DocEnhance video – interview with Michaela Aschan

Transcript of video

By Per Pippin Aspaas and Michaela Aschan, UiT The Arctic University of Norway

Date: December 2022

My name is Per Pippin Aspaas. I work at the university library at UiT the Arctic University of Norway. I'm here to talk about Data Management Plans and I'm joined by Michaela Aschan, a professor in fisheries, biology, and management and also vice dean for research education at one of the major faculties here at UiT.

First of all Michaela, what is a Data Management Plan?

Oh it can be rather short and simple but it can be complicated as well. So it depends on how big a project you are in. But for a PhD project, I would expect that you have a very clear and good name for your dataset or your datasets, and you probably refer to the project that funded this, so you would have a reference that complies with the other project members as well. But then you need to describe your data: How did you acquire these data, and what kind of variables do you have in here.

And very often when we talk about Data Management Plans we think of datasets: Numbers or some kind of descriptive information like lengths of animals or something like that. But also text can be considered data if you're a social scientist and have an interview or something in place.

So you describe your data, what variables you have or what format you have for your data, how it has been collected and so on. And then you should also try to state if you follow certain standards when sampling or collecting them. And if you have several datasets you are likely to provide how each relates to a metadata base (or source) with information on where to find what.

The metadata base may also be provided in the project if you provide a lot of information. I would like to highlight here that if you work with big data for instance, you have a special challenge in identifying all the data that you wish to use that is out there. And by combining new datasets you actually develop new information. And especially for social scientists, but also I see more and more in technologies and natural sciences, that you need to get into and consider the GDPR, to ensure that personal data protection is in place.

Next you would need to look also into the data sharing, if this is open data, and include information on where you are going to archive the data. And finally also perhaps indicate how the data can be used, e.g. for visualisation or in decision making and so on.

It sounds to me like this is a big task for a PhD candidate at the very beginning of a project. And it also resembles somewhat the entire project description. Is there a link between the two?

I recommend that you include the Data Management Plan in your project description. Because that ensures that you can revisit it regularly. I suggest that you revisit your project description, including your Data Management Plan, annually when you report to the faculty that you are in, in time on time and in shape.

And that gives you the opportunity to elaborate and develop this Data Management Plan as a living document throughout your PhD period of three or four years.

A living document you say. So it is possible to change your mind?

I think that things are moving fast these days, and even if you have a plan, it's a plan and you need to update it during your project, so that it's accurate enough and precise enough, so it works as a kind of a guideline for you and your supervisors (very often many of them). So you have something written, and so that you have agreed on how to perform.

For instance, the discussion on open data and open science. It really is good for you as a PhD candidate if you can provide a DOI number and say there is the dataset that I made or developed. And if you have this written it's easy to follow that plan, and you may avoid ending up in a kind of awkward situation where one of your supervisors suddenly finds out that, no, we keep and we protect this data and we apply for IPR rights or something. Because that will prolong your own process and the finalising of your thesis.

Yes, so there are several ways then in which this can actually help you progress through your thesis and be finished in time?

Definitely, and I think – especially I think many PhD candidates, when you go there and talk about the Data Management Plan with your supervisor, the supervisor might not always be very well prepared. So you may perhaps end up in the situation that you have to go to your university library or the homepages and find out about what set of rules you have, and what kind of guidelines you can find, and even present it to your supervisor. And when you are there then, during the lifetime of the project, make it more specific.

Can it also be helpful to you after you finish your PhD?

Definitely, because if you have uploaded this – now, let's say a database … If you have an EU project or your university doesn't have its own repository, you would perhaps put your data in Zenodo or another open source database, perhaps connected to your field of study, and you would then get a

DOI number. So that is a referable dataset. And of course making this available, perhaps after a year or so, may show that three or five people, or even a hundred have used your data for further publication and use, and that's of course a good thing on your CV.

You mentioned the existence of different repositories. Some are general repositories and international ones, some are perhaps located at the local university library, but they are also international in terms of DOIs, so that it can be searched for and found anywhere in the world. But regarding the research Data Management Plan: Do you actually need, as you start your PhD, to plan ahead to the publication of the data? Do you need to plan everything till the very end?

I definitely would suggest that you do so. And as the plan is there – it's a plan so you can change it and adjust it to ensure that it fits your aims. But bring it up annually at least with your supervisors. And perhaps when you bring up your project description – go through it and ensure that it is in line with all your expectations so that you're all on the same page. And then it saves you from a lot of hazards or barriers in identifying how it could be done in the last minute.

