chem-bla-ics

Open{Data|Source|Standards} is not enough: we need Open Projects

Published November 7, 2008

Citation
Willighagen, E. (2008, November 7). Open{Data|Source|Standards} is not enough: we need Open Projects. *Chem-bla-ics*. https://doi.org/10.59350/kmk49-mj610

Keywords
Odosos, Chemspider, Workflow, Cdk, Bioclipse

Abstract
The Blue Obelisk mantra ODOSOS, Open Data, Open Source, Open Standards, is well known, and much cited too. Jean-Claude Bradley popularized the Open Notebook Science (ONS). This has always been nagging me a bit, because the CDK, Jmol, JChemPaint and other chemistry projects have done that for much longer, though we did not use notebooks as much, so called it just an open source project.

The Blue Obelisk mantra ODOSOS, Open Data, Open Source, Open Standards, is well known, and much cited too. Jean-Claude Bradley popularized the Open Notebook Science (ONS). This has always been nagging me a bit, because the CDK, Jmol, JChemPaint and other chemistry projects have done that for much longer, though we did not use notebooks as much, so called it just an open source project. It really is no different, IMO, though surely, there are differences.

Anyway, the key thing which ONS and CDK and Jmol share, is that they use an Open Notebook. Not every Open Source or Open Data project does. Actually, many scientific Open Source are not open Projects! They are more like the Cathedral than the wished-for Bazaar (see The Cathedral and the Bazaar). So, Open Source (science) projects are certainly not ONS projects by default!

Now, the CDK actually is ONS, it is a Bazaar. The notebooks we use include:

- open project via mailing lists
- open methods/results via subversion
- informal reporting via blogs (e.g. Rajarshi, Christoph, Thomas, mine)
- informal reporting via CDK News

What more would you wish for? That's not a rhetorical question. Remember that every reader of this blog is in my advisory board !

Unfortunately, I do not create work at a workbench myself, so I do not produce new knowledge myself, other than extracted from existing data. That's really a shame, and I really do hope that Jean-Claude or Cameron will send me a box to measure solubilities (see here, here, and here, here for first data exploration), even though I cannot participate in the challenge. (hint, hint :)

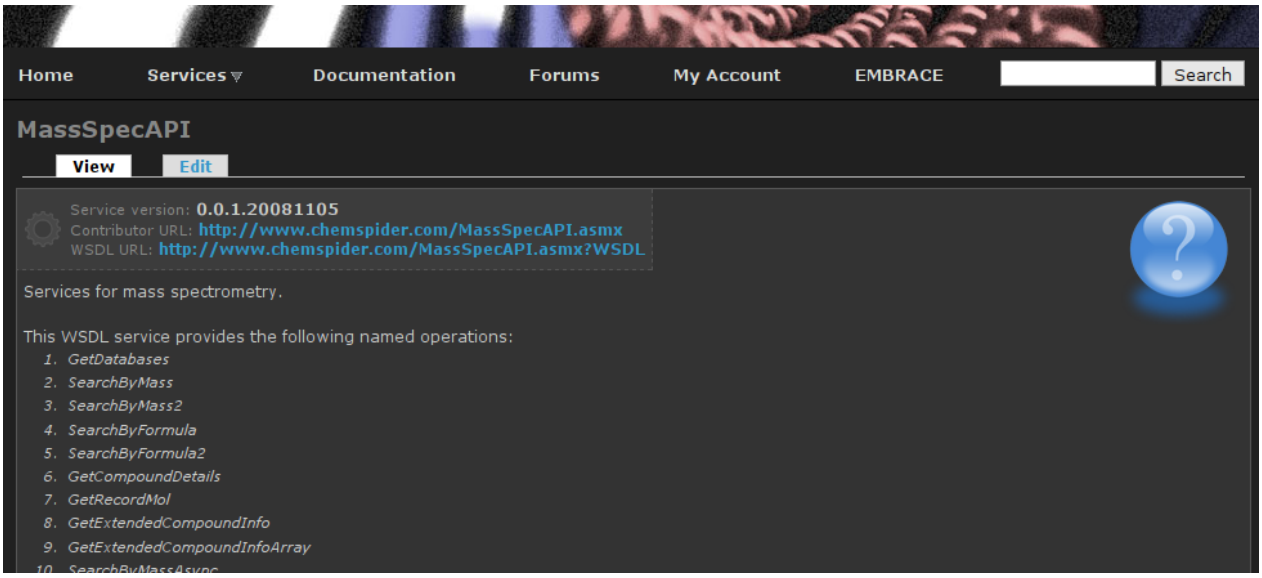
From Cathedral to Bazaar in Life Sciences

