Q: Therefore we are also using it. (laughs) #00:00:03-8#

R: Yeah. #00:00:02-8#

Q: Okay. #00:00:05-2#

R: I...I just...I actually just scanned the consent form. I don't know if it's really important that you have them like now-now, before you start. #00:00:11-0#

Q: No, it's not that important. I think you are... #00:00:19-2#

R: I am consenting, yeah. (laughs) #00:00:21-4#

Q: (lacht) So, I started to record now and, yeah. I think if you have no further questions regarding the project, we can just start with the first question block. #00:00:31-9#

R: Yeah, so just for the record, I saw that you sent me the interview guide, which is really nice, so thank you. The only thing is that I have...I've, I've been extremely busy in the last weeks, so I only, like, just had a first look at it. So I...I looked a little bit at the type of questions you're going to ask but I...I haven't had time to really deeply prepare them already. #00:00:52-3#

Q: No problem. Okay. So in the first block, I would like to take a closer look at secondary data use from your perspective as a data user. So then the first question would be: How often did you use secondary data in the past? From your lab and also from other labs. #00:01:12-8#

R: Yeah, let me (...) if I'm going through my publication list...I think in the times that I used data, I think at least a third is with, yes, secondary data, which is collected either by me or by someone else. So, I...I guess that's also your...your definition of secondary data, right? It's, like...because I have one large dataset that I collected that I have already (...) in several different papers about that dataset. #00:01:49-0#

Q: Okay. And only you re-used these data or also PhD students and (...) students #00:01:55-3#

R: Also just other people in general have used it. #00:01:59-0#

Q: Okay. Mhm. #00:02:01-6#

R: I think, you know, not my PhD student (thinks). Let me see. I can actually just... (...) check it right now. (...) Yeah, so...this is...new data...new data...new data (she skims through her list). Maybe I overestimated how much secondary data I used. (...) Secondary...this is also new data...secondary... Yeah, I'd say a quarter to a third of the projects are with...with secondary data, I think. #00:02:54-6#

Q: Okay. Good. And for which purposes did you use these secondary data? #00:03:04-0#

R: Mainly to ask new questions. #00:03:05-6#

Q: Okay. #00:03:03-7#

R: So it were often large datasets that you can use for different purposes and, yeah. So to answer new questions. So I'm not sure if this would fall into your definition as well, but I'm also involved with a couple of reproducibility projects in which we really sort of take published data in order to see if we can sort of find the same numbers as they report in their paper. So it's one project but there is ... how many papers did we check (she reflects), I think...well, I can look it up but I guess twenty papers or something that we really use all the data to check whether these results are reproducible. #00:03:47-8#

Q: Okay. And which criteria have you chosen for the papers to use? #00:03:57-6#

R: They were from one journal, from... *[Journal 1]* is the one I'm looking at, and they have an Open Data badge. #00:04:05-0#

Q: Ah, okay. #00:04:03-2#

R: And I think within a certain year. But if you...yeah, if you want to know, then I can look up the details later. #00:04:12-9#

Q: Yeah (lacht). I'm just interested in it. (laughs) #00:04:15-1#

R: Yeah (lacht.) #00:04:14-3#

Q: Okay. And, yeah, what kind of metadata would you need to optimize your work? So regarding new questions and re-analysis of the... #00:04:28-4#

R: Mm. Well, the...the first and maybe most important thing, I think, would be a detailed codebook. So really a list of all the variables in the dataset and what exactly they mean and what skill they have. And sort of what are the potential possible values on this variable. #00:04:47-9#

Q: Yeah. #00:04:46-8#

R: Yeah. #00:04:48-2#

Q: And you would also need the argumentation, so-to-say, behind the variables? So why the authors used this kind of scale and not another one? Or would it be sufficient just to provide the scale? #00:05:03-6#

R: I think for the...the purposes for which I've used secondary data now, the arguments behind it would not necessarily be super interesting. I mean, yeah, no. #00:05:17-0#

Q: Because for PISA and so on, for...for these large datasets, you get also the argumentation behind, right. #00:05:25-5#

R: Yeah, well...for me, it's not necessarily the argumentation I would need but more maybe the exact method with which the data are collected. So that...that might help in, like, especially if...if it's not my own data that I'm using for secondary data analysis but someone else's, and I want to...like, for these reproducibility projects, it doesn't...in the end, it doesn't really mean...really matter what it means. Because I'm just checking if I can get the same number, so it could be anything, it doesn't really matter. If I want to use it to answer a different substantive question, then I do really need to know what the variables exactly mean.

