Transcript

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

Introduction

Today, I am thrilled to present Dr. Egon Willighagen, who will be talking today about Open Science and chemistry. And what can we tell you about Dr. Willighagen?

Dr. Willighagen is assistant professor in bioinformatics at Maastricht University. He studies how mathematical representation and computation can maximize outputs from theoretical and experimental development as applied to drug discovery, metabolomics, and taxology. And he is a leader in the open science movement, specifically in the context of chemistry. So without further ado, Dr. Willighagen will be talking about open science and chemistry, giving a retrospective for the past 25 years and on to the future.

The floor is yours.

Lecture

Yes, thank you. Thank you very much for the introduction.

Yes, I'm going to give an overview of how I got involved in open science. And if I am a leader in open science, this is more because I've been involved in it for many years than something else, I think. That said, I'm advocating for Open Science is very important as well. I've been doing open science, and I learned a lot from this, and I'm very eager in sharing, what I learned about open science. There's a lot to learn, and there's a lot to do. And what I'm going to show in this presentation is some of the open science things I have been involved in, and what I will try to do as much as possible, explain why I'm bringing up that particular bit. I'm proud of all the open science that I did. But, each of these things have unique aspects which I think are useful for science in general and for chemistry in particular.

The past and next 20 years is an attempt to commit myself to actually say a bit more about, what we still need to do, because we actually did achieve a lot in the last 20 years. But, I'm realizing efforts in Germany here, like the Forschungsinstitute for Data, with the chemistry here. There is so much happening more now than there was 20 years ago. So let's see where we are.

(Speaker starts sharing screen)

I'll start with this slide. This is a slide of all the people I have collaborated with, some more than with others, but these are all co-authors on papers. And the reason why I'm showing this slide is that I have my contribution, I'll try to be as clear as possible where my contribution was, what I did, but as it should be probably with science, and definitely is with open science, these are collaborative projects. So there are a lot of people involved, as you can see from this slide.

The slide, the list was made by Scholia, one of the tools that we have been working on, based on data and Wikidata. You can see the link to actually my scholarly profile based on Wikidata there, and you will find information about many of these authors there in Wikidata too, as well as the papers that are the base of that.

So the abstract of this presentation is this long text, and this abstract is coming back a couple of times, and each time I highlight some aspect of the abstract. So we're going from the top to the bottom a bit, and we'll start with Open Science.

Open Science, less than Open Access fortunately, but there are different views on what open science is, and what is inside the scope of Open Science , and what is outside the scope. So I want to make clear what I consider open science. And importantly, and this is actually how it started all for me, is actually it's something being reused. Open Science is only useful for me if someone else is reusing what I do, or if I can reuse something of something else. To me it feels a bit awkward to think about this, because it would be as quite equivalent to actually, can you use the laws of gravity? That is research output from someone else. Can I use that? Yeah, fortunately we can. It's old enough perhaps that you could argue is not a reason.

But open science to me should be about learning new things and disseminating this. But this is not nowadays, not universally agreed on how to best do that. Open science is one of the solutions, but there's no hard evidence that it's better than knowledge going into industry, and industry making things that benefit society. But reuse that is important and being able to learn from earlier work and particularly if the earlier work needs adaption for example.

We will see that in a moment. Right now basically, not just theoretical reuse but real reuse and for that a particular scientific setting you need three rights. You need to able to use it at the first place, but you need to able to modify it and reshare and reshares getting attention, but the notion that you need to be able to modify something, for example to improve it, to remove typos in text, to translate it, all kinds of reasons why you want to be able to modify it.

For source code, this is very obvious. If you have a tool that you want to extend, and that is the first example that I will show, if you can modify it, then you can stand on the shoulders, if you like. You can extend on previous work, you can improve things, you can repurpose things, and we can create new use cases for things that things were not originally created for.

What are aspects of FAIR? So “FAIR” and “Open”, they're said not to be the same thing. They complement each other. FAIR is not a requirement for being open. Open is not a requirement for being FAIR.

But: open data that is FAIR is definitely better than open data that is not FAIR. FAIR, the findable and accessible, that is advertisement, in a good sense. If we have knowledge, if we have research, but we don't disseminate this, we don't communicate it, then no one can use it.

So FAIR, findable and accessible, for me, is important as well. But if we want to reuse, we need the content, the research output, that we make FAIR. It needs to be interoperable and practically reusable. Something can be very easy to reuse, but be completely useless for what you want to use it for. And that reusability, the scientific reusability, that is something that is seeing a lot of discussion right now in various communities.

And that is important because in every community, this is different. What we actually find in our group, is that the notion of reusability is not even the same for the same research output, depending on how you want to use it. So data is not intrinsically useful or not useful. It actually depends on how you want to use it, what you can do with it. That is getting more clear, and that can be formalized, actually.

