DocEnhance video – interview with Kenneth Ruud

My name is Per Pippin Aspaas and I work at the University Library of UiT The Arctic University of Norway. And I'm here today to talk with you, Kenneth Ruud, Prorector for research and development here at UiT and also deputy chair of the board of the Research Council of Norway.

Besides these functions you are also a Professor of Computational Chemistry and you've been in this research sector for several decades. So what are the main differences, if you look back to the 1990s when you did your PhD and today when PhD candidates are embarking upon a PhD project, in terms of the research data and how to handle it?

Well, I would say that when I was a PhD candidate, the research data was largely a private set of data, of course shared amongst the members of the research group, whereas today I think everybody sees the importance of data in advancing the scientific process, and therefore the importance of sharing data, or at least to a larger extent. There are, of course, still those that consider research data to be an asset in terms of reaching publication, reaching new scientific insight, but I would say there's been a huge shift in terms of openness when it comes to research data.

Yes, openness, and also another familiar term is transparency. Sometimes research data cannot be shared for sensitive reasons, if you work with patients for instance. But you can still be transparent. Do you have some examples?

Well, there is of course the opportunity of sharing information about the data. It's also much more customary that you provide, say, information on how access to data can be obtained. Or in the case of sensitive data, whether you can extract non-sensitive information from that kind of data, and in that way at least make the world aware of the fact that the data is there, and can possibly, given the circumstances in terms of legal boundaries and so on, also then perhaps be utilised.

So you can still be transparent even though you are obliged to keep the data as such behind some sort of wall of access?

Absolutely, and of course included in transparency is also the fact that you should provide more information on how the data has been obtained. So for instance, although we are probably not quite there yet, there is an increasing push to make sure that if you provide analysis based on data, you also provide the programs you use to reach those conclusions, so that others can actually go in and verify that, yes again assuming the data are available. That you can check that the analysis is also appropriate and based on the right assumptions, because you often have to make assumptions, in the analysis of data – that these are legitimate. So that also increases the transparency in general.

We mentioned that you're part of the board of the Research Council of Norway, and the Research Council of Norway, exactly like the European Commission and other funders of research that indirectly fund lots of PhD candidates across Europe, have this emphasis on as "open as possible, as closed as necessary". There is also this slogan called FAIR. Findable, accessible, interoperable, and reusable. So I suggest now we go through them quickly.

Findable. What are the requirements so to speak, about research data? Do they have to be findable, and how?

Well, if nobody can find the data, then obviously they are not of much use. And it's also important that you can find the data in the future. So you need to have some kind of identifier that is unique for a given set of data, and that is persistent. That means it will still exist if you search for a given set of data say 20 years from now. And that I think is also important for reproducibility because, for instance in climate science, long time series are key. And that means you can't risk that part of the time series disappears. So persistency in addition to being findable is very important in terms of the reproducibility and also the progress of science in general.

Accessibility we mentioned earlier – that sometimes it can't be directly accessible but it should be findable so that people will know that they exist, and know how to access them. How about interoperability then, which is the third part of the FAIR slogan?

Combined with the fact that you need to be able to find the original data, it doesn't help if you can't really read or understand it. And we don't know what the computers will look like in 20 years from now. So it needs to be in formats that can work on different kinds of computer architectures, and can be analysed by different programs. For instance, if you end up storing your data in proprietary data formats, which then requires you to have a particular set of tools for analysis, that prevents interoperability and access to the data. So you can in a sense also make them less accessible in this way.

So that's an important point as well, and there I think we have a huge mental shift that we need to go through. We often, still, perhaps think of data much the way we as think of articles, that they are for humans. But really, data is more for computers, so things that make it easy to analyse by algorithms rather making it easy to read, are important in terms of interoperability, and making sure that we can data mine through all kinds of data across all the servers around the globe.

I would guess that in some research disciplines this will be very difficult, if not impossible, if you work in the humanities or other sectors. So there are some PhDs out there that can listen to this program and realise "Oh, this doesn't apply to me". But still this transparency applies to all. Is that a correct assumption from my part?

Absolutely, and I think we should also keep in mind that what we currently think is possible, or impossible for that matter, may become possible in the future. I mean, the big revolution now that is going on in terms of digital humanities shows exactly the fact that perhaps by digitalising old records used within Humanities, we can provide new cross-links by our ability to data mine them. So I would say any kind of data, and I think we also need to be very open-minded on what data actually can be. This can be images, this can be text, this can be numbers, all these should be made available to facilitate future research that we don't see today.

