Q: And in the first place, I would like to elaborate on secondary data use from your perspective of a data user. So the first question would be: How often do you reuse datasets from your lab and other labs in the past and I would like to ask you to quantify your specification by providing the relative frequency for reusing data compared to producing primary data. #00:00:31#

R: Okay. So just two sets of clarification questions. When you mean secondary data, reusing data, do you really mean primary data because if I'm doing a meta-analysis, is that secondary data? #00:00.47#

Q: Yes, that would be secondary data. What would be another understanding? #00:00:52#

R: Well, so I think really the raw data but if I'm reading an article and I'm extracting effect sizes or means and so on – Would that be considered also? (…) #00:01:07#

Q: Yeah, also. I'm considering all cases including either research synthesis methods or re-analysis method. So in the case of re-analysis, it would be, I think, rather the raw data thing (R: Mhm) and in the case of research synthesis, as you already said, more effect sizes and and aggregated data perhaps. #00:01:33#

R: Okay. (…) Yeah. That is now about the kind of studies that. So again, I just want to make sure that I really understand the question – because you are asking about how often do I use secondary data, right? So --- in in my research, right? So, okay. Alright, mh. Okay. So Mh (laughs). Of course I would say often because I – first of all I not am a statistician, so I don't usually do primary research. And – in my own research, so if it's, it's also a bit tricky now for me to define what is sort of mine research versus I'm helping colleagues analyze their data, right, so it's a little bit difficult to clearly make that distinction. And because in my own research I would say basically zero percent of the times I'm using primary data and in 100 percent of the times I'm using secondary data. #00:02:43#

Q: Okay. So you are the first with this percentage. (laughs) #00:02:45#

R: (laughs). I don't generate primary data. I I I use secondary data for meta-analysis, for illustrations, for re-analysis, and – I guess once in a while (…). Okay maybe if I'm I'm doing a meta-analysis, I'm also then using secondary data, right? #00:03:13#

Q: Yes, of course. #00:03:14#

R: So I'm trying to remember the last time I actually generate primary data, and that would be a very, very long time ago. So basically a 100 percent. Now if I'm thinking more broadly – colleagues that I'm helping, students that I'm helping, supervise, they generate primary data of course. That's much more difficult now for me to say out of my whole research how much of that is – yeah. It may be more like fifty-fifty, I would say. But but if it's really just about me, then it's basically a 100% of time that I use secondary data. #00:04:06#

Q: And for your collaborators or students…it would be fifty-fifty? #00:04:10#

R: Then (..) To put it all together. Yeah I would say it would be fifty-fifty. #00:04:14#

Q: Okay, good. That's really good. Even the fifty-fifty is really good because other researchers I've interviewed talked about percent of just about like 10 percent or five percent. (laughs) #00:04:31#

R: Okay (laughs). #00:04:33#

Q: So, okay. Then I would like to know for which purposes you used secondary data. You already said for meta-analysis, illustrations and re-analysis. #00:04:46#

R: Yes. #00:04:47#

Q: Is that all or are there other purposes? #00:04:40#

R: So, yes (…). Primarily I would say meta-analysis, so either I am actually involved in a meta-analysis pretty much directly, or I'm getting meta-analytic data to illustrate methods, to compare what authors have done with a different approach to analyze the data, and then another source of secondary data that I've been using quite a bit is data that have been

collected by my colleagues in the past. They have put together a very big dataset using a certain data collection technique which is called experience sampling or ecological momentary assessment. Yeah, it's a very big dataset that has been constructed based on all of these studies conducted in the past. Now I'm using also these data for re-analysis. #00:05:45#

Q: Ah, okay! So you created something like a big data file? #00:05:50#

R: It was not created by me but also by my colleagues in the past and it's a huge data file *[deleted for data protection reasons]*. The studies are slightly different but it's, it's about ten studies that have used a similar design and a similar questionnaire. And pretty merged into one (…) dataset and we have using it for some re-analysis. #00:06:23#

