chem-bla-ics

Open Data: license, rights, aggregation, clean interfaces?

Egon Willighagen

Published May 18, 2009

Citation

Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/hvqxm-xnq47

Keywords

Opendata, Nmrshiftdb, Rdf, Dbpedia, Bio2rdf

Copyright

Copyright © Egon Willighagen 2009. 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.

A recent post by Cameron on his visit last week with Nico, Peter and Jim, discussed Open Data licensing. This lead to an interesting discussion on these matters, and questions by me on why people care so much about only public domain data (or licensed with PDDL or CC0).

Open licensing for data has not as much matured as for software, and international law seems to be more confusing about the issues. I guess that is because data aggregation has been around for way before the computer era. The PDDL and CC0 both try to overcome this fuzziness. But there is another issue we need to keep in mind. A lot of useful Data was aggregated and made Open *before* these licenses came about, and use, for example, the GNU FDL license, such as the NMRShiftDB.

Rights

Right now, there are two Open Data camps, much like the BSD-vs-GPL wars in Open Source: one that believes in waiving any rights on the Data, indicating that facts are free; others that believe that data must be protected to not be eaten by big companies and lost to the community (e.g. the WolframAlpha arragnements are suspect).

Of course, both camps are not that far apart, and both believe Open is important. Interestingly, there are some noteworthy differences with the Open Source wars. I see parallels between the two, which details an important difference: Open Source has algorithms (uncopyrightable) and implementations (copyrightable); Open Data has Data (uncopyrightable) and aggregation (copyrightable). Open Source talks mostly about the implementation, not the algorithm; it's Open Source, not Open Algorithms after all. In cheminformatics it is even often the case that the algorithms are not even specified and that there only truly is source.

However, Open Data in title does not make distinction. Data is fairly cheap and acquisition can be automated and computerized; Aggregation, on the other hand, requires human involvement: curation and thinking about data models, etc. This is where added value is. Consider an assigned NMR spectrum or the raw data returned from the spectrometer.

It is this added value that people want to protect, not the data itself. I think.

Aggregation

One important argument that tend to show up when people argument for PDDL and CC0 is that it makes data aggregation easier. This is most certainly true: if you can do whatever you like with a blob of data, that also means aggregate with any other blob of data. However, copyleft licenses, like the GNU FDL, require the aggregation to have a compatible license too. It is the license incompatibilities that make this impossible. Or … ?

Open Source has matured to such a point that it is fairly clear what the intended behaviour is, regarding derivatives. An aggregation of software (typically refered to as a distribution) is only a derivative under certain conditions. This makes it possible to run proprietary software on top of GNU/Linux, which uses the GNU GPL but does not require software to run on top of it to be GPL

too. Unless... unless, not a clear well-defined interface has been used, indicating a strong dependency. Now, surely, these things have not been confirmed to match actual law in court, but the intentions are clear.

Clean Data Interfaces?

Now, if we would translate this to Open Data, would there be the equivalent of a clean interface? Can we build a data distribution with data of various licenses? I think we can! I am not a lawyer and please consider this an invitation to discuss these matters...

Let's start simlpe... if I put a GNU FDL image in this blog, by linking to it with a open, free, clean HTML interface (`<img src=""/>`), would that make my blog GNU FDL too? I don't think so. Surely, I would need to list copyright owner, and actually would be required to put the GNU FDL in my blog too, but hope linking to the license text would suffice too. (Let's skip fair use at this moment, and assume the use goes beyond fair use). Question: am I not using a clean interface, and would this not make the image's license no infect my blog?

A more difficult example, consider rdf.openmolecules.net, which surely aggregated facts, including data from the NMRShiftDB and DBPedia. I am using a unique identifiers here, the NMRShiftDB compound ID, and the DBPedia URL, which surely is GNU FDL, and use this to make a `<owl:sameAs>` statement. Again, please do not consider fair use, which this certainly is. But, let's say I put in some more DBPedia and NMRShiftDB data in this aggregation. The GNU FDL data on rdf.openmolecules.net would be separate RDF blocks, with proper dc:license, dc:author annotation. But the block would be part of a larger aggregation. The clean interface here is Resource Description Framework.

This second case does not only affect my rdf.openmolecules.net website, but, for example, bio2rdf.org is also in the same situation and aggregated and distribute DBPedia's GNU FDL data (e.g. hexinanose. Does that make the whole of bio2rdf database GNU FDL. They too use RDF as clean interface.

Call for Discussion

Despite what one of the two camps like to see, the mere fact of added value when making data aggregations will keep copyleft license stay around, and instead of trying to convince everyone of the virtues of PDDL- and CC0-like licenses, we should think about to what extend it really matters.

I can do my data analysis with data sources of various licenses. I can search and retrieve data from various sources with various licenses. What obstacles are really there that disallow us to do science? Do the data interfaces we have now not provide enough technical means to address the license incompatibilities? They have in Open Source, why would that not apply to Open Data too?