

Extracting RDF from Chem4Word documents

Egon Willighagen 

Published January 21, 2010

Citation

Willighagen, E. (2010). Extracting RDF from Chem4Word documents. In *chem-bla-ics*. chem-bla-ics. <https://doi.org/10.59350/jc9ak-pdr79>

Keywords

Cml, Java, Rdf, Chem4word, Strigi

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.

chem-bla-ics

Joe has released the first [Chem4Word demo file](#), and has written about how to [extract the CML with Java](#) and [with C#](#).

I haven't actually gotten around to fiddling with Java, but ran [Strigi](#) against it to extract RDF, while having the [Strigi-Chemistry](#) plugins installed. This is part of the [RDF](#) that came out:

```
<example-doc.docx>
  <http://freedesktop.org/standards/xesam/1.0/core#title>
    "acetic acid",
    "(8R,9S,10R,13S,14S,17S)- 17-hydroxy-10,13-dimethyl- 1,2,6,7,8,9,11,12,14,15,16,17-d
    "testosterone";
  <http://freedesktop.org/standards/xesam/1.0/core#version>
    "2",
    "2";
  <http://rdf.openmolecules.net/0.9#atomCount>
    "8",
    "49";
  <http://rdf.openmolecules.net/0.9#bondCount>
    "7",
    "52";
  <http://rdf.openmolecules.net/0.9#molecularFormula>
    "C2H4O2",
    "C19H28O2";
```

I believe there is quite some room for improvement, but it's a start :) Thanx to Joe for posting the public domain test file, so that other projects can start play with the exiting new technology. I should note, however, that I am not running a Microsoft OS nor MS-Word, and the saved documents source are the only way I have access to the [CML](#) right now.