So what about Open Science? What are the core components there, I think? Well, what we've seen there is that where we see nowadays open access still as the primary part of science, how we communicate it, this is really broadening right now. Open data is coming along. Again, it needs an open license because there are legal constructs that are limiting the openness of data. Facts are free, are said to be. Databases can be protected to some level. There's a lot of confusion, expensive confusion. And then there are a lot of things that we actually want to...

What we need as well, we need the history of the results of an experiment. How was it done? What machine was used? Depending on the machine, the interpretation of the data can be different. So a lot of protocol information, metadata information that needs to be complemented as part of that open data. This is one aspect where FAIR has done a lot of work. It's done a lot in the last 10 years.

We need open standards. We need standards. I've had discussions in the past whether we should say open standards or open specifications because standards tend to be associated with lengthy, expensive processes where we try to find a solution that fits as many people as possible. I'm still not entirely sure if we need that for open science because open science, a research process, is something that is continuously moving, and by trying to come up with the open standards, we might be trying to make shortcuts that inhibit us from doing the research, from doing the science.

But in chemistry, we have a good number of open specifications, open standards. We have SMILES or OpenSMILES, for example. We have the chemical markup language that is based on the extensible markup language, the XML framework. We'll see that we have identifiers, the InChI and the nano-InChI we were thinking about, and there are things like how do we share standard ways of sharing metadata, like bioschemas. All things that provide us languages to communicate what we want and in this way make things more interoperable and more useful.

Open source is very important here to me because a lot of things that we do is actually automation. It's calculation on the fly. It is interpreting things. It's analyzing things, and we need to be able to study that, and we need to be able to see how it works, how it can be improved, why it's failing in certain cases, why it has bias. So the open source component is part as well. Interestingly, actually, we see that the link between open source and open data is quite close. Very often, software packages are distributing open data.

They need open data as part of some of the algorithms in the software. And we'll see a bit more about that link, how we have been looking at that in the open science chemistry community. Open repositories, we need infrastructure, and open repositories are part of that.

And we need ways of repeating things. This is still really hard, repeating an experiment, whether it's digital or experimental. And we have to put effort in making sure how to reproduce something. See what we need to do to make the experimental error as small as possible. Definition of standard operating procedure, for example.

And very importantly, and this is where the work from you, from Monica, comes in so much, is the notion that, well, we need to peer review. That these are not things that are isolated. They're an Helio of open science. And I really learned a lot from how to talk about these things from a model like this. A model like this as a standard, a way to communicate things, to make sure that we have a better understanding of what we're talking about. So that is for me open science, open research, if you like, open knowledge, and open science. And the various aspects of that.

And we're now going to look at a number of things, Jmol, JChemPaint, Blue Obelisk, CDK, moving towards NanoCommons, SafeByDesign [SbD4Nano], Scholia, some more recent things where these things come in. And I will try to indicate at each moment where these aspects come in and why I think they are aspects of these, of open science, worth keeping.

So for me, this, the dictionary on the chemistry, this was a Dutch website. I created that. I started that in 1995 and it lasted until about 2003 or something. It was a website that I started because of two things. I wanted to learn how to make websites. There were no courses there. It wasn't a financial business at that time. And I was learning organic chemistry. And I realized that, well, I mean, I'm probably not the only one. I'm trying to find this information for this course, name reactions, trivial names for chemicals. Why do we call this acetic acid and not etanoic acid? That kind of thing. What is the provenance? What other databases have more information? Or what is the ontological classification? It was in Dutch. So the slide is in Dutch here. I actually did start even translating that. And then you want to translate acetic acid to the Dutch “azijnzuur” and not acetic acid to “etanzuur”. It's still all the same thing, but you want to capture some of the etymology as well, perhaps. At least if you're interested a bit in the history of chemistry, which I was. But we see a picture here of “etanzuur”, of etanoic acid. And yeah, that was an image, a PNG or maybe even a GIF image at the time. And it was a drawing. And I didn't like that. I wanted to have it more machine readable.

So I started looking around and I found tools that could do that. Java applets specifically. And it was Jmol. This was done by Dan Gezelter in the United States and Notre Dame. And that could depict three-dimensional structures. Nowadays, it can draw large protein structures as well. Computing was not that powerful. And there was JChemPaint from Chris Steinbeck here in Germany - in Jena at the time, back in Jena now - which could draw two-dimensional structures. That was JChemPaint. So I had a way of having the chemical information in a machine-readable format.

And then these tools could visualize it. Now, I realized that I don't want too many files about the same chemical. At the time, the data from the web page was in HTML. Actually, it was in XML. It was already separated at some point. The data from the display. The model from the view. So I wanted to have the chemical structure information as well. That's how I got into the chemical markup language.

