

chem-bla-ics

Chemical Archeology: OSCAR3 to NMRShiftDB.org

Egon Willighagen 

Published September 8, 2006

Citation

Willighagen, E. (2006). Chemical Archeology: OSCAR3 to NMRShiftDB.org. In *chem-bla-ics*. chem-bla-ics. <https://doi.org/10.59350/am2k8-ygc58>

Keywords

Oscar, Bioclipse, Acs, Chemistry, Textmining

Copyright

Copyright © Egon Willighagen 2006. 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

Chemical Archeology (see [Christoph's comment](#)) is the process of extracting chemical information from old journal articles. Some time ago, [Peter Corbett](#) from the group of [Peter Murray-Rust](#) visited the [CUBIC](#) to talk to us about [Oscar3](#) which can do just that. That day, we already [hooked OPSIN into Bioclipse](#) .

Oscar3, however, is capable of more than the name2structure of OPSIN (see also [10.1039/b411033a](#); it can take a plain text file with an experimental section with details on the synthesis of small organic compounds, and analyze the chemistry in that. This functionality has been available as [an RSC authoring tool](#) for some time now (see also [10.1039/b411699m](#)).

Unfortunately, what publisher put online (PDF and HTML) is much more difficult to process with Oscar3: those formats are often optimized for display, not for machine processing. The HTML can be cleaned up, but there is no general approach.

[Christoph Steinbeck](#) is going to present at the [upcoming ACS meeting](#) the use of Oscar3 for extraction of NMR spectra from old journal article, in preparation for submission to the [NMRShiftDB.org](#) (see the [abstract](#) of [CINF 101](#)).

Since the full Oscar3 was not hooked into [Bioclipse](#) yet, I had some work to do. It took me some time to figure out how to properly configure Oscar3, and what additional things I had to do to clean up the HTML used by publishers to get Oscar3 to extract NMR spectra (thanx to PeterC for hints!). I also had to tweak the Oscar3 code itself here and there, but that's what opensource is about :) (Peter, if you are reading this: I have a number of patches for the Oscar3 code in [bc_oscar](#); let me know if you're interested in them.)

This is the end result:

The screenshot shows the Bioclipse interface. On the left, the BioResource Navigator displays a file tree with folders like 'eupatoriochromene.cml', 'furan.cml', and 'jo060803q.scixml'. The main window shows the XML structure of a molecule and its 2D chemical structure. The XML structure is as follows:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<molecule id="m1" xmlns="http://www.xml-cml.org/>
  <atomArray>
    <atom id="a1" elementType="C" formalCharge:
    <atom id="a2" elementType="C" formalCharge:
    <atom id="a3" elementType="C" formalCharge:
    <atom id="a4" elementType="O" formalCharge:
    <atom id="a5" elementType="C" formalCharge:
    <atom id="a6" elementType="C" formalCharge:
    <atom id="a7" elementType="C" formalCharge:
    <atom id="a8" elementType="C" formalCharge:
    <atom id="a9" elementType="C" formalCharge:
```

The 2D structure shows a bicyclic system with a cyclooctane ring fused to a cycloheptane ring, and a phenylpropanal side chain. The bottom console shows log messages for file loading and parsing:

```
Bioclipse console
'jo060803q.scixml' loaded and parsed successfully.
'ir.spectrum1.cml' loaded and parsed successfully.
'hrms.spectrum1.cml' loaded and parsed successfully.
'ir.spectrum1.cml' loaded and parsed successfully.
Parsed file in 9 ms.
'3-(2-Oxocyclooctanyl)-3-phenylpropan-1-al.cml' loaded and parsed successfully.
```

Note especially the hierarchy in the resource navigator on the left. The misc folder contains all the chemistry found in the article. But more importantly is that for six molecules it fully detected the experimental section! For 3-(2-Oxocyclooctanyl)-3-phenylpropan-1-al (InChI=1/C17H22O2/c18-13-12-15(14-8-4-3-5-9-14)16-10-6-1-2-7-11-17(16)19/h3-5,8-9,13,15-16H,1-2,6-7,10-12H2) it derived the molecular structure (with OPSIN), and a few spectra: H-NMR, high-resolution MS and IR.

So, if you attend the ACS meeting: make sure to visit Christoph's CIN 101 presentation!