chem-bla-ics

Quality of Chemical Database

Egon Willighagen

Published June 19, 2007

Citation

Willighagen, E. (2007, June 19). Quality of Chemical Database. *Chem-bla-ics*. https://doi.org/10.59350/49wqj-62k11

Keywords

Opendata, Chemistry, Pubchem, Rdf

Lately, Chemical blogspace has seen an interesting discussion on the quality of opendata and free chemical database (over 32 free resources now), such as the NMRShiftDB.org. For example, see Antony's view on the NMRShiftDB and Robien's analysis.

Opendata makes such quality assurance possible, and I am happy that the NMRShiftDB was explored like this; the found problems can be reported and corrected. If correcting them upstream is difficult, opendata allows one to make a better derivative; that's what opendata is about. For example, BioMeta (DOI:10.1186/1471-2105-7-517) took data from KEGG and corrected a lot of molecular problems (like reaction balancing, stereo chemistry, etc).

I have contributed almost 900 spectra to the NMRShiftDB, and I am sure I may have made a mistake here and there. But my submission is verified by a reviewer, and furthermore, users of the database can report inconsistencies via the NMRShiftDB.org website. Now, I have focused on uncommon NMR nuclei, like ^{11}B, ^{195}Pt and ^{29}Si (see the stats), which tend to have only one peak. Nothing much that can go wrong; still, one or two errors were catched by the reviewer.

Ensuring data quality

Humans make errors, but not even only when data is entered; they make mistakes checking data too. Nothing much that can be done about that, other than using computers to find patterns. This is exactly what Robien did: he used his software which implements common patterns to find entries in the database that did not comply to those patterns.

Automated quality assurance requires a easy to use, machine-readable interface. For example, CMLRSS (DOI:10.1021/ci034244p) can be used for running new entries in databases against known patterns. But other interfaces are most welcome too. Rich recently discussed the new PUG interface , which offers an interface to PubChem.

German scientists offer a RDF interface to Wikipedia: DBPedia. Informal semantic markup in Wikipedia, such as the Infobox template, are used to create triples. It's a shame that the ChemBox is not used yet, which would make detecting molecules in blogs even easier.