There was a project by Peter Murray-Rust and Henry Rzepa in the UK. They had a machine-readable and semantic model for how to store chemical information. The chemical graph. And the chemical coordinates. 2D and 3D. In one file. And I could put this file inside the file that I already had. Perfect, I thought. Because I also had the tools to visualize that. This is how I actually really got into open science.

Because these two projects, Jmol and JChemPaint, they had source code. They had an open license. There was no version control system at the time. They were just files that you downloaded. You unzipped them or untarred them. I don't know. I don't remember that, unfortunately.

And I started making patches and I approached the various people. Chris Steinbeck, Dan Gezelter for Jmol, and Peter for CML. And I said, well, okay. Do you accept a patch for Jmol and for JchemPaint if I add editor support? If I add reading support for CML, for chemical markup language files? They did. And that was my first real open science contribution. That was about 98, so about 25 years ago.

If that makes me a leader, I'm happy to be a leader. Importantly, I was still a student. And the reason I bring that up is that these were small contributions. They were small things. These were things that I could do. And one of the things that I really hope for science is that we don't... Don't stop accepting small things.

And open science here really is awesome. Because Jmol still exists. JChemPaint still exists. Open science doesn't go away. If it has these permissions of modifying use, it can be extended. Neither Jmol nor JChemPaint, they still look like that.

Well, JChemPaint hasn't changed that much in terms of look. But Jmol, for example, it's no longer a Java template. It's fully JavaScript now. There's still Java code. There's still Java code. It's underneath it. And it has been rewritten at least twice since the original code base. But it evolved over time because of the rights of use, modification, and redistribution.

Okay. So another interesting thing happened because we had this network. We had these people collaborating, and we had each other reviewing. Dan and Chris also realized, well, actually, there was a lot of overlapping code. There was a lot of overlapping code there between Jmol and JChemPaint. How to represent the chemical graph, the data model of how to represent the molecule in the computer. They both had that. Why do we need that twice? That is twice the same code. It's both Java. It doesn't have an advantage. We thought so at the time. Actually, it did have an advantage because what JChemPaint needed, and the chemistry development kit needed later as well, is more flexibility.

But Jmol, if you want to visualize a protein structure, with so many atoms, you need something really, really fit for purpose, really efficient. So we merged it, the code bases, and it diverged at some point again. So the Chemistry Development Kit, that got founded in 2000.

We had an... There was a conference, a chemistry and internet conference in Washington. I was a student there. I got a bursary to present, well, actually a presentation there about some other work where I was using Jmol in the context of using content negotiation to provide the chemical information in the most semantic format, depending on who was accessing the website. That is actually a failed project. I gave that presentation. There was some interest after that, but there was so little interest that the... So a few questions, I think I had two questions, was so disappointing that the project just died. That happens to open science as well. The code is still there, though.

But the Chemistry Development Kit is a very nice example. I remember when we first started pitching it, I started pitching it at a research group at some point, a cheminformatics research group, and I got the question, why are you doing this as a researcher? Will your work actually be cited? They can just download it. Well, yeah, it did get cited. It got cited quite a bit. Not as much, perhaps, as some other open source cheminformatics toolkits, but it did get cited, and it's used a lot, and we're now 23 years after the founding of this project. That says a lot about the sustainability. Now, the other interesting thing here is actually all the people involved here.

So Chris Steinbeck, Dan, and me, we founded the Chemistry Development Kit. It was based on Jmol. We founded that in Notre Dame. After that conference in Washington, we went there to Notre Dame, and we started this. We wrote a campaign to be based on this new code base and what's not, but what you can see here actually is a lot of people since then that were not part of that original team.

There is John Mayfield. He did a PhD with Chris Steinbeck at the same point. Actually, so these enormous performance boosts that we see, that is all the work from John. We see an enormous improvement there, though, better algorithms and better code. We focused on what we needed to have the algorithm openly available in the first place. We needed to do to solve problems. That's where the grant money was coming from. But at some point, we were past that point. We had all the basic functionality, and we could start focusing on making things better.

Jonathan, Arvid, Lars and Ola here, they were in Sweden, and we'll see that in a moment. BioClips was a tool integrating cheminformatics and bioinformatics toolkits, and they needed a good cheminformatics toolkit. They were looking for something in Java, so they actually came to one of the CDK meetings. They got involved, and they were a heavy user.

In fact, Bioclipse also still exists. Let's see. Yeah, and then a number of other people. Nina Jeliazkova, I highlight her here for a second as well, because we'll see a lot of nanosafety projects later on when I come to the new work, and a lot of work that is open source work from her company is actually being reused in those projects.

What happened here basically is people got together with a mutual interest, with a willingness to invest time in open standards, in open source and open data as an early infrastructure. We didn't need conferences. All we needed basically was a web browser, a place to email, a place to share code. SourceForge at some point was really helpful, because we started using version control. That improved things.