I think it's a very interesting example from one of our local research groups here that for instance used machine learning on patient journals, and actually discovered that the biggest indicator of whether a patient may have complications after an operation was what the notes made by the nurses in these patient journals said, not necessarily temperature or other physiological variables.

How then about the fourth part of FAIR, reusability?

This is of course, to some extent, at least the way I formulated it previously, related to interoperability – the fact that we must be able to reuse the data in other contexts. They may be created for a given purpose, but by providing enough metadata on the context of these data, they can be applied to new areas of questions. And that, I think, is also very important in order to also make science more effective. That instead of several researchers gathering the same kind of data but possibly for different purposes, we use the same kind of data for multiple investigations. But again, this requires a new way of thinking, but at the same time it creates vast opportunities. And I think it will transform the way we will do research in the future. To a larger and larger extent, we will base ourselves on data science more than necessarily gathering data for a particular purpose.

But this sounds like a lot of work. I mean, if I were listening to this and I was at the beginning of my PhD, I'd say "Oh, how, how will I ever find the time?" But this investment in ensuring all this quality in your research data, is it worth it for a PhD, would you say?

Absolutely. I believe that in the future the article, and thus the analysis of data, will probably play a smaller and smaller role, but the quality of the data will be more and more important, because that will be what people will use. And once you make it a common practice to cite data, and this is now

possible by giving data unique document object identifiers, investing in a good dataset that others want to use will improve your scientific standing, and have a much higher impact on your PhD work. Normally while you will do your PhD work you get a few papers, write a monograph, or collect them into a PhD thesis, and then it's there and some people may cite it. But your dataset may live for much longer than I think in general the article will. But that requires that your data is of high quality, so although yes, it is a lot of work, I think it's an investment well-made.

Investment for eternity, it sounds.

I would hope so, yes. Definitely, yeah.

You mentioned the importance of having good research data as a basis for whatever you do as a PhD candidate, for your research articles, etc. But would you say, as a reader of research articles – I mean, every PhD candidate has to read lots of other papers – would you say that papers that do not cite any data are not trustworthy?

No, I wouldn't go as far as that. I mean, the basic premise of science is that we have great confidence that people doing science have learned methods in an appropriate manner, that they follow a general research code of ethics, and don't put bias into the way they analyse the data. Having said this, no one is error-free, so errors can happen. And we do know that there are those that try to cut corners. So if a paper surprises you and the claims made cannot be substantiated by data, then, clearly, this is a warning sign, I would say. So I think for the benefit of science, again having the data available would make it possible for you to go in and make your own assessment of the data, see if you reach the same conclusions as the authors, and make sure that they haven't made an error in their analysis. And all these things can also be, not be . I mean, we do have the peer review process, but also that may miss some of these incidents. And again here also transparency is so important, because without being able to verify the original data, cheating is of course much more easy to do.

So again that's an important point. And in my field – I do computational science – a long standing controversy in the field on two different programs giving different physical results, turned out to be due to differences in assumptions on some of the modelling parameters. And only when both groups were willing to open up the code and see the differences they made, which were not central in the theory per se, it was possible to show that there was no discrepancy. And again, that means, in that particular case this was a debate that went on for seven years, and that's not the way to advance science. I mean you could have done great science for seven years instead of fighting over that particular issue. And just because the data were not, and the program was not available.

The investment in time spent by a PhD in making quality data that you prepare for publication when you finish your PhD, will this also pay off in terms of the ability to get a job? I've heard of something called the DORA declaration?

Yes, so the San Francisco Declaration of Research Assessment tries to, or it addresses this issue that the article is not the only way of contributing to science. There are so many other important contributions to science, like data. So, the DORA declaration states that when you are being evaluated for a researcher position, you should consider not only the papers you published but also contributions like computer programs or research datasets and so on. And that is for the purpose of exactly giving credit to the time you spend on producing high quality data. And as I said, this can be measured in terms of the impact the datasets will have, and we're working on ways of making it clear what it means that the dataset is of high quality. How can we verify this?

And there is work at the European level with the European University Association that tries to set up all these kinds of different dimensions that you can contribute to science. And this also includes things like outreach and so on, and innovations and so on. So it really tries to put emphasis on the point that a researcher's contribution to science is not defined by articles alone, but it's so much more. But this is a work in progress. Several institutions still haven't signed the DORA declaration, but

there are more and more institutions that sign this, so I think this is definitely going in the right direction.

Kenneth Ruud, thank you so much for your time.

My pleasure, thank you.