Q&A
13 October 2021
Towards Biological Plausibility Using Linked Open Data
Dr Egon Willighagen (Maastricht University)

Q1: How can a database like ChEMBL be helpful to build our own ML system for drug discovery? To do that you need a defined threshold to differentiate your dataset between 2 states like active and inactive, but on ChEMBL you will get a collection of e.g different substrates which were obtained and published in different research labs?

I've just read the question about how to use ChEMBL, and the question where you put your threshold between active and inactive. We wrote an article some years ago about this and we did not use classification of active/inactive, but we made a regression model. This work was done in Uppsala where they had a good bit of experience with proteochemometrics already. Just classification is possible, but you can do a lot more with ChEMBL, and then you don't have this problem of a threshold at all because you're just making the regression. The semantic part is where we use information about the quality of the study. It's a bit crude information in ChEMBL, but at the time they had a classification of 1 to 9, it wasn't quite linear, but it at least gives some direction.

The effect of taking into account that assay confidence can be seen in figure 11.

Q2: Are you planning to use any SHACL in the development of your further ontologies or queries? And what do you think might be the advantages of that compared to the current method you're doing with SPARQL?

SHACL is an alternative implementation of the idea of shape expressions. The short answer is no, because we use ShEx for that. But the same approach is quite comparable. They behave slightly differently, but if the question is are you using shape expressions? Yes, we are them for quality control where it represents the minimal amount of information that we need in the data source. Or as in our paper about the protocol for adding data to Wiki data where it is also about ensuring conversion went OK, so not the data itself is in the way we wanted but also that we can also use it to monitor the conversion process itself. So, shape expressions is definitely something that we are using and will continue to use.

Q3: You mentioned conversion, if you are converting data into linked data, what methods do you use?

A variety of things really; it depends on the format in which data comes. Personally, I'm quite fond of using the Groovy scripting language because it can handle XML quite well and with Bioclipse or the current version of that bacting. I have access to a number of other libraries that allow me to read Excel spreadsheets and Google spreadsheets. And the advantage of using the scripting languages that you're not just doing the structure, reformatting, destruction, reorganization, but you can also do a bit of a data curation as you go, a couple of tables that normalize some labels into either a central type or in harmonized label. So, the script is a combination of the reformatting and automated curation. I do prefer to automate the curation and conversion as much as possible just to be able to not have to repeat the work when a new version of the input data comes. This automation process is something that I really quite cherish.

Q4: Definitely. I've done quite well with R2RML scripts and libraries are combined that with JavaScript for conversion, but that also took quite a while, but once you least once you had the script, you could then run them for new data.

Yeah, Ammar Ammar in our group, has been playing with RML a bit; so here the mapping is in a more formal mapping language that you can use to convert relational databases and other structured data formats into RDF.

Q5: My question is whether you think extension of pathways is something that we could apply my kind of learning approach too. So, having watched the talk before, because you're obviously working a lot with pathway information, do you think there's any applicability there?

Yes, I think so; There are a couple of interesting things happening. One bit of machine learning, well, the equivalent of text mining, is OCR on pathway diagrams. The team in San Francisco, Alex Pico's team, have been using pathway OCR and this has given a website where they made this available and you can search there for gene names and you get pathways from figures from articles. Another thing that that is of interest is the work by Andra Waagmeester. He was also the one that created the first version of the Semantic Web representation of Wiki Pathways and a lot of the Wiki data work, and he has worked on a Pathway Loom and this has the point of extending pathways by using knowledge from other databases. So that could be partly look-up, but I can quite easily see how you include other information there. So, one thing where we sort of see this is protein interactions, where data can be based on a co-occurrence in an article in literature or based on computational protein interaction strength. Here, they already have that information cashed in protein interaction databases, so you don't really have to include the machine learning directly into the pathway growing.

Second, we do classify our pathways with the pathway ontology and in our data analysis. Pathway similarity at an experimental data level, for example at a transcription level or at a proteomics level can be combined with ontological information in the Gene Ontology, the processes there. Laurent Winckers published an article where Gene Ontology information is used as a filter to not get this really huge hairball of data points but really a zoom-in on particular biological processes. I think there are some interesting links there with what you have been doing.