One Cathedral we ran into with Bioclipse was BioCatalogue, which will serve as website where people can annotate and categorize (web) services. While the project has been around for a while, the website was rather uninformative. Fortunately, the projects is going to open up, and be more Bazaar-like. For example, they now started a wiki and a mailing list. I hope these efforts will continue, so that I can contribute from my point of view!

The EMBRACE Registry is a project with similar goals and a rather nice outcome (which I learned about on Monday). It is actually anticipate to be replaced by or merge with BioCatalogue. So, all data I entered, cheminformatics workflows (look, no 'o'), will later be available from BioCatalogue too. That is already my first contribution to BioCatalogue. One enormously interesting feature of the Registry, is that is allows uploading of code to test the service. This will mean the Registry will not only poll if the service is still online (by checking the WSDL file), it will also test if the service behaves properly. Now, immediate thoughts are mashups with MyExperiment. Each WSDL entry in the Registry points to MyExperiment workflows that use them, and the workflow page would indicate the status of all used WDSL services. This integration was already anticipated long before I thought about it, as the involved Cathedrals were nicely located in the same floor in Manchester.

chem-bla-ics

Below is a screenshot from the EMBRACE Registry for the ChemSpider WDSL entry for a workspace I uploaded about a year ago to MyExperiment:



BTW, ChemSpider has an Advisory Board of which I am member, but it is also a classical (and intentional) Cathedral project. We do share common interests though, which makes us collaborate.

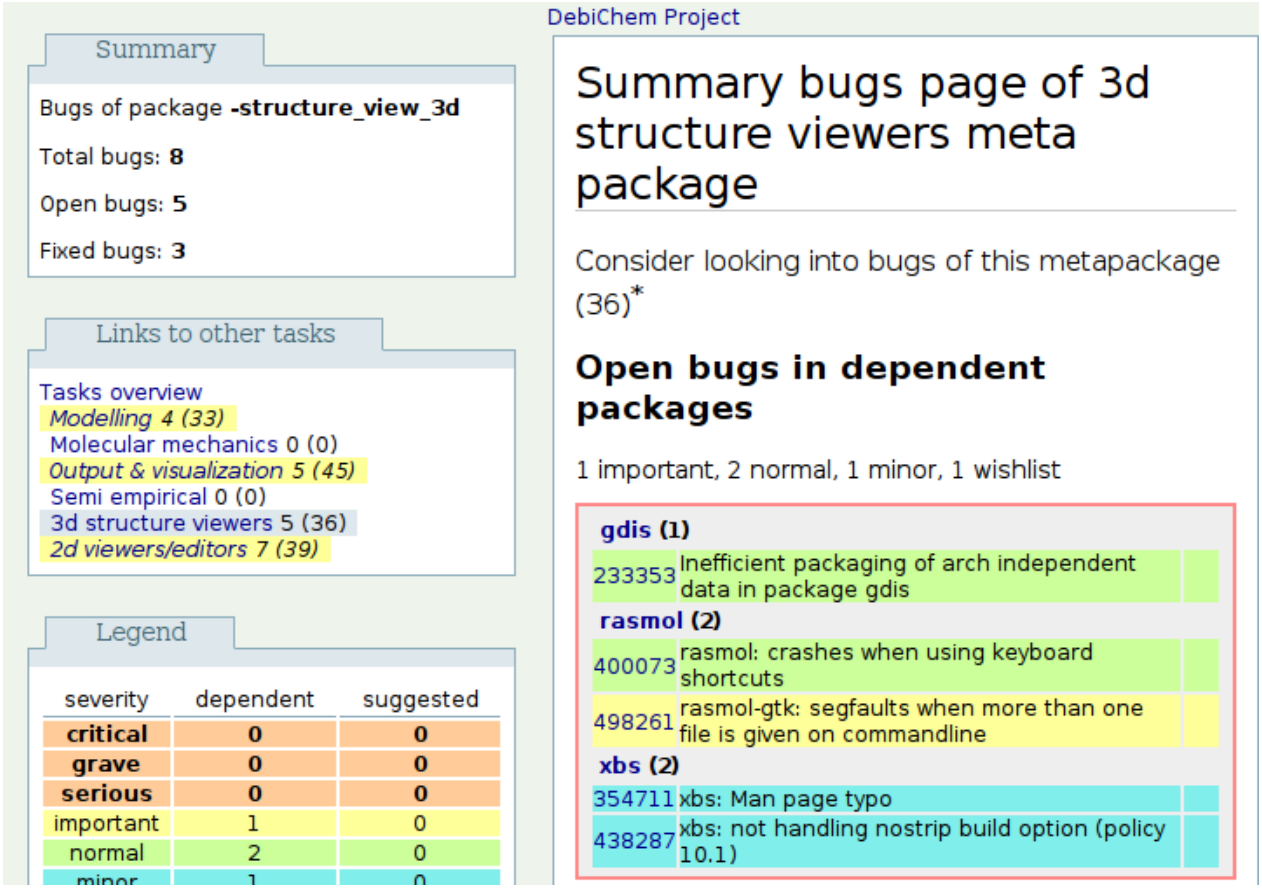
Why Important?