Q: Ah, okay, nice. It sounds really interesting. #00:06:26#

R: Yeah, actually it is. #00:06:28#

Q: Okay, and which metadata would you need to optimize that work? #00:06:37#

R: Okay. So, again, I would have to make a distinction here between meta-analysis and sort of these secondary-primary data that I'm reusing. So for meta-analysis, so what kind of information do I need there? What I need to know exactly what the ins.. what does the dependent variable if it's a mean or if it's a correlation, so what what were the two variables? #00:07:04#

Q: So the design? #00:07:06#

R: The design, the instruments. One big issue especially in this sort of more complex analysis that I'm particularly interested in, where you may be extracting multiple effect sizes or multiple correlations or multiple means based on (which) you compute effect sizes, is the correlation among variables. That is something that I often not reported and that really hampers the ability of people like me or others who want to do a more – this type of multivariate analysis, to do that properly. So inter-correlation, so the association between variables. #00:07:57#

Q: And would it be sufficient for this purpose just to provide a script with which you can generate the inter-correlation-matrix? #00:08:08#

R: It basically, while, so I'm assuming in a meta-analysis, you have the articles that you are extracting information from, so you don't have the primary data. So, authors report this information or they don't, you can of course approach them and ask them for this but ideally that information is quietly provided in the supplements or somewhere. #00:08:35#

Q: Okay, but we try to create a standard for data. Of course it's also important to know what has to be written in the paper but for us then to make your life easier (laughs), it would perhaps be both, so including the inter-correlation matrix in the paper and additionally perhaps providing the scripts, so that you can just check whether the matrix is also correct or something like that. #00:09:12#

R: Yeah, so if the, yeah. If the goal is the sort of – communicate to authors how they should provide their data essentially, then. If they providing their data directly of course, then I can easily compute any inter-correlation that I would like, of course if authors have already done this before me, like, or have created a script. That's nice. That makes my life easier. But again, if it's really sort of more this this (summary level) information that is typically used in a meta-analysis, then I would want some kind of information about the inter-correlation. So in the data in a meta-analysis at least, there are of course also these primary data meta-analysis but that's not what I'm talking about here. To me, the data of a meta-analysis is the effect sizes or the information that you need to compute the effect sizes, maybe some study-level characteristics and then this p-studies usually missing which is about the inter-correlations of the variables. Yeah, so if I'm really working with primary data, then I'm re-analyzing, then I can generate correlations myself, then I just need to know really what do these variables represent, what kind of instrument, what is their scale, what units were there measured, are they part of a kind of standard instrument, that would be nice to know, mh, what else? --- Yeah, so I think these are primary things that I would like to know. So what is. So if it's a questionnaire, what was the actual phrasing of the question, so instead of saying, you know, Item 1, Item 2, what is the actual phrasing of that question. #00:11:32#

Q: Yeah, and that would be something, for instance, that could be written in a codebook? #00:11:37#

R: Yes, exactly. A codebook would be great. #00:11:41#

Q: And are there other methods, other purposes for reusing data that you know about but haven't used in your past, and would require other metadata, so other than the ones that you have already mentioned? #00:11:58#

R: Mhh ------ (laughs) ---- So, that's difficult for me to say because (…) of course, once the data are there, then I can do all kinds of analysis with that beyond having the data and the codebook, what else can you do what isn't already covered by that? #00:12:36#

Q: Perhaps you could provide your hypothesis or author information so this (…) this bibliographical stuff. #00:12:44#

R: Yeah, okay. I guess, in principal, it could be interesting to ask questions like who collects what information, so then you may want to know something about the people who have actually collected the data. So what is their background, what were their questions irrespective of what I want to do with the data. So about the original, more information about the original study that generated these data. What was the purpose of that study, the context. #00:13:24#

Q: Okay, yeah, good. Any other ideas? #00:13:34#

R: Mhh, no. Can't think. #00:13.39#

Q: I don't want to force you. (laughs) #00:13:40#

R: Yeah, yeah, yeah, I. #00:13:42#

Q: Good. Then what kind of data are you generally using for the different purposes, so rather physiological, behavioral, video, what kind of data are you using? #00:13:58#

