chem-bla-ics

NMRShiftDB enters rdf.openmolecules.net #2: SPARQL end point with Virtuoso

Egon Willighagen

Published September 4, 2009

Citation

Willighagen, E. (2009). NMRShiftDB enters rdf.openmolecules.net #2: SPARQL end point with Virtuoso. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/nv925-tje87

Keywords

Rdf, Sparql, Nmrshiftdb, Cheminf

Abstract

About 6 months ago I reported about my efforts to RDF-ize the data from the NMRShiftDB.

Copyright

About 6 months ago I reported about my efforts to RDF-ize the data from the NMRShiftDB. Since then, time was consumed by many other things, but now that Bioclipse can query SPARQL end points, that I want to contribute the triple set (it is GNU FDL-licensed) to Bio2RDF, that a student started working in my group (now larger than just me :) on reasoning on life sciences data, and that I recently contributed my 1000th NMR spectrum to the database, I thought it was time to finally reinstall Virtuoso.

There are precompiled binaries for Ubuntu and Debian, but Michel encouraged me to use version 6 when he visited us. And so I compiled and install 6.0.0.TP1 on the public server, while I do have the binary debs for 5.0.12 on my laptop. With some basic Apache magic, I hooked up the SPARQL end point of the server to the web:

```
<Proxy /nmrshiftdb/sparql>
  RewriteEngine On
  Allow from all
  ProxyPass        http://localhost:8890/sparql
  ProxyPassReverse http://localhost:8890/sparql
</Proxy>
```

Nice thing about this is, that I can set up multiple servers, allowing me to keep incompatibly licensed data sets apart (see Open Data: license, rights, aggregation, clean interfaces?), which is the same approach Bio2RDF is taking.

The end point now offers about 278887 triples, but this will soon rise as I make more content from the database available in the original SQL database. The data is from the 1.3.3 release by Chris' team, and does not include my 1000th spectrum.

Getting the data into the database was not trivial either. The documentation suggests WebDAV, and that indeed worked for me once, after using the curl approach suggested here. But upon a second upload, it did again not enter the store. The ultimate solution was to use the iSQL interface, with the following SQL

```
DB.DBA.RDF_LOAD_RDFXML_MT(
  file_to_string_output('/tmp/nmrshiftdb.rdf'), '',
  'http://pele.farmbio.uu.se/nmrshiftdb'
);
```

Scientifically, this progress is not overly interesting, although it makes it very clear that you really should not have to be happy with proprietary and non-semantic formats for anything. But, to me, this is mostly a technological success of great importance: I can now share really large sets of RDF data.

Querying this data is a simple with SPARQL, and the results are available in various formats, such as JSON, which makes it easy to integrate in third-party applications or Google Wave robots (did I hear someone say NMRShifty?). As I have blogged before, SPARQL is an excellent tool to aggregate scientific data prior to data analysis. And I will demo more interesting queries later this month.