In the chemistry development kit actually that is also... I mentioned peer review earlier. One huge scalability thing in the chemistry development kit was that we got so many patches with so much work that was done, that our main line source code branch broke too often. The code base had become so large. We didn't have continuous integration, continuous testing yet that the manual, the expert peer review was no longer enough.

And at that moment, we started doing things. So we started writing unit tests and integration tests, but we also formalized the peer review a bit. We wanted to make sure that there was a senior CDK developer that reviewed the code, that made sure, that double-checked that the source code compiled. So there were always two people that had the code compiling, the new patch. And that really helped a lot with the scalability.

All these things are quite normal nowadays, actually, 20 years later, because you can, on GitHub, you have the pull requests, and you can block your main branch, and you can formalize the whole peer review process there. Really great infrastructure that makes all this work a lot easier. We had to use harsh methods like only having a very few selected elected people, if you like, with push permission to the main branch. That sort of is now also, but much more transparent than in those days.

And the Blue Obelisk, that was a platform founded by a number of people, Peter Murray-Rust, particularly at a meeting in San Diego, at one of the American Chemical Society meetings. And it's called the Blue Obelisk because they were at a restaurant. At a small square with a large blue obelisk in the middle. That's the etymology of the Blue Obelisk movement. And this diagram highlights what the Blue Obelisk is about. We worked on a number of things. Standards, those are the diamonds. Data, those are the rectangles. And the oval things, those are source codes. And we decided, okay, let's try as best as we can to reuse things where they exist. Isolate things.

So one of the things that we realized is, well, we had two open source toolkits. Well, actually multiple. There was JLib, there was Java, there was the Chemistry Development Kit, and there was already Open Babel. And, of course, actually, RDKit here is this one as well. This is the article from 2011. The first Blue Obelisk article was in 2006. RDKit was, I think, started just two, three years before that. We did not have that on the radar at that time. So we have alternative implementations doing the same thing. RDKit and Open Babel, they're in C and C. This is C and this is C++. Easily accessible in Python. And these were in Java and easily accessible in those frameworks.

And there are other toolkits in Java. So JChemPaint is Java, so it's much easier to base that on a Java library. Opsin is also Java. This is a Java wrapper around a C library, not depicted here, but there's an in-sheet right below this, etc.

One thing that we realized is there were multiple toolkits that needed isotope information. If you want to calculate the mass of a molecule, you need the weights of the isotopes. If you have the weight of the isotopes, you can take the natural abundances of the isotope and you can calculate the natural weight of the molecule. Because the isotope distribution can actually differ over the world, so can this mass. But they all need the same numbers for those isotopes. Those values are established by organizations like IUPAC, the International Union of Physical

and Applied Chemistry. They calculate these things in a lot of detail. But that needs to be maintained, that list.

We started the Blue Obelisk Data Repository. We factored that out from the code, so we started separating source code from data in this case. This data set actually got a very nice spinoff because it started getting used in Calcium, a periodic table viewer of the KDE desktop. This data that we did in this project got repurposed and ended up on quite a lot of computers around the world. Very nice example of, I think, the power of open science. It is there, and it gets used in all sorts of ways that you did not anticipate.

I'm going back one slide for a second. At the top you see Bioclipse. You see Bioclipse integrating a lot of libraries, and this is just focusing on the chemical informatics. The bioinformatics was this platform to provide a graphical user interface, but at the same time, particular Bioclipse too, still the ability to script things. It was actually recording things, or you could turn on recording while you were visually doing things. It would create a script on the fly in the background. You could save that script, and you could then repeat the same process later on. They really worked that out. Actually, the two actual Bioclipse papers are not here, but the scripting was, if I remember correctly, started by Jonathan. Because the scripting was done via Java, and Java had support for multiple languages, you could script the same things that you were doing in Bioclipse in JavaScript. That was what they first implemented. We added Python later, and I think the third one we added was Groovy. I like Groovy a lot because that syntax is very close to the Java syntax. Because I was coding a lot of Java, I was much more fluent in that. Here we have a use case that is quite relevant to all the research that I have been doing in the last 10, 15 years, which is predominantly around the toxicology.

In my introduction, you hear about drug discovery. In Uppsala, we were doing more around drug discovery than toxicology. But I knew, well, I mentioned Nina Yaskova already. She was doing predictive toxicology. This paper was a crossover between their OpenTox project and Bioclipse. And why? Well, they were using RDF. They were using semantic web technology. They were using a REST, or at least a REST-like interface. They were using open infrastructures. They were making their, their chemical structures and the toxicity information available in a machine-readable way. They were making their predictive models in machine-accessible documented API calls. So that was easy for us. Well, that we can integrate. That we can work with. And so we did. And that resulted there. So this book chapter is the best description of that.