Yeah, but I don't think I would necessarily have to know why the authors chose (...) a certain scale (…). #00:06:08-3#

Q: Okay. And would you also be interested in having the scripts, so... #00:06:14-4#

R: Definitely, yeah. (laughs) Yeah. #00:06:19-0#

Q: Okay. So if I get you right, you would be interested in the codebook for raw data or aggregated data, both...? #00:06:27-0#

R: Mainly for raw data. It depends a little bit on what the aggregated data is. I guess if they aggregated the data, it would be great to have an analysis script that has those steps included as well. So, ideally the...the...the rawest data, and with rawest data I would specifically mean the data as they were first, like, entered into some sort of spreadsheet kind of format. But ideally, every notation or aggregation or...or summary statistics that are calculated based on that should be included in an analysis script, so that also those facts were reproducible. And in that case, you would need a codebook for the summary data. #00:07:08-8#

Q: Okay, mhm. Good. And are there other methods you know regarding secondary data use which you haven't used in the past but you know about them? And you would say, okay, for...for these methods, if I would use them in the future, I would need more than just the codebook and...and the script and so on. #00:07:33-8#

R: Mm, what kind of methods are you thinking about there? #00:07:35-0#

Q: For instance, illustrations, meta-analysis, systematic reviews, yeah. #00:07:46-5#

R: Yeah, I think... (reflects) I think most of the things that you would want to do with secondary data, for most of them, it would be enough to have the raw data and a detailed, yeah, detailed information about what exactly is in the data. And again, but maybe I'm overlooking a particular kind of use of secondary data but for...for a meta-analysis, for instance, I'm not sure if you...yeah, yeah, maybe some argumentation about why a certain scale was used might help in determining what the quality of the study was. But yeah, no, again I think...I think the main things are the raw data that is well documented and preferably an analysis script. #00:08:31-9#

Q: Mhm, okay. Good. Then we switch to data sharing (laughs). #00:08:40-4#

R: Mm (agreeing). #00:08:40-7#

Q: Okay. First one. I’m sorry. Ah, no. I missed a question in the first block. What kind of data are you generally using for the different purposes? Are these more behavioral data or rather physiological data? #00:09:05-6#

R: In my case, they are usually archival data in the sense that they're often data about journal articles. So, for instance, in one of my projects, I am looking at *[project type]*. So...so, yeah, my...my participants are the *[type of statistical coefficients]*. #00:09:23-1#

Q: Yeah, okay (laughs). Good. And do you or have you perceived any differences in the documentation quality of these archival data? #00:09:37-6#

R: Yes. So if I...so let me maybe clarify a little bit. So if I collect data, it's often archival, so I...I go through large...yeah, a large bunch of scientific articles to...you know, systematically document some...some kind of element. If I use other people's data, it's either the same type of data that I collect because I'm then doing similar research questions or want to say something about just meta-science things or it's for reproducibility purposes, and then it's often data of psychological experiments, so then it would be behavioral. #00:10:16-7#

Q: Okay, mhm (agreeing). Good. And there are differences in documentation quality, so...? #00:10:26-3#

R: Oh yeah, sorry, that was your question, yeah. Mm (reflects), I think, yeah, so for...but this is not necessarily very representative. So the...the...the data that I use for meta-research are often either collected by myself, and I'm very focused on documenting my own data, and of course, I know more about my own data than about someone else's data. So those are generally documented well. The documentation in...yeah, the more applied studies can vary widely. So yeah, I've...I've seen, like, perfectly documented both datasets and scripts, and...and things that are completely incomprehensible. (laughs) #00:11:09-6#

