Anonymized Questions Transcript

Title: Open Science in Chemistry – The Past and the Next 20 Years
Speaker: Egon Willighagen, Maastricht University

Questions

Participant A: *Thank you very much for this really nice talk, especially about the history of a lot of open science and chemistry, it was very interesting. I have a question about, you said in the beginning something about the open standards, so that they should not be or they must not be complete when using them.*

Willighagen: Yeah, it was a discussion that came up two weeks ago as well, and I was presented with the statement slash hypothesis, we were discussing that, whether you can capture something in one standard. And I think the answer is no. I don't have arguments or evidence to that, except for the following. My best argument is no. I don't have arguments or evidence, except for the following. My best argument for the no is actually, well, we're researchers, so we're always looking for new answers. And if you're confined by a standard that limits what questions you can answer, what studies you can do, then I don't see how that works. Then we're limiting our freedom to do research. So, to me, a single standard doesn't work because it inhibits the intrinsic nature of research. That would be my best argument. I'll leave it to you whether that's convincing or not.

Participant A: *In principle, yes. I mean, for me, a standard must not be closed or ended at some point. So, I think it's, especially in science, like you said, it's very important to have these standards open and, yeah, so that they can be adjusted to new findings in science. So that we can just put a new topic into them, a new category, if we need to.*

Willighagen: Yeah. The stability is something, of course, that is of interest because you can standardize on something, but you need people to learn the standard and to implement use cases around it. You can define a very nice new language, but if no one speaks it, but learning to speak the language and to write it, that will take time. If things... So, there's this balance, indeed, with being able to argue. You can update it and stabilize some things. But this is, I think, why we can have multiple standards overlapping next to each other. If I compare it to science itself, I think we have this all the time. If we think about biology, and biologist thinks about metabolic reactions often in a different way than I think about them in a chemical way. We see this reflected in databases as well. So, in science, the notion that we have different more of what we call the models rather than standards, but I think the parallel is quite high because these models are simplifications of reality that make it possible for us to standardize our communication about the reality. I will try to not get too platonic here.

Participant B: *You started with discussing the various dimensions of openness before diving into the various infrastructures you worked on. One dimension you did not mention is openness in terms of participation, opening up science practice and community to all. The infrastructures you mentioned are highly complex. How do you envision openness in terms of participation being built into them?*

Willighagen: Very good question. This is something that is getting a lot of attention right now. A very nice example I see there which actually links to that question about single standards and not whose knowledge. And they had an awesome presentation some time ago. If you have a standard, you have a small group of people that decided how things should be done. Maybe that's my second argument for why one standard I would say no. If you have a single small group of people, [no matter] how wise they are, if they dictate how it should be done, then we limit that aspect as well, even if it's correct.

How infrastructures there fit in, I think in a similar way as in open standards, because those choices will be adopted. The open infrastructures let us know, tell us, dictate us what we can do or not. The big publishers, for example, they have an internal data format, which they typically do not share so much. We have JATS now, but you can't submit your article, not even in JATS, a standard for submitting articles to central archives. I would love, for example, to be able to submit my article, not in LaTeX, not in Word, but in their native format. I would expect that that would greatly reduce the cost of the typesetting that we're seeing right now. That could be so nice an infrastructure, but we don't have that. We stick with the infrastructures that they have or are developing, what their interest is, and with this, this has clear implications on the people that have to use this infrastructure. So I totally agree with the point there.

The same thing we see, for example, with GitHub. I don't have solutions, so I'm not going to be able to give a clear answer to that. But another area where we see this problem is with GitHub, for example. I mentioned WikiPathWays, and we have had researchers from Iran studying biology in our group using WikiPathways. Back in Iran now, but WikiPathways at this moment is hosted on GitHub - the website at least. They can still download it because the archives are available from a different service. But the notion that infrastructures... There is a political angle to this. They don't have access because the United States government forbids GitHub, an American company, to export their infrastructure to Iran. That has direct implications to us, for us as researchers. So yeah, a very good question. Unfortunately, I think we need more solutions. I don't think we have enough solutions there yet.