But there is some more work around that. This is how that got integrated in a lot of detail. So here we have computational processes running on a remote server, somewhere on the Internet. And if there is someone here that remembers one old way of communicating things on the Internet, like the USB plugs, but then all the time, and electronic XMPP.

XMPP, we extended XMPP. That was done by, he should be on the previous slide, I think. No, he was not involved in this article. We extended, Ola and I, we worked on the Bioclipse client-side implementation. And I'm really sorry. I can't. It's on that slide with acknowledgment. His name is somewhere on there. He wrote an I.O. extension to XMPP. And we could have computational models behind an XMPP interface asynchronously.

That's the most awesome part. So solving the problems of SOAP, which was predominant in bioinformatics at that time, the servers would actually send you a message when it was done calculating. And then you could fetch the data. We had that implemented all here. All things that are still, well, we're still writing similar solutions right now, which is good. You need alternative solutions. And here you see some more integration. So if you can do it for one molecule, you can highlight in one molecule where something is wrong. You can do that for multiple molecules as well. You can scale that up. And you can then start playing with how do we visualize this big data. And then in toxicology, you can visualize it with red and green indicating, okay, this is a molecule you need to pay attention to. There is something wrong with it. There is a potential risk there. Another reuse here, taking advantage of, it's there. And we can use it for work.

It's recent work by a Ph.D. in our group, which is focusing on single nucleotide polymorphism, as in peas, in proteins. So the missense, as in peas, binding with ligands. And the ligand here is in green and the protein is in gray. And what happens if you have a mutation that causes an amino acid change and change in that protein? And he did that for about 600,000 combinations of ligands and proteins. I think 24 proteins and 600 bindings with small molecules for those proteins. And he made that all available on the web. Why was that possible? Well, because of all the open source, including the visualization. And with this, you can make a high... Well, there was really a lot of computing behind this. But you can make the results visually. Visually available.

Why was that in the scope of a project of a master intern, actually? Because all the open science provided so many building blocks that could be easily reused. So this is the actual reuse where I focused on. Putting an open license on something is not enough. You really need to put focus also on making it interoperable by using standards, really making it easy to reuse.

Let's have a look at the FAIR principles. I don't have a lot of slides to that, and a couple of aspects of it we already saw. But a couple of things that I want to highlight that we worked on as well. Again, stressing this collaboration. So one of the projects we had with people from around the world, Michel Dumontier, he's now actually at Maastricht University. At the time, he was at Stanford, if I remember correctly. Jenna Hastings in the European Bioinformatics Institute. Nico was in Cambridge. I was... I don't know where I was at that time. Leonid was in the United States somewhere, and I also do not remember where Chris was at that time. An international community, and we had a common interest. How do we capture the aspects of computation so I mentioned earlier the calculation of the mass. So you need to know what isotope information was used.

So how can we share this information? How can we make this metadata findable and accessible and interoperable? And then you need a common language, an ontology that describes that. And we worked that out in this article. This ontology, so this paper is now 13 years old. We're still working on this ontology. We're still using this ontology. Where? Well, for example, in the eNanoMapper ontology.

So this is when Nina and I really started working again, but now in the field of predictive toxicology eNanoMapper is in that sense a continuation of the OpenTox project that I mentioned earlier where we interacted, we from Bioclipse and Nina and Barry from the OpenTox side. We extended all this work, but now for nanomaterials. And there we needed an ontology.

We started using rich ontology. And Jenna Hastings from the EBI that I mentioned, we started collecting bits from other ontologies. We're reusing ontologies, but actually selecting them, reviewing them, selecting the bits that we want, indicating what we selected because that is what with the web ontology language you can easily do. You can isolate. And we worked out a workflow of, okay, we're using this ontology. We're selecting these terms as a means of unique global identifiers as you use in the web ontology language. There's a good provenance of where we are. This is still ongoing. We're still doing this. We're still improving the code. There are more software tools around this. Robot is a very nice one that is being used now for this ontology. The third example here would be the nanoparticle ontology that also needs maintenance, and that is all possible because of those Open Science things.

So what about standards? I mentioned the chemical markup language already. So this is actually one of the original articles. And look at the time here. So XML was introduced in January 1998, I think. I started my patching of Jmol and JChemPaint slightly before that actually because they were working drafts, open notebook science at the time. It didn't have a term at that moment yet. But there were early drafts of the standards of the World Wide Web Consortium. So people could start playing with XML. And CML was one of the use cases of XML. This is how Peter and Henry were involved in the XML definition and the chemistry use case of that. I already explained how I got in contact with both of them. And actually, we started using that a bit more for other things as well. And one of the things I still like very much, that's why I have that example on this slide, is that we took the news feeds.