Q: Okay. Of course the script is missing and codebook or only thing is missing or... #00:11:15-7#

R: There's actually...about Psychology ...this is the project that I'm working on now with the reproducibility is still in progress. I'm actually not sure what the study of that is at this point, but there has been a similar project that has systematically documented this. Yeah, maybe if you've seen it, it's by *[name of the project manager]*, (...), and they, yeah. So I think, so far...yeah, subscripts (?) are generally not shared or not often. And the data are and the documentation of the data is, in my experience, usually pretty minimal. #00:11:55-3#

Q: Okay. Mhm. Good. Now we switch to data sharing. Here I would be interested in what sorts of metadata you usually provide about a dataset when you upload. If you have uploaded in the past. I don't know. #00:12:19-1#

R: Yeah. I actually have a (...) checklist from a...let me see. (checks it up). Right here, yeah. Yeah, so what I share are both...is both information about the data itself but also about the way it was collected. So, for instance, what's in my metadata are, for instance, author roles, so who did what in the project. Where the data came from, so you know, what...what...what the sample was and everything, so this is pretty much a very short summary of a method section almost. Whether there was ethical approval (sort...) and obtained. A history of the data, so when was what collected and by whom. And then just a list or so of file descriptions because usually in my case the...the data consist of multiple files, so I always make an overview of, like, this is the file name, this is what's in the file. And...yeah? #00:13:29-8#

Q: And...sorry if I have to interrupt you. What means multiple files in this case? So you have multiple (time points or...)? #00:13:39-1#

R: Mm, well for instance...yeah, multiple data files or a raw data file and different analysis scripts for different purposes. And maybe also summarized data which you effectively or...or in theory could also obtain by just looking at the raw data and the script, but for the sake of completeness, I also just upload a...a file that I created. And sometimes materials as well.

And...and, and, and then the actual description of the data files itself would be just an overview of...yeah, pretty much as I described the...the file or the variable name and exactly what it means. #00:14:25-4#

Q: Mhm (zustimmend). And this is something you're describing in a separate text file. #00:14:31-2#

R: Yes. #00:14:36-7#

Q: Okay. You're doing the work most of the researchers don't want to do. (laughs) #00:14:44-6#

R: (laughs) #00:14:48-6#

Q: Yeah. Do you think that these metadata are sufficient for re-using your dataset? #00:14:54-9#

R: Yeah. Yeah, I know, I think...because actually people have used my data as well and, yeah, seemed to be able to do so. So that seems like a good sign. #00:15:07-0#

Q: Good. And how many people have re-used your data? Do you have a...? #00:15:10-6#

Q: Mm, that's a very good question. I think, for one particular dataset I have in mind, I think at least three different people. Yeah. Yeah. But I...yeah, it's...it's hard to keep track of...people also don't always, like, sometimes I find out after they've actually already done the analyses, which again is probably a good sign that they were able to use my data without my help. (laughs) #00:15:41-3#

Q: True. (laughs). And what kind of data do you produce? So what's the thematic background of your research? #00:15:49-5#

R: So...so my own background is in Psychology. But my research is really focused on...on meta-science, so I'm doing a lot of research about how we can increase both the...sort of the...the computational reproducibility both and the replicability...how can we find out more about the truth in Psychology? How we can use, yeah, pretty much Open Science-related strategies to...to get there. #00:16:19-8#

Q: Ah, okay. And your datasets that the people re-use are related to these topics, right? #00:16:27-4#

R: Yes, yeah. #00:16:25-2#

Q: Ah, okay. I just want to know because you can perhaps afterwards say something about what kind of data are important for...for re-use. (laughs) #00:16:38-2#

R: Mm, yeah. #00:16:40-8#

Q: Good. Then: Have you used certain metadata standards for annotating your data in the past? So, for instance, DDI or Dublin Core, something like that? #00:16:57-8#

R: No. No, we have sort of, within our department, tried and...yeah, informally developed some sort of, yes, standard is a big word already. We also have within our faculty rules about data sharing or at least data archiving and making sure that at least two people have access to it. And there we have a checklist of things that need to be in there. But those are not guidelines that are used at other places. #00:17:27-6#

