chem-bla-ics

The Molecular Chemometrics Principles #2: be clear in what you mean

Egon Willighagen

Published August 12, 2010

Citation

Willighagen, E. (2010). The Molecular Chemometrics Principles #2: be clear in what you mean. *Chem-bla-ics*. https://doi.org/10.59350/dzqvt-ynv20

Keywords

Mcprinciples, Chemometrics, Rdf, Cml, Semweb

I noted earlier this week that *[d]uring the week [in Oxford], someone (name and address is know at the editorial office) commented on the fact that my blog posts are somewhat difficult to follow; that is, it's often not clear why I am posting what I am posting*. This triggered the start of a series of principles in the field I coined Molecular Chemometrics, and the promise that I will try to indicate in each blog post to which of these principles it relates. Just to put things in a bit more perspective; to make a bit more clear why I am blogging about that bit; just to be clear in what I mean.

Now, the first principle was about the need for access to data (McPrinciple #1). This principle goes without saying, one would think, but is not widely accepted yet. This is why Open Data promotion is still needed. For example, data in papers still is not freely redistributable, as Peter points out once again.

Anyway, this post is not about McPrinciple #1, but about the second principle.

Molecular Chemometrics Principles #2: In order to reproduce cheminformatics studies you need to be able to understand the input data.

Readers of my blog will surely recognize this theme. Clearly this theme explains my past fetish for the Chemical Markup Language, and my more recent work on the Resource Description Framework.

And it is so easy to jump to conclusions. Easy to make mistakes. And this is not just at the received side; the sending person may have accidentally made a mistake, or left something accidentally unclear, causing incorrect assumptions, and therefore errors in the cheminformatics computation. Now, if the data was semantically (clearly) annotated, and the meaning was clear, it was also trivial to see when a mistake had sneaked in. Think of it as a check bit.

"Well, isn't this a bit exaggerated," you might say. Perhaps, perhaps not. An simple, recent example. We all know SMILES, right? And we all know that lower case element symbols indicate aromaticity, right? That is, c1ccccc1 is aromatic, right? So, what's the problem then?

Now, consider the SMILES string c1ccc1. Lower case carbon element symbols, so aromatic, right? Oh, wait...

Therefore, be clear in what you mean. It saves us from a lot of trouble.

Further reading:

- The Molecular Chemometrics Principles #1: access to data
- Molecular Chemometrics, 2006 (doi:10.1080/10408340600969601)

References

- 10.1080/10408340600969601