We have Atom feeds now and we have RSS feeds. Well, actually, we have various flavors of RSS feeds. But one of the RSS feeds was an XML document. And CML was XML. So we could embed that with it. And so I wrote plug-ins for both Jmol and JChemPaint. I was writing plug-ins for these two tools anyway. That would actually... You could, just like in your podcast app on your phone, you track episodes of your favorite podcast or similar things for videos and sorts. What you could do here is actually... This paper is from a bit later. This paper was from 2005, I think. You could actually follow these new feeds with chemical information.

At the time, we envisioned that journal articles in their RSS feeds, which you're still using, that they would embed the chemical structures there, that they would embed identifiers of proteins, that they would say, well, we have knowledge here that we want to share and it's about this science. No. We're still doing text mining after 20 years. We showed at that time actually how such a thing would work. And Jmol and JChemPaint in those feeds, you could register to those feeds. You could actually even read an OPML file indexing those news feeds. It would actually recognize the chemical markup language in it. It would extract that. And it then would show a table of chemicals like this. You could click one. And it would show up in 2D or in 3D, depending whether you're in Jmol or JChemPaint. Awesome interoperability that never got used.

This is closer to Bioclipse, again, that I mentioned earlier. And here, we were using the reproducibility. I mentioned that how do we make this machine modeling of what quantitative structure, activity relationship. It's a mathematical model linking the chemical structure of a molecule, so what molecule are we looking at, with their physical chemical properties or their biological activities. So it's regression modeling, but understanding the chemistry and trying to figure out how, if two molecules are same or quite different, how that affects their properties. There's a lot there that you want to share. So extending on the chem information of ontology, the chemistry development kit, and actually JLib, we implemented that as well.

The idea was that by this ontology annotation, you could rerun the calculation, not just with the same tool, but even with another tool, because it would specify, well, this is the input that you need. This is all the things that you need to get together. And then it doesn't matter what library you use for that. Just repeat it. In that sense, a bit like a Jupyter notebook. It doesn't matter if you run it in MyBinder or in Google Colab. But it's complicated because you need to know, well, which mathematical description do you need for your molecule? Those are the descriptors. You need the input data and a lot of those aspects.

So moving to the end. So what we saw basically now is how we, this large community, worked on components, on open science components in open science. And a lot of these things are returning now. The thing that I don't have slides about, but if you look at the number of GitHub repositories in chemical literature, this is really growing very fast at this moment. There are a lot of young chemists that realize that we've passed the critical point in chemistry where open science benefits their work, their experimental work. And we see all sorts of fascinating things. So in that sense, I think the Blue Obelisk Movement has succeeded. Our work is done in that sense.

So what I'm moving towards next now is going towards that next phase. So what are the things that we're doing right now? Why am I so interested in the nanomaterials in projects like NanoCommons, SbD4Nano. What is the role of Scholia and Wikidata? What problems are they solving? And aren't they solving yet that we need to continue solving in chemistry? And one of the big elements there that I don't have a slide, so I just say it up front now, is still data. And this, I think, applies to a lot of fields.

There's a lot of data underlying knowledge, data knowledge not available in a machine-readable way that we would like to use, that we can use, but simply is inaccessible because digitizing it is manual work and it's too expensive. So that is one problem.

But that was a problem 20 years ago already, and it was actually 60 years ago from the first day that we started collecting things in a digital way a problem anyway. So one thing around nanomaterials and this is an example where Wikidata and Scholia is helping is that sometimes we need identifiers for things that are a bit complicated to identify. So a chemical at least has a chemical graph and they're quite stable. And because they're stable, we can identify them over a longer time. So chemicals, they have a CAS registry number, they have an InChI and an InChI key. They have a computer representation in OpenSmiles, for example, or in chemical markup language. And because that chemical is stable over time, this is how we can use it. This is how they have their function. Because they're stable, we have medicine.

If they would break down too quickly, the medicine would not have its beneficial function anymore by the time we take it in. For other materials, this is not always the case. And nanomaterials, as an example, they're not always that stable. So we need to track their stableness. If you keep them on the shelf, which actually is a problem for some classes of chemicals as well, if you keep them on the shelf, they degrade.

The other thing is that nanomaterials, they're not well-defined. A chemical graph has a fixed number of atoms. We can quite accurately describe that. That's for materials, that's not the case. So we have a new identifier problem. And we worked on that. And Wikidata is very nice because that is an open... Well, Wikidata is the machine readable, it's Wikipedia. It's a public knowledge graph that everyone works with, can interact with. Wikidata has APIs. So it's an also fair resource where, as a community, we can define an open standard, if you like. So one of the things that I've been doing is, well, we had these industrial... representative industrial nanomaterials from the Joint Research Center, one of the centers of the European Commission.