R: I would say the large majority is questionnaire data that I use. So that definitely applies to that when I'm working with secondary raw data, then this is usually related to SEM and so that's questionnaire data. Now to that you can add of course other types of (sensors), so

people sometimes use more passive (sensors) to that, so then you have, I guess, what you could call more physiological measurements. So accelerometer (…), heartrate, – #00:14:43#

Q: EMG also? #00:14:49#

R: I'm sorry? #00:14:49#

Q: EMG also, or? #00:14:50#

R: I've also, so if I'm thinking more broadly what I've been with what I help my colleagues: EEG, genetic data, imaging data – #00:15:07#

Q: So also psychobiological data? #00:15:11#

R: Yes, for sure, yes, yes, yeah. Then the more medical side of things when it's more meta-analysis, then it can also include a sort of – just to give an example, I've been involved in quite a bit of research related to smoking cessation and you of course ask people whether they've quit but then theirs is usually some kind of conformation of that by accelation of CO2 or. So I guess that's more physiological again. #00:15:50#

Q: Yeah, I think also it is more physiological. Okay, and how would you or how do you perceive the quality of these data or how well are they documented? #00:16:06#

R: Mhh (thinks). #00:16:11#

Q: Are there differences between the different type of data you are reusing, so? #00:16:17#

R. Yes. – Mh (thinks). That's a good question. That's hard to say in general, it really depends on the particular study where these data are coming from. So maybe one thing that I see quite a bit and where again that … makes meta-analysis difficult is what is often not documented clearly in these studies that you want to include in your meta-analysis the type of drop-out that occurred. .. So you have a sort of flow chart of patients, but but still, once you really dig in, you realize that the degrees of freedom can't that, they don't really match the sample size that they claim they have analyzed. So then depending on certain outcomes, actually the

sample size was different because while maybe on this outcome some data were not usable, so that's one thing that is often not documented very clearly and that applies I would say to all of these types of data. Is that worse for particular things? I don't know. That's too hard for me to say. I don't know. #00:17:54#

Q: So it's the transition from raw data to aggregated data that is not very well documented? #00:18:00#

R: Yes. #00:18:01#

Q: Okay, so in this case perhaps once more a good analysis script would help or a data preparation script? #00:18:12#

R: Mhh (agrees). Yeah, yes.

Q: Okay, good. And do you know any reasons for these differences in data quality? #00:18:31#

R: So when I see these differences, mhh – yeah, what are the reasons for that people, they are just sloppy or they are just. It's strained of course by how much information they can put into their paper, so they (...) these details. – It's maybe. It might be related to a sort of people's understanding how important it can be to have that information available. It's probably related to people's background – their statistical training that usually leads to a sort of heightened sensitivity to these issues but that's just a guess. Of course I don't know always what kind of training people have but I would assume that's often related to that. #00:19:38#

Q: And would you presume that implementing a standard – so what we try to do – would help to overcome these problems? #00:19:50#

R: Yes. It certainly make people more aware of some of these problems. And, yeah, for sure (…). #00:20:03#

Q: Okay. Then let's come to secondary data use from your perspective as a data provider. What sort of metadata do you are generally provide about a dataset when you upload? #00:20:20#

R: Mhh, so then I would say the. So this really now applies to meta-analysis when I'm really primarily involved into a meta-analysis because again: Raw data I don't really generate. So what do we provide when we actually provide the dataset: All of the effect sizes, the codings of the moderators, so the study level characteristics that we're interested in #00:21:00#

Q: (…) #00:21:02#:

R: We don't usually provide a codebook with that because that basically should be in the article. I must admit that (...). You would basically have this dataset and then based on the article that writes up the meta-analysis, it should be clear what these variables are, but maybe it's not always so clear. So we don't usually provide a codebook now that I think back in the past how we've done it. – So the actual data used for the meta-analysis that's what is usually (...) not always but usually provided. #00:21:52#

Q: Okay. But the data are not documented? #00:21:58#

R: Yeah, not properly, I would say, yes, yeah. #00:22.02#

Q: Okay, and yeah, you already mentioned something like this but would you say that your metadata are sufficient for other researchers to reuse your data? #00:22:14#