One recurrent theme in Open Source is given enough eyeballs, all bugs are shallow. This surely applies to science as well. The difference between the two is that in current science the eyes only inspect with a delay of at least 6 months. Current practice is that research is finished (delay), and when decided publishable written up a paper (delay, and loosing valuable information in the process, as you can read in my blog all the time), and published (even more delay). ONS changes that, and so do Bazaar-like open source projects, such as the CDK, Jmol and Bioclipse. They bugs are present, whether we like it or not, not just in source code, but in science too. Theories get overthrown, but why should we like the long delays current scientific good practice? Hate it! Work around it. Use the Bazaar. Use ONS!

Now, ONS actually needs Open Source, allowing them to deal effectively with the data they produce; to allow extraction of new scientific knowledge from the measurements. If Rajarshi and Pierre would not have made their efforts, other could not easily join in, leading to those much hated delays. Bugs should be shallow, and openness allows us to make those bugs visible. We can prove that there is a bug, without having to reproduce data ourselves, leading to those nasty delays again. Just copy the data, compare it to your own, do your analysis.

One recent project in open source chemistry dealing with making bugs visible, is the web page set up by Andreas Tille for the DebiChem project. His page summarizes the bugs listed for the chemistry in Debian (which includes the Blue Obelisk projects Avogadro, BODR, CDK, Chemical MIME Data, Kalzium and OpenBabel):

chem-bla-ics



This data analysis helps the projects being analyzed.

Packaging

This brings me to a last topic, for this blog: packaging using Open Standards. In order to allow those eyeballs to spot bugs, it is of the utmost importance to package your results in Open Standards, and not just one, but likely many. For Open Source projects this ultimately means Distribution Packages (deb or rpm). If that goal has been achieved, you know your results can be read by anyone. Software should be installable (make, ant, cmake, etc), and Data should be readable (no PDF, but RDF, XML, JSON, or whatever standard). Preferably not Excel, as this is too free format (as Rajarshi also indicated), but with some added conventions it may do well. Blue Obelisk project are generally doing well in terms of packaging.

For the CDK, which already is reasonably well packaged, I am currently working on Eclipse and Maven2 packages. The former is already being used by Bioclipse, while the second aims at Jumbo (which has just seen a new release. Jim, I'm happy to see the CMLDOM/Jumbo split!), CDK-Taverna, and possibly a third (Paula, what for do you plan to use it?). The POM export is not fully working yet, but with four research sites involved in this Open Project, I'm sure we'll work it out.

The bottom line is, scientific progress would benefit so much from a Bazaar approach. And the key thing is not collaboration; that's something you can do in a Cathedral-like fashion too. No,

the key thing is to be Open and allow anyone, even your worst nightmare, to comment on what you do. Let him prove you wrong, openly, that is.

OK, there it is. My open notebook entry for this week. Now you know what I have been up to this week.