And Dave started out with these materials as they had batches. So if you order this one, you always have the same material. So if you use this JRC material, you know that you have a stable version of that compound, confirmed to be synthesized in the same way. So all the properties are meant to be comparable. So that is interesting because then the next thing is, well, what is all the literature then about a material? So I started tracking that and then you can actually see that, well, I mean, it's still then less than 100 articles in total. So collecting all this knowledge, this chemical knowledge together, is pretty much limited by the speed at which we do experiments.

All the more reason for me to make, I think, to make it as much as that openly available because if it's limited by that, then at least let it be that limit and not even just a fraction of that. Particularly in collaborative research, because this is all in the context of the European Nanosafety Cluster, a lot of research projects that are together trying to come up with an answer to how do we govern the risk of the potential risk of these nanomaterials that have a lot of beneficial effects.

Think of quantum dots in TVs. Think about the lipid nanoparticles used in vaccination. Think of metallo-organics like zinc oxide or titanium dioxide that have been used in skin protection, for example.

And with these identifiers, we can do things. We can use standards like bioschemas here to make metadata around datasets available, but not just the core metadata. Well, it's not visible on this slide, but we can actually say that this dataset is about one of these particular JRC materials. And then we can start doing... finding data in a much more efficient way. We don't have to browse thousands of articles about titanium dioxide. We can just ask, now just give me the articles about this JRC material or this dataset about this JRC material, this project deliverable at this JRC material, this predictive model around this JRC material.

So the use of open standards or for fair approaches to make things more reusable, because if they have an open license, even if they're reusable, if you don't know where they are and you don't have the time to spend six weeks in the library, then you would actually like a Google search to solve that for you. Does this work? So bioschemas is actually interesting for us and it works for nanomaterials, for chemicals, for datasets.

At least for the datasets, Google has the Google dataset search and these things are becoming findable there. And now at least if we can make those JRC codes part of the keywords, it's not ideal yet. It can be better, but at least you can start doing things. Give me all datasets or all, well, in this case, the datasets for a particular material.

Were we there? Well, no. That's just the JRC materials. We're studying a lot more materials. So we realize we need this identifier thing for other things as well. So Jeaphianna van Rijn, she's a postdoc in my group, and we started the European Registry of Materials, minting identifiers. Why couldn't we use Wikidata there? Well, actually, this is set in the title of the article here because at the start of the project, what we saw is that projects were not eager yet to share which materials they were studying. Often for valid, actually in retrospect, obvious reasons. For example, because they weren't entirely sure if they could get a copy, a sample of that material. So they would have an interest in that material. They wanted to talk about it. They wanted to link resources. But they may not have had the material. They couldn't physically, chemically characterize it yet. So there was no information yet about it.

But you still want to have an identifier so you can start using that in reports, in datasets, in spreadsheets, in whatever format you have to share, share knowledge. And just an example. So here we have six ERM identifiers. And with the data that we were looking at that we had available at the time from the various projects, this was all mostly internal data at that moment, these six identifiers, they actually had 60 different names altogether. And then you see titanium dioxide, another form of titanium. Oh, yeah, there's a titanium dioxide somewhere in the database with an underscore behind it. And then there is titanium dioxide in full letters. There was somewhere an internal code, P25 used. So, yeah, issues. More names than identifiers.

The classical thing that's why we're using global unique identifiers as much as possible. Why we have the ORCID for researchers rather than using our first and last name. One new thing here is, so one of the comments that we had on that paper is, well, it's just, you just have the identifiers. How is that useful? Well, our point was that, well, that material will become available. But we couldn't prove it yet. But now we can. So this is actually what that looked like. So we really had just an identifier. We indicated this is a chemical substance. And it had a label. We need to track in some way what that material was to some extent. We made the metadata that we were collecting minimal. But the people registering this identifier, internally they would make, that was the promise, they would make more knowledge, more metadata available at a later stage. But we did wanted to know which, who requested it. So that's reflected in this label. So in this case, it was the NanoSolveIT project.

But this is what we're collecting right now. And we started collecting this because, despite all this progress, it's a lot of work. So the articles, they're not fair. Journal articles are not fair. They're quite unfair. The title is searchable. You can access the abstract. The keywords, if you're in the right domain, they're a bit more fair. They might convert your keywords to mesh identifiers and controlled vocabulary in the medical sciences. But it's still a lot of work. Data sets as well, the metadata can be... There's a lot of research. Actually, the fair data objects, the research output crates, they're all looking into how can we make the machine-readable provenance, history, context of data and other research outputs more machine accessible. This is not the case yet. We don't have a consensus yet. We don't have an open specification yet. We have proposals there.

