Q: First of all, we would like to elaborate on secondary data use from your perspective as a data user. How often you reused datasets from your lab and other labs in the past? Could you quantify your specification by providing the relative frequency for reusing data compared to producing primary data?

R: 100% because I am not producing primary data.

[Hier startet die Audioaufnahme.]

R: (...) also the meaning of the numbers are important because sometimes people reverse the scale without mentioning that and then also, I guess, like, time stamp for collection of the data because it's not always the case that data have been collected at the same moment, e.g. (...) results... (thinks) Again, I guess also, like, the protocol behind, behind the collection of the data, whether data (...) collected at the same place by the same person using the same instrument (...) #00:00:51-3#

Q: Okay, so, so everything regarding methods, so to say, procedure. #00:00:55-7#

R: Yes, yes, indeed. #00:01:00-1#

Q: Okay. And for which specific purposes have you used secondary data? I just missed what you said... #00:01:13-0#

R: (laughs) Well, both for teaching and also for illustrating the methods that we, we test or create within my own research group. #00:01:25-0#

Q: Mhm. What kind of methods are these? #00:01:27-5#

R: These are mainly methods for dimension reduction, like principal component analysis, but then in the context of high dimensional data (...) selection , (it’s) towards machine-learning more. #00:01:45-0#

Q: Ah, okay. Mhm. Okay, nice. So then we can (switch?...) to the meta (-data...) So what (else...) do you use/need? #00:01:57-3#

R: What...? #00:01:57-0#

Q: So, (more...) when I summarize it now, I would say, you need an extended codebook, something like that? #00:02:07-8#

R: Yes. Yes. #00:02:10-6#

Q: And you would need the scripts for analysis and data preparation also. Is that right? #00:02:20-1#

R: Erm, yes, in case that the data have been processed before making them available, yes. #00:02:31-0#

Q: Mhm (agreeing). #00:02:32-6#

R: So, preferably, I would have the raw data with scripts used for processing the raw data but I see that in many cases, people just already report processed data and it's very hard to get back to the original raw data. It's kind of repeated because there's often quite a lot of information loss. #00:02:58-0#

Q: Yeah, that's true. Okay. Good. Then the next question would be: Do you know about any other methods where you need secondary data, you haven't used these methods right now? What you know about them and would you say that you need other metadata than the ones you have already mentioned for these methods? #00:03:24-8#

R: (thinks) #00:03:29-9#

Q: So, for instance, meta-analysis or something like that. #00:03:33-6#

R: (reflects) Oh, I think in meta-analysis, people often work at a level of the statistics that have been reported in (...) and then you would need the code that has led to the scripts, preferably, because I think there's a big difference in reporting a statistic and actually how the statistic was calculated because people sometimes then deal with problems of missing values without reporting how they treated (them...) e.g. So, well, I would really prefer we just have the raw data with all the codes that led to the final (results...) that's reported in the paper. #00:04:23-9#

Q: Yeah. Okay. And would you also need a link to the paper, so would this also help you? #00:04:33-7#

R: Yeah, it would help definitely. But (laughs) that may be an issue because (they publish...) probably. (...) let you provide the link to the site of the publisher but then (it's...) not sure that the paper is really available without a cost, yeah, yeah. #00:04:51-2#

Q: Of course, if it's not open access, yeah... (laughs) You have a problem. (...) Good, nice. Then let's come to the second question. (thinks) Oh, no. We are already at the third question. What kind of data are you generally using for the different purposes? Are (...) more physiological data, for instance, or more behavioral data? And then I would also like to know from your perspective which kind of these data are best and which are worst documented in your point of view? So, do they differ in quality, so to say? #00:05:36-6#

R: Okay. Yes. Well, I, I think I use for 50% I use bio-medical data, (...the) data and also (...) data, and then the other 50% is behavioral data. So, (typically...) like personality questionnaires. And concerning genetic data in medicine, they have I think already taken a (...) a lot of disciplines are creating repositories with (raw data?...) standards for uploading data there and providing documentation. #00:06:26-4#

Q: Yeah, the BIDS standard, for instance, right? Or (...) for fMRi data. #00:06:33-0#

R: Aaahh (seems to understand). Yeah, for fMRi, I don't know but for the genetic data, there is, there is a standard but I don't know the name of it. You have the NCBI website where you have the (gene...) expression of omnibus (?) (...) #00:06:55-1#

Q: Okay. #00:06:57-6#

R: And you will see that there they are all data are stored under different formats (...) and they are the same for every dataset that has been (...put there?) #00:07:14-2#