Q: Ah, okay. Are these guidelines publicly available or could you send them to me? #00:17:33-6#

R: I think so, yeah, I can definitely send it. I'll just send you a checklist that we...let me see if I can find it now. I was actually just checking it out as well. Checklist, there we go. (checks it up) Let me see if I find here...Yeah. So this is a checklist for data audits. So what we do is...so I'm...I'm part of this committee as well. So we have rules at our faculty that we have to...yeah, share your data, sharing is not necessarily the right word but we want to make sure that data are...are safely stored and that multiple people have access to it. And there are some new guidelines to do that and we audit people randomly to see if things are going well. So every...every semester, I think, we're auditing two researchers from every department within our Faculty of *[science type 1]*. And these checklists are used during the...the interview where we check whether people have...have done what they (...) did. #00:18:40-4#

Q: And do they use the checklist? #00:18:42-8#

R: Well, we get they impression that they start using the checklist the moment that they find out that they are going to be audited because we always announce that we're going to be auditing them. But the...the departments do, yeah, there are differences between departments but they do develop their own data sharing standards and it seems to go in the right way. But at this point it's also a lot of extra work for researchers and, yeah. So, there is an extra hurdle to overcome. #00:19:05-4#

Q: Mhm, okay (laughs). Good. And last question: If you were to create a metadata standard, what do you think is the most important information that should be included in such a standard? Perhaps you can just think about this question in terms of the JARS. (...) So the Journal Article Reporting Standards from the APA. Because I think that most people are familiar with this kind of standard and... #00:19:42-9#

R: So I'm actually not, I think, or maybe I don't know the term that you use. So what...what kind of categories do you mean? #00:19:50-1#

Q: Just which kind of data should be provided in a research article. So how to write an introduction generally, which data should be provided in a method section and so on. #00:20:00-4#

R: Mm (understands). Yeah. I think, within an article, I don't necessarily think that you have to report every single detail of every little thing that you did and every...every tiny secondary data analysis. What I do think is that then, in your article, there should be a link somewhere that refers to supporting material that does contain all that information. Is...is that the direction that you were...? #00:20:29-2#

Q: (...) the direction that you would like to think about (laughs). #00:20:32-9#

R: Yeah. Yeah, well, so that's pretty much, like, I think it's really important that if I want to dive into an article deeper, that I have the ability to do so, preferably without contacting the original authors because, yeah, there is, both in personal experience and empirical data shows that, yeah, contacting the authors usually doesn't resolve many things/anything (?). Yeah. So...so pretty much all these elements that I said about raw data, scripts, metadata, those should be (...) but not necessarily explained within the article, I think. #00:21:08-4#

Q: Yeah, of course. And would it be enough for you to have just one link within this research article? Or would it also be possible just to link different keywords to different information in the dataset, so that you have perhaps a link from the research design directly to the variables in the dataset? #00:21:32-1#

R: Mm, yeah. I guess that could work. But in the end, I...I think as long as you're clear enough about what your supporting materials contain, like, if...if there's at least some sort of summary file, like a readme somewhere that explains, like, okay, this is in my data package or whatever you want to call it. Then I think it should be sufficient. Or at least you should aim to make it sufficient, yeah. #00:21:53-3#

Q: Mhm, okay. Good. Nice. Then I thank you for your time. #00:22:03-1#

R: Yeah, you're very welcome. I will send you the consent forms and the...the data file, the data check audit list thing, yeah. (lacht) #00:22:12-0#

Q: Thank you very much. Have you any further questions before we end? #00:22:16-0#

R. Mm (thinks). Yeah, so let me know if...yeah, I'm...I'm interested in the project. Let me know what happens. #00:22:24-0#

Q: Yeah, no problem. We will do that. (laughst) #00:22:27-2#

R: Alright. #00:22:29-1#

Q: Okay, thank you. #00:22:29-3#

R: Well, thank you. And talk to you later. Goodbye. #00:22:34-1#