And the one probably with the most reuse will actually become it. But we can see for this material, we can see some information right now. We see the ontology annotation that we figured out at some point. We see a bit of a hint at the chemical composition. This is in blog post from this database. So this was when... Basically describing when this data was added in the database. The blog post we added now to highlight the different uses of it, to give people an example, a motivation of how to communicate these identifiers in different ways. And then we find a data set and an article here. The article is just a DOI. The data set is just a DOI. And by means of open infrastructure, APIs and open data, we just fetch the data. It gets formatted by JavaScript library called Citation.js. And this is done on the fly, on demand. So the metadata, the knowledge actually, is all embedded in that identifier.

The last thing that I just want to quickly touch upon is WikiPathways. So taking chemistry to biology. Also because WikiPathways is one of our main projects in our research group in Maastricht. And there our biological pathways look like this. And all these blue things, they're chemicals. Well actually, the black things, they're enzymes. So they're also chemicals. Those are proteins. But those are the small molecules that are closer to the chemistry databases. We want to know chemical properties about them. Not just the biology. We want to link the chemistry with the biology. So we need the identifiers.

And our tool there is BridgeDB, which is the identifying mapping tool that we have behind WikiPathways. And there we need Wikidata again. Because we need to map those identifiers to other databases. the CAS registry number in an experimental data set. If you want to link that, you need to know what other databases it links to, which pathways it needs to do. We need the identifying mapping to implement FAIR in this area.

You might suggest, and that actually works for proteins, why not pick one database and use the identifiers from that database. And then if people want to find things, they use just the identifier in that database. Why this complexity? Well, it's chemistry. Not all chemical databases, they have the same scope. And we've just found out that our biological pathways, they have a diversity in the chemistry that people want to describe in those pathways, where a single existing database could not provide that.

So we started using Wikidata to complement that, to fill the gaps left by the other databases. I don't particularly have a slide about that, but how that is picking up, Wikidata is used a lot in chemistry right now. And I see that one of the, the lead author of the LOTUS project is online,

for example, which is explaining which chemicals are actually found as natural products in species. And that data is in a very well, nice annotated way. Not a direct result of this Wikidata work, I think, but it does show the synergy. We have a community that is interested in the chemical structures on Wikidata.

The power of Open Science. And we're using open standards. We have APIs, and I'm pushing this a bit. So I mentioned semantic web technologies, RDF a bit, but it nowadays has a powerful language, Sparkle, which only recently learned is partly the result of discussions of people using it in the healthcare and life sciences community. Wanting to query this knowledge in the RDF, resulting in something that fed into the discussions about the creation of Sparkle. So all the time, open science on top of open science on top of open science.

Sparkle language is very powerful, and we're doing a lot of fun things with it. So concluding there, so what is left? Well, we need more reuse. We have a lack of data, but also the things that we need, we have, we need it reused. Why? Because that reuse is the best way to have people in and with enough attention peer review what is out there. If you're investing time in reusing something, you pay a lot more attention to why you're using it.

And why is this such a success story? I think in combination with open science is that with open science, you're not limited to two or three peer reviews. As we are used to with journal articles or with conference proceedings. Open science can be reviewed at any moment at any time, ideally during working hours.

Two other things that I wrote down and that has to do with that reuse is actually the community that you're creating. So like the symposia in the past as a main way to communicate new insights, it's now actually our open science projects that are this platform. There is this discussion, should journal articles exist? Well, we don't need them anymore. We have open science projects.

And we can do the peer review. In the CDK, we started doing that formal peer review in 2003, 20 years ago already. Peer review is not limited to journal articles. And the other thing, and that reflected by the different shapes, is this integration of different kinds of research outputs. They're so closely connected. It's really hard to see them independent from each other. It's not Bioclipse and then JChemPaint separately. No, it's actually Bioclipse. But immediately with that you get CML, you get BODO, you get OPSIN, the CDK, you get Jmol. And this is what all these things.

For the open source developers here, we know this as dependency hell. But cherish this luxury that we have this hell, that we have this reuse in the first place. It came up already, but one of our PhDs, Jente Houweling, she's really looking into one of our safety projects into FAIR aspects. And in the first year, based on what she found and the things that she did not find there, we should stop thinking about isolated things of output. We should stop talking about research data. And I try to highlight that with these dependencies as well. There is this data in source code. You can extract that a bit, but you can't fully do that. There's always this interaction. You can't calculate masses with the CDK without those isotopic information.

So we really need to start managing all our research output much better than we're doing right now. Because even the things that we are managing, we know all the problems there, but there's a lot of research output that we're not managing at all. So those are the things that we... Well, these things are starting, are happening, but they have not landed yet.

These are things that will happen in the next 20 years of open science. Is it going to take 20 years? Yeah, probably. With that, I'm wrapping up. I referred to the opening slide with all the names of the people. Some of the names I highlighted in these things, but all of them, they're involved in some area closely related to this. And these are the projects, the grants that have funded a good bit of the research that I have been presenting.

Thank you so much for your time.