We have also identified examples where both papers and data have been uploaded in a place where you don't get the DOI number. And that is of course not then applicable or compliant with for instance the Norwegian national research council rules or the European Council.

What about the Data Management Plans as such – is that more like a personal thing between you and your supervisor then? So that's not what you publish afterwards just to make that clear?

I don't consider a Data Management Plan something that you publish. But you will … the data management plan, the context of that plan – it is very nice to have it attached to your data as an introduction and as a description of your data. So much of that work is done in your data management plan. So when you upload your data, you need to give it a name, you need to tell how the data was acquired, so you need to provide some information and that is of course then updated in your latest version of your Data Management Plan– and makes it easy.

I also want to highlight the fact that very many journals these days requests that you provide the data and the code so you can repeat – anybody can repeat the exercise. And in that sense it is also a good thing to have, ready to go for anybody to use.

Most PhD candidates plan to themselves collect data, or they collect them in a group – in a research project. But still, it's data that was assembled during the project period. But sometimes I know that PhDs, they merge their data with other people's data, or even use only other people's data. Yes it's getting very common now, because with the PhD programme of three years – you might not even have the time to go out at sea. And now during Covid-19 we see that many students have lost their whole period of going to the field. They don't get access to the surveys and so on. So it's

more likely that we have even more students working on big datasets that they receive from statistics, e.g. from the FAO or from national statistics or even from colleagues. In health science and when you work with machine learning and so on, you very easily collect data from different sources. And of course when you then develop your Data Management Plan you have to be really careful to explain what data you have received, where you got it from, who owns it, was it openly accessible or not, and then show and demonstrate how you compile new datasets by using this previous data.

And of course in this connection I again want to stress the issue with big data and machine learning, that when combining datasets you may easily end up in a situation where you reveal information about individuals that you shouldn't. So this is something that you should keep in mind.

The GDPR – and it is now being talked more and more about – what is that in general? I mean what is it that you need to be aware of?

So in general you would say that the GDPR, General Data Protection Regulation, it is to ensure that the personal information is not distributed to anybody. So if my information is about my address, my health condition, and is somehow combined – I'm even a Finn, and I'm probably the only one living on that street being a Finn, and if there is some health information – even if my name is not there – the information about me is available. So by combining my address and my health condition and my nationality that might have been in three different datasets, when they are combined there is delicate information about me.

So in the Data Management Plan this has to be described: How to avoid these delicate situations where you reveal too much about somebody.

How then about the rights of data? I mean, if you merge with other people's datasets, can you be sure that you are allowed?

You should always ask for permission and explain for what kind of a project you are using the data. If there are openly accessible international and national statistics it's perfectly fine. But if you get data for instance from a health institution, from a private institution – all environmental data should actually be openly accessible by now from all your national institutions. But still the institution would like to know for what purpose you're going to use this data, and it's also a way for them to promote that their data is used and made useful for the community.

So it's a good habit actually to tell the provider of the data that you wish to use it, for what you want to use it, and also for what time frame. As soon as there is some delicate information or if it's privately acquired, you are usually asked to delete the data after project end. But with the open accessibility – and now making data available – you would of course try to bargain and say I ensure

that there are no problematic issues with this data, that I don't break the GDPR and that it will be fine in making it openly accessible.
One thing: if your data is not openly accessible, you need to argue why. So let's say you go out there and interview businesses or you get access to their books for bookkeeping. I say okay – some do better than others. The businesses don't want you really to make that data available. That is a very good reason for not making data available. But you can still conclude on the basis of having access to that information.
But you wouldn't explicitly say this business is doing bad, or this one is doing wrong or you are doing better. But you would perhaps indicate the success rate among businesses in a certain field, and thereby you would probably promise to delete the data after the project.

There are some disciplines where perhaps the PhDs don't feel that they are into data at all because they work in humanities for instance or arts. Should they write a Data Management Plan if they don't plan to collect any data?
I think that one should think of data as a bigger concept than just numbers. And of course if you work as a social scientist and interview people and so on, you will probably transcribe that interview and put it into a text analysis. So in that sense your text is a set of data.
But if you work with arts – if you provide a piece of art: pictures, videos of your installations or other things may be relevant for documentation. And in that sense, a Data Management Plan for how to document and keep track of these pictures may be a good solution.

Michaela Aschan, thank you very much!
Thank you, it was a pleasure.