Q: So (...) is rather high? #00:07:20-4#

R: Excuse me? #00:07:19-3#

Q: So quality is rather high or...? #00:07:25-5#

R: Mmm (thinks). Yes, yes. #00:07:28-4#

Q: Okay. #00:07:30-0#

R: Yeah. #00:07:32-5#

Q: And it's also higher than for behavioral data that you use? Or would you say it's rather the same? #00:07:41-4#

R: Well, it's higher because for behavioral data, in my experience, often I have to email (...) the data (laughs). #00:07:52-1#

Q: Okay, good. And the last question of the first block. Where do you think are the reasons for these differences in documentation quality? Do you have an idea? #00:08:09-5#

R: Mm, well, I think it's probably related to the publishers. (So in...) bio-medicine, there are often strict requirements concerning communication of the data together with the paper. And this is especially true for the high impact journals. #00:08:34-6#

Q: Okay. So they already require something like a standard, so to say? #00:08:38-1#

R: Yes, yes, yeah. And many of these journals do not allow publication without making the data available. #00:08:53-5#

Q: Mhm (agreeing), yeah, true. Okay. Good. Now (we...) switch to secondary data use from your perspective as a data provider. And there I would like to know what sorts of metadata do you generally provide about a dataset when you upload, if you upload, I don't know? (laughs) #00:09:15-3#

R: (laughs) (...) the scripts that I use to create synthetic data, so (the...) simulation studies. #00:09:29-5#

Q: Yeah, okay. #00:09:31-4#

R: Mmm, but, well, this kind of, it's also data but it's not empirical data. (laughs) #00:09:38-5#

Q: No problem. #00:09:35-0#

R: Yeah. What I (...) there is just provide the script together with the seed for the random number generator because that also influences a bit the results. But I do not provide the data themselves, and the main reason there is that it's just too much. It can become several gigabytes, just for one study. #00:10:05-2#

Q: Ah, okay. And (...) repositories are not sufficient to... #00:10:13-8#

R: No, they're not sufficient and I, well...the code is there, so if you push on the (…) run, you will (generate data?...) yourself. On the other hand, when the computing environments are changing constantly (...) changed but also software packages change and sometimes it's (not) possible to create the data again. #00:10:41-8#

Q: What kind of software packages do you use for your analysis? Mainly R or...? #00:10:44-7#

R: R and MatLab (?). #00:10:46-4#

Q: Okay. Mhm. Yeah. Yeah, in both cases, that can be a problem, right (laughs). #00:10:57-1#

R: (laughs) Yes. Indeed. #00:11:01-1#

Q: Okay. Do you think that providing these scripts and the seed for the random number generator, are these things sufficient to reuse your data, so to say? #00:11:18-2#

R: Well, I think it's also good to provide some description. (...) simulation study was set up because, (well the code...) often becomes rather long and messy. It requires quite the big effort to go through it. Well, in principle, well, someone could just push the button and get everything but then you have meaningless data, so... #00:11:43-3#

Q: Mm (agreeing), yeah. So does that mean that you do not document your code, so that you write, okay, here we do this analysis and here you can do this one, and so on. #00:12:00-5#

R: I do put a bit of comments in the code. (It's a?...) sort of documentation, and furthermore, I usually, well, I publish everything on GitHub. #00:12:10-6#

Q: Ah, okay. Mhm. #00:12:12-4#

R: And there on the main page for each project, I provide a description of what has been done in the study, which files (...) each of the different steps. And also what the order of the files it is, so there's a first file which you have to run and then a second one and so on. #00:12:30-4#

Q: Okay. Nice. #00:12:31-2#

R: And sometimes I also, like (I have the...) process data, I also publish those results on GitHub, (...) e.g. with genetic data when the paper ends with a final selection of a few candidates, I provide the list of candidate genes. #00:12:56-5#

Q: Mhm (agreeing), okay. #00:12:57-0#

R: Yeah, but not the data themselves because these are not mine, so...yeah. (laughs) #00:13:03-3#

Q: Good. Okay. Then...have you ever used a certain metadata standard in the past for annotating your data? #00:13:17-3#

R: No. #00:13:14-0#

Q: So, since you have no data, probably you haven't used (...) #00:13:20-7#

R: No. (laughs) I don't. #00:13:23-9#

Q: Okay. And if you were to create a metadata standard on your own, what do you think are the most important information then? That is which additional information on your data is most important for an optimal reuse? Perhaps you can think about this question in terms of the JARS, so the Journal Article Reporting Standards from the APA because I think all researchers know about this standard and (...) you can just transfer these categories to the data documentation standard which you imagine or would imagine. #00:14:02-1#

R: Okay, well, standard (...) (laughs). Mmm (agreeing). I must say that it's been many years since I consulted the APA manuals. (laughs) #00:14:17-7#