R: Probably not. No, no, so I think we should do better on this that it is much clearer what these data represent. So I think that that could be improved even much. Yeah, yeah, yeah, so (thinks). #00:22:37#

Q: And how you would improve it exactly? #00:22:40#

R: So, yeah, good question. So I think a codebook would be a good first step to really have a proper codebook. – You could also, because the the the data in the end come from a sort of coding scheme that you applied to to the individual studies, you could also provide these

coding sheets that you used when you distracted the information, that you used for distracting the information. So at least providing the coding sheet itself would be useful or even how you have filled it out for every study. That would, I think, also be quite useful. Then in in some cases you you have to do additional steps to actually compute the effect sizes, so so the link between the information you extract from an article and how that leads to the effect sizes, sometimes you can really automate the whole process, so you really have a dataset where you have all the means, and sample sizes et cetera et cetera, but – and that's the idea – but if you, if that's not possible or if there has been another sort of processing to compute the effect sizes using certain converting equations, then that needs to be written down at least somewhere and then that making that also part of the data would also be important, I would say. #00:24:25#

Q: Okay. Good. Have you ever used certain metadata standards for annotating your data, so something like DDI or Dublin Core, Darwin Core? #00:24:44#

R: No. #00:24:46#

Q: No. Do you know about these standards? #00:24:49#

R: No. #00:24:50#

Q: Okay, good. #00:24:52#

R: That's an easy answer, not a good one, but okay. #00:24:56#

Q: Yeah, that's good to know because if we want to implement a new standard, we also have to know about the status quo of researchers regarding standards. (…). Okay then, last question: If you would have the task to create such a standard what do you think is the most important information that should be included in it? Perhaps you can think about it in terms of the JARS standards from the APA, so the Journal Article Reporting Standards, because I think you generally use it, so you know about the different aspects included in it, and if you would transfer it to a dataset what would be the most important information? #00:25:45#

R: Ouh. – Well, the most important information would be the meaning of each variable, so that, yeah, a codebook because sometimes again a dataset, where it literally just says like:

Item 1, Item 2, Item 3, and that's essentially useless at that point. So a codebook is the most important thing probably from my perspective when I'm working with raw data. #00:26:20#

Q: Yeah. Okay, good. Then I thank you very much for your time (R: you're welcome), and for these good ideas and impressions. #00:26:34#

R: Are you doing actually all of the interviews in English, or? #00:26:38#

Q: Not all, two are in German. #00:26:40#

R: Okay. Wir hätten das natürlich auch auf Deutsch machen können, aber ich weiß nicht, ob das (...). Wir haben jetzt aber auch die ganze Zeit auf Englisch kommuniziert, aber gut. Ja, okay. #00:26:53#

Q: Naja, ist egal, gut zu wissen. (laughs). Nee, ich wusste nicht, dass du jetzt deutsch sprichst. #00:27:00#

R: Gut. #00:27:02#

Q: Aber wir haben es ja hingekriegt. Okay, dann ja vielen Dank und dann wünsche ich dir noch einen schönen Tag. Vielleicht sieht man sich ja einmal. #00:27:12#

R: Ja, sicherlich. Ich hoffe, dass ich … ich hatte schon mit *[Person 1]* gesprochen, dass ich mal irgendwie für ne Woche oder so nach *[Stadt 1]* kommen möchte, um einfach mal in Ruhe arbeiten zu können, und ja, da war er sehr aufgeschlossen. Die Einladung steht. #00:27:34#

Q: Du bist jetzt noch in *[Stadt 2]*? #00:27:36#

R: Ja, genau. #00:27:37#

Q: Und da kann man nicht in Ruhe arbeiten? #00:27:39#

R: Ja, ich meine, ich übertreibe jetzt. Aber es ist halt auch so, wenn man woanders ist, dann wird man auch eher in Ruhe gelassen, so nach dem Motto. #00:27:51#

Q: Ja, das stimmt. Okay. Gut, dann vielen Dank noch mal. #00:27:57#

R: Jo, tschüss. #00:27:58#