we can at least quickly see the differences in the results.