Participant C: *From your point of view, what's the biggest obstacle for interoperability between different groups or labs? And along that side, for repeating a chemical experiment? And what's, from your point of view, the best idea of resolving these issues?*

Willighagen: For the second question about repeating the experiment, I think the role of Open Science there lies in the first place in communicating carefully how the experiment is being done. If that is done, I think that is a requirement for us to be able to try to reproduce it. And in doing this and in communicating this for those experiments where we don't know it yet, we can start actually looking at the differences. What is different in one situation and another situation. We can start, I think, we can start isolating the things that determine the variance, the inter-lab differences, those kinds of things. Even if it just finds that, so you have the description of the experiment and then how the experiment is being done. The open infrastructures around electronic lab notebooks can help with both of them. Even if the description is the same, but we still get different results, then we have a clear starting point of, well, we need to sit down. Something very weird is happening here. We don't, apparently, we don't know what is happening here because we, neither of us have thought about how to do it.

With regards to the first question: that's a very hard question, but I'll take, for me, the hardest one that I practically have. So you can see that on [the last slide] with the nano safety projects that I have been involved in. In grayed out, those are the grants that have ended. In the eNanoMapper, NanoSolveIT, NanoCommons, and RiskGone, and SafeByDesign-for-Nano (SbD4Nano) (that is the ongoing one) one of the interoperability problems here really is, how do we communicate our knowledge about these nano materials, what they are. Often because the biological experiments that are done at the same time as the physical chemical characterization, we know that they come from the same chemical batch, so we can match them later. The problem, the bottleneck there is really, really figuring out how do we do these experiments in the first place? There are round robin experiments being run. How do we communicate this to the regulators, which is part of the RiskGone project? How do we use, and what information do we need to use to do predictions, which is the NanoSolveIT project, for example. And in NanoCommons, in eNanoMapper, we try to... The obstacles that we're trying to overtake there is at least improve our language, our communication, to be more precise, so we can start identifying what we can see on that side, so that we can translate our gut feeling, our expert knowledge, into something tangible, related to the underlying data. So that is a huge, huge bottleneck.

I think in chemistry, another bottleneck actually is our lack of community agreement on how we want the field to move forward. So we still have a lot of things that are not Open Science in Chemistry, though this is changing rapidly at this moment. But there's still a lot of knowledge in resources that are not accessible. And one interesting thing there is, a couple of years ago, I was looking at the best-selling drugs in the United States, and apripiprazol was one of them. And I thought, “oh cool”, let's start at the top of this list, and let's get together some information about these compounds. And I started looking at the apripiprazol, and I was looking at what metabolites can I expect in the human body after you administrate this compound. And then I was looking, and I was reading the literature. Ah, ah, they have some information here. Okay, where did they find that? Because they were repeating it from another source. Oh, okay, this paper. Ah, ah, this paper. Oh, oh, oh, what is happening here? They're citing a conference poster. Okay, that's not useful. Okay, where's that conference poster? Oh, here's another source. Oh, oh, oh, yeah, but they're actually citing the information leaflet shared with the package with the drug. Is that a scientific valid resource? So there are a lot of social aspects here in how we do science, how we communicate science, how we trust each other. This is a huge bottleneck that is of a more psychological and social science. So, yeah, that is something that we need to work on and where we need more research on how to do these things properly. But that's a bottleneck not so much from chemistry, but more from the chemist, if you like.

Participant D: *I always thought that this is one of the beauties with the linked data semantic web, the ability to link multiple standards together. Together we have a bigger freedom about which standard to follow or even to create our own, which I agree with you, Egon, is very important to be able to do.*

Willighagen: This is another aspect of the open standards. If we think about standards as "okay, let's try that, let's see how it plays out", a lot of things need sufficient reuse to determine what the bottlenecks are, for example. You need it to be used in anger. So that we did [use] XMPP, for example: no one uses SOAP anymore, I think. That's not entirely true. Every now and then I see it show up, but not so often anymore. But XMPP? No, everyone is using REST or SPARQL, actually. But the discussions that all these things led to and the effects on how we think about all these things and what is needed, that always helps a lot. And that's that community aspect, I think, that Participant D was also referring to.