Q: Okay (laughs). #00:14:21-1#

R: I'm not very sure what they say with respect to reporting. So I know that, like, in reporting the statistics, they say (about...) reporting the degrees of freedom (and so on...) (laughs) but concerning data, no, I don't know what the standards are. #00:14:39-6#

Q: Yeah, but if you, so if you were to create a standard, what would you (...the basic) standard you could imagine for your data? So what would be the best way to document your data that others can optimally re-use them? This is (...) the way you are documenting your data right now, is it already the gold standard or do you think there could be done something more? #00:15:14-6#

R: Well, I think more can be done, so it would be nice if there is a standard that is used by everybody, so that when you're working with public data, you kind of recognize...for a new dataset, you recognize the environment immediately, and you know what kind of information is available and in what ways you can require specific variables or specific data on a specific topic, and also that the documentation provided is, is just the same for every (data...) because in my experience, often, you lose a lot of time just trying to, to see what is provided. If I go to Google, it's very different from the Gene Expression Omnibus, which is then different from how it's done on OSF, so... (laughs) #00:16:12-2#

Q: Yeah, yeah, it's true. #00:16:13-9#

R: It takes quite a lot of time to, to get into that. Mmm (thinks), yeah, well and I think having, like, the raw data, the labels of the variables, the meaning of the numbers (it's...) for me already very important. And then if there are different waves in the study, you also need a documentation on the waves (thinks). Probably if there's more information available on, on, on the samples, it would be nice to document that, like, for this dataset, we have also... #00:17:14-6#

Q: Mhm (agreeing). So (...) setting and so on. #00:17:17-0#

R: Yeah, yeah. #00:17:16-2#

Q: Exclusion criteria, inclusion criteria. Mhm (agreeing). #00:17:22-5#

R: And then I think also all the steps that you need to, to clean the data, like, with fMRi data, e.g., that's always (laughs), like a forest, it seems cleaning. #00:17:36-6#

Q: Yeah. #00:17:37-9#

R: So it would be nice if, if you have the raw data available together with the clean data. #00:17:45-1#

Q: Yeah. And then... (...) #00:17:47-2#

R: Because cleaning... and the steps in-between documented in some way, either (...) with some file. #00:17:56-7#

Q: Okay. #00:17:58-6#

R: And study protocols, I think study protocols are really important. #00:18:03-4#

Q: Mhm (agreeing). What includes a study protocol in your understanding? #00:18:10-1#

R: Mm, well, the way the study was set up and (how the...) data was collected, (how the...) sample was created, who did what, what kind of instrument was used. #00:18:28-1#

Q: Mhm (agreeing). Okay. Good. I thank you very much for your time (laughs). #00:18:35-3#

R: Yeah, you’re welcome. (laughs) Hope it (will be) useful. #00:18:41-3#

Q: Yeah, it's really good. Do you have any further questions, concerns, ideas regarding our project? If not, you can also email me in the future. #00:18:49-6#

R: Erm (thinks). Not for the moment. I think I would be very happy with the development of some standard for, yeah, for a repository for code, for providing (...) because this is not yet something which is commonly done. But I hate it as a reviewer when I get a paper where people have been creating their own methods and they have been using a lot (...), and they (did...) not provide the functions which are (the...) basis of the paper. #00:19:26-5#

Q: Yeah. Yeah, that's true. #00:19:31-2#

R: But it's...it's not yet common practice, so... #00:19:34-7#

Q: Yeah. Okay, so #00:19:36-4#

R: But this may be...may be future projects (laughs). #00:19:40-0#

Q: Yeah, I think so. You cannot manage all these things in one project. There will be upcoming projects but...we try. #00:19:49-9#

R: Yeah. Okay, well (laughs). That's good to know. Good luck with the research. If I think of something, I will let you know. #00:19:59-9#

Q: Okay, wonderful. (...) #00:20:07-1#

R: Oh, nice. Well, in my...in my group, there are a lot of meta researchers, so I don't know if you are also interested in... #00:20:18-4#

Q: Of course. #00:20:20-4#

R: Like, *[person 1]*, I don't know if you...if you're... #00:20:23-0#

Q: Mhm. #00:20:25-6#

R: Well, I can send you the email. #00:20:28-8#

Q: Mhm (agreeing), okay. Then I write him. And can I refer to you? #00:20:34-9#

R: Yeah, sure. #00:20:37-9#

Q: Yeah? Okay. (...) Thank you. Wonderful. Then we can perhaps see you in the future, in the *[institute 1]* or at your institute #00:20:49-1#

R: Nice. (laughs) Have a nice day, bye! #00:20:53-1#

Q: You, too, bye! #00:20:55-9#