chem-bla-ics

Script logs as HTML+RDFa: mix free text reporting with CSV

Egon Willighagen

Published July 19, 2010

Citation
Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/tb3sc-ggw42

Keywords
Html, Rdf, Sparql, Rdfa

Copyright
Copyright © Egon Willighagen 2010. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Richard (Talis) wrote up a three-step tutorial on how to publish your data. I think I would be more than happy if scientists reached step 1. Related, Ola asked me a while ago if I was interested in using the computing facilities of UPPMAX, and I was. But until this weekend I did not have the time or energy to give it a spin. If you are puzzled how the heck I see those two items related, read on :)

Two days later, today, I ran my first analysis. Still a test run, but using the CDK to perceive atom types on the first 2.5 GB of PubChem data. The full data set is now 80 GB, and I will start doing this analysis today. You might remember this already two years ago (see Wicked chemistry and unit testing) for a small subset, but only now have the power to analyze all compounds. The UPPMAX system I work on has 348, each with 8 cores. Each core has 3 GB of memory, but I am using the IteratingPCCompoundXMLReader class anyway. Analyzing the 2.5 GB of data was done using 50 nodes, and finished in about a minute. Nice :)

Now, this first run dumped the results as a plain text file, looking like:

```
CID 200234: Ti  1
CID 200235: Ti 1
CID 200237: Sb 1 Sb 2
CID 200365: S 1
CID 200761: Hg 1
CID 201374: Ce 1 Ce 2
CID 201395: As 1
```

Simple and effective.

Or? And this is where the two items outlined in the first paragraph meet. No, this is not useful. Since the output is from an analysis of PubChem, I'm sure you already figured out that the first two columns indicate the compound being analyzed. You might also work out that then the elements are given for which the atom type perception failed. You may even figure out that the number is likely to be the index in the connection table representation of the molecule. Right?

But what about machine readability? I could, of course, write the output as CSV, but then I would loose my ability to write the report in human readable format. And moreover, the list of failing atom types does not have a fixed length, as you can see in the example lines given earlier.

Now, this is where RDF comes in. If I create my output as HTML+RDFa, I can do fancy stuff. My results page could link directly to PubChem, so that I can inspect the actual compound. Though I could do that even with merely HTML. But with RDFa, I can actually make my free text log output machine readable. I can accurately annotate what bits are informative:

```
<div about="#200234" typeof="um:Compound">CID
  <span property="um:cid" datatype="xsd:integer">200234</span>:
  <span rel='um:hasProblem'>
    <span about='#error0' typeof='um:Problem'>
```

```
      <span property='um:hasElement'>Ti</span>
      <span property='um:hasIndex' datatype='xsd:integer'>1</span>
    </span>
  </span>
</div>
```

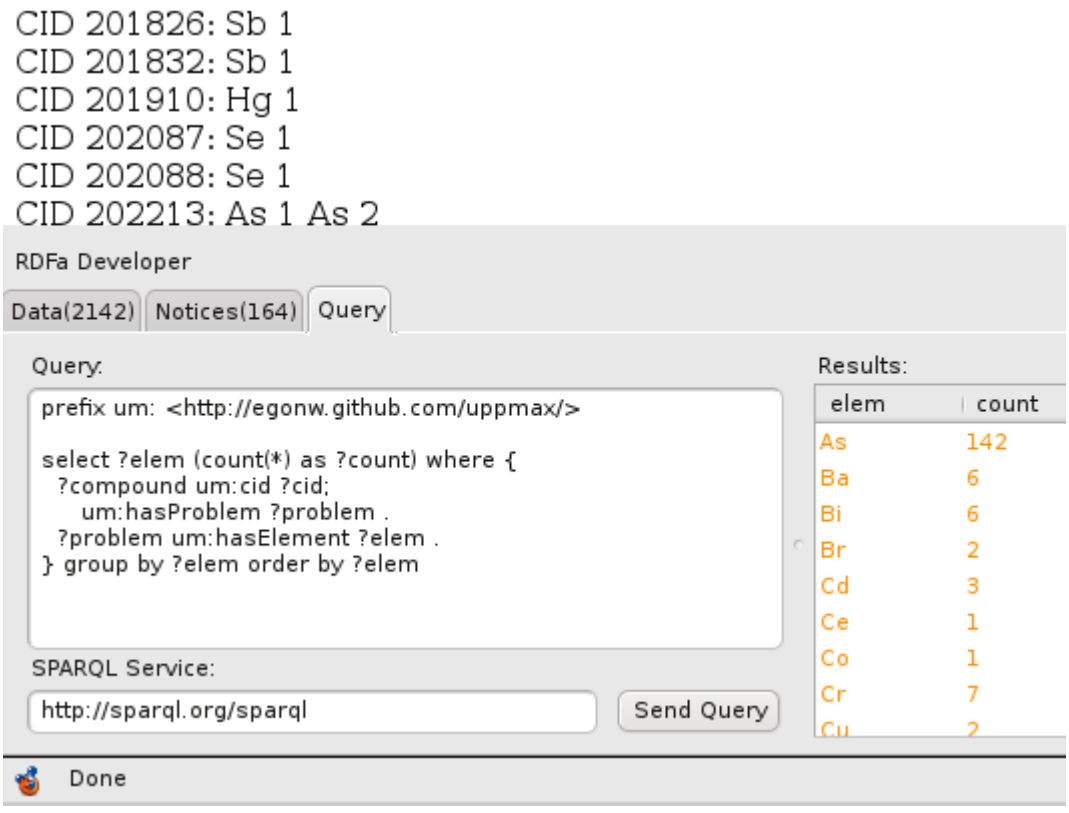
The file is not backed up by an OWL ontology, but where possible one would do that. Reuse of ontologies is a good thing (e.g. use a service like Schemapedia).

Now, I can easily open up this file in a web browser (follow this link) and get the same view as above. But I can also import the file directly into Bioclipse (see Semantic Web features in Bioclipse 2.2), or in any other tool that supports RDFa. I can then use SPARQL to do some first analysis, for example, with:

```
PREFIX um: <http://egonw.github.com/uppmax>

SELECT ?elem (count(*) AS ?count) WHERE {
  ?compound um:cid ?cid;
     um:hasProblem ?problem .
  ?problem um:hasElement ?elem .
} GROUP BY ?elem ORDER BY ?elem
```

Combine that with the RDFaDev tool I wrote about last week (see RDFaDev: HTML+RDFa development with FireFox). Now you should get some feeling of the advantages of using Open Standards: I can do some initial analysis of the results, just right there in the web browser you have open anyway:



Therefore, next time you ask your data analyst to perform some calculation, insist that he sends you HTML+RDFa log files with results. Better, ask him to put it online, and you immediately reach Step 3 in the analysis by David.