Participant A: *I have one more question. You said also in between the talk somewhere that some of the software does remote calculations of molecules. And there directly came to my mind, how is this funded? So, how are the machines provided with power where the calculations are done?*

Willighagen: Thank you, I can say a lot about that. First of all, everything costs money. So, Free Open Source Software is more about freedom than about free beer. So, that aspect is obviously there. But… And the nice thing is… So, yesterday we had a PhD defense from one of our candidates. And he worked on two projects not here, EU-ToxRisk and OpenRiskNet. And OpenRiskNet is a bit in the corner of projects. It was extending on earlier projects. And then we had another project, also earlier, OpenPhacts, where we used SPARQL interfaces and REST APIs for pharmaceutical data. And at the time to host this data in the projects over a period of, I think, three, maybe four years, hosting that data, that costed practically 300,000 euros. Fast forward about 10 years. VHP4SAFETY were actually pretty much doing the same thing, but only at the cost of about 5,000 euro per month.

Why? Well, Docker images. Evolved standards. Simpler things. Agreement about how to do things. Multiple implementations resulting into… I mentioned CDK (Chemistry Development Kit) and OpenBabel, for example. Why do we have two libraries or multiple libraries for cheminformatics? Well, actually, there has been a lot of detailed discussions about implementations of algorithms and comparing of results and implementation between the two libraries, which made both libraries better than some of the commercial cheminformatics tools there by means of that open. So, things still cost money, but what I think is what we're seeing here now is that open standards, open data, work really well with the growing computing power that we have, but the open things make it easier to integrate things that we can scale up things better.

And effectively, the cost goes down. So, in VHP4SAFETY, we can now do something for… What was it? We were… I said 5,000 per month. I think it's actually half of that, but the 5,000 is for the doubling that we anticipate in the next year. For a fraction of the cost that it cost at the OpenPhacts project, we can do the same thing now. This is because of Open Science, because of Open Source, not having to reinvent the wheel, an infrastructure provider being able to use Linux machines and video cards with Open Source on top of that, with open drivers, doing things much more efficient. So, Open Science is bringing down cost, I think.

Why can the Chemistry Development Kit, after more than 20 years, why does it still exist? What cost it? Why companies, in particular, in cheminformatics, went bankrupt or bought out and then bought out again and bought out again because it was not sustainable? Why can Open Science do something that commercial companies could not? Well, I think by bringing down the cost enormously.

Participant D: *What is your take on the status of linked data today? Did it get its breakthrough into common use? If not, what are remaining challenges?*

We first started having the first RDF data sets in the life science at least 10 years ago. We had that American Chemical Society meeting in Boston. We had a mini conference on linked data in chemistry. We set out as an afternoon. We actually, I think, it ran one and a half day, three full sessions instead of the one session that we originally planned. It was quite successful. But it sort of plateaued at this moment. And I think two things are affected here. People started looking at the use cases, started using it more and more. OpenPhacts is an example of that. And so we saw more and more use cases. But, yeah, I mean, combine three resources, you can do a lot more if you combine four resources. You get more interaction. You can ask more answers. So I think the plateau is partly answered by people having so much more opportunity by having things available as linked data that it kept them busy.

The other thing is, is that the tools were still pretty much under development. People were trying things. And we see this, for example, with triple stores. We have Virtuoso Open Source. We have Blazegraph. Blazegraph did some things really, really very conveniently, but it's sort of abandonware right now, causing problems. We have a couple of other triple stores. And we have some companies coming and going, providing services around triple stores. And, well, we have a commercial company in Amsterdam. I'm not sure how they do it. But there's TripliDB, I think, there. They're, again, providing commercial services around making linked data available. If I look at ChEMBL RDF that we did 10 years ago, it still exists. We're discussing the evolution of that, where it should go next. We have Wikipathways RDF for three years. UniProt has been running linked open data for a very long time now. And there are a number of other resources that provide RDF. And the whole Japanese life science community doing a lot of work in linked open data in this moment in the life sciences. So I think it has been perhaps a bit silent. But there was a lot of attention. I think this is going to change very quickly. And with that, actually, in a couple of weeks, there's a spot for life sciences, this very nice conference about semantic web languages and applications in Leiden. That will be an awesome place to go to if you want to see where everything is with linked open data in the life sciences.