R: Do you want to ask me the questions or should I just go through? (laughs) #00:00:04-0#

Q: No, I...I think I'm going to ask you the questions and, yeah. #00:00:10-4#

R: Okay. #00:00:09-3#

Q: Then we just talk about the facts. Okay. Yeah. As you can see, first of all, we would like to elaborate on secondary data use from your perspective as a data user. And first of all, I would like to know: How often did you reuse datasets from your lab and other labs in the past? And please quantify your specification by providing the relative frequency for reusing data compared to producing primary data. #00:00:40-2#

R: So, mmh (thinking). Yeah, for (...) data from my own lab, that's actually hard to estimate. Mm, I can recall, I think, maybe two or three cases. #00:01:00-0#

Q: Okay. #00:01:01-9#

R: Where I, I basically, like, ran a study, collect the data or...and, let's say, I don't know, something didn't work out, data ended up in a file drawer, and then, sometime later, (...) oh, that's...I can test, like, something different with that or take another look. So that now I don't know, (...) two or three times. Once, there was, like, from data that I collected for a different study, I realized that...that the data that we had...that we could actually test a novel hypothesis that was, like, completely unrelated. #00:01:32-2#

Q: Mhm (agreeing). #00:01:31-1#

R: That was actually the only time that the data (ended up?) in a...in a paper and that we ran, like, (...) studies. But just, like, compared to the...mmmh...like, total number of studies that I've done in my life, entire time, like, percentage-wise, it's maybe, like, for all the...the studies that I've done, that's probably not even the one percent. #00:01:56-3#

Q: Okay. #00:01:55-0#

R: I don't know. Like, what...what kind of, like, (...) certain percentages or, like, I mean, I don't know, like, let's say, hypothetically, it could happen, like, three or...three times. #00:02:09-9#

Q: A year or...? #00:02:09-6#

R: Like, like, for, for, for me or for myself. #00:02:11-1#

Q: So for your whole career three times or...? #00:02:16-2#

R: Yeah. Yeah. #00:02:19-8#

Q: Okay. #00:02:18-3#

R: And I...I'm sure I ran more than 300 studies over my career, so that would kind of, like, smaller than one percent. Using data from other labs, it was actually just recently, it was the first time. And, like, my knowledge is actually (...) limited because it was mostly my...my grad students have done all of these analyses. I mean, I know a little bit about their experiences, so I can tell you a little about that. But that was basically the first time, and this was basically three existing datasets from the same research team. #00:02:52-0#

Q: Okay. #00:02:53-7#

R: Erm, where we basically felt that the way they analyzed the data actually does not provide, like, a sufficiently nuances view of, like, with advanced statistical procedure, you could actually get more information out of the data. And, mmh (thinking), so that, yeah, basically, like, once for one project that was all related to the same question and we used (?) pre-existing datasets. #00:03:21-0#

Q: Mhm (agreeing). Okay. Yeah. That's it or...? #00:03:26-7#

R: Yeah, it's like, it's difficult to quantify, it's like, basically, that is like (laughs). #00:03:28-8#

Q: Yeah, okay. (...) #00:03:27-0#

R: It's easier with, like, number of studies of my lab compared to the number of re-analyses. Yeah. #00:03:38-0#

Q: Yeah, yeah but it's okay (laughs). Good. So then the main research purposes for which you used secondary data were mainly applying new methods to existing data? #00:03:55-0#

R: Yeah, mmh (thinking), like testing, testing new hypotheses with existing datasets that can provide answers or, like, have the relevant data to test these hypotheses. That would be one. The other is analyzing existing data in a...in a new way to get more nuanced information. I think that would make another, like, an abstract wave, I mean (?). #00:04:22-2#

Q: And for these methods, which metadata would be most relevant for you to...to optimize that work? #00:04:31-3#

R: Yeah, so one issue that they ran into in this, like, I don't know, this is not like, I...I don't know, a review and sometimes I look at other (...). The biggest difficulty is that sometimes the...the...there's insufficient variable information. It's, like, completely unclear. Basically, people have their idiosynchratic way of labelling things, it's completely unclear (...) what, like, acronyms are referred to. So, like, imprecise labelling. The second is...is, I think that has become more common but, like, variables from raw data have been created, so there is, like, often insufficient record of, like, data code (...) There's, like, analysis code but then people (...) the full dataset but it's unclear...like, the...the...the dataset basically includes the raw data plus the aggregated data. But the...the step from the raw data to the aggregated data is not well documented. #00:05:46-9#

Q: Yeah. Okay. #00:05:45-0#

R: And, so, yeah, in some sense, it's kind of, like, the aggregation code is missing. Mmh, yeah, that's... #00:06:02-5#

Q: So you would need a codebook and a... #00:06:06-1#

R: Yeah, yeah, like, I mean, I don't know, it's...I'm jumping a little bit ahead, like, what I'm doing in my...my lab as a...as a default is actually...I only provide raw data, and the code that I provide on...on OSF is basically, includes everything including the aggregation, the code for the aggregation, and then the analysis with the...that are relevant for the...the reported results. So it's basically, every step is basically transparent in the code. #00:06:36-9#

Q: Also the variable labels for...for the different variables? #00:06:39-2#

R: Exactly, it's like basically what refers to what, and, like, what variable is (...). And ironically, that actually...that very specific documentation and, I don't know... (thinking) Yeah, I don't know. You're not really asking...it is, like... *[person 1]* recently actually identified an error in our aggregation code of our IT database, so exactly that...basically, the way we aggregated the data, they included actually, like, one typo that basically led to, like, an omission of, like, (...). And then, there's a... basically, like, one, like, syntactical error that basically presumed, I don't know, correct responses on all trials, which was not the case, and things like... And so, like, basically, based on this response (...) everything was fine, but I think that's exactly the kind of transparency that I think is very hard in the field. In our case, it was pretty minor in the end. Are you still there or...? #00:07:55-6#

Q: Yeah, yeah. I'm still there. #00:07:56-7#

R: Yeah, yeah, okay, good. The screen went dark (laughs). #00:08:00-7#

Q: (...) for the record (laughs). #00:08:02-3#

R: Yeah, and...so that basically (...) to correct those errors, do the re-analyses, but I think that's exactly the type of transparency (...). You could go back and see what we've done and then... And I think actually the reason for that is because he re-used raw data from our lab for an entirely different project. But in...in doing so, he realized that he actually could not reproduce our own. The findings that we reported and then the...we realized that the source of the discrepancy is actually the syntax in the aggregation. And...so...I think that's exactly the type of documentation that we need. #00:08:40-3#

Q: Yeah, true. Okay. Are there any other methods you know of but you have not used them on your own in the past and do you think you would need other metadata for...for these methods? So, for instance, meta-analysis or (...). #00:09:01-1#

R: Mmh, very vaguely, I know that, like, *[person 2]* here in our department and, (...) have been, like, have been involved in, like, creating, like, like, a massive dataset of *[discipline 1]* data. I mean, it's technically not meta-analysis but it's like, it's (...) basically where people provide the raw data and they...they try to identify larger...large-scale patterns because particularly in *[discipline 1]*, it's, like, everyone running these studies where there are super small samples. But...yeah, I mean, that comes to mind but other than that... #00:09:38-9#

Q: Okay. Good. #00:09:42-7#

R: Yeah, and for that, obviously, you need, like, also proper documentation, yeah. #00:09:46-9#

Q: Yeah, and proper documentation in this case also means code for you or is it more than...? #00:09:55-0#

R: I'm...I'm not versed enough with *[discipline 1]* data but I...one thing that I do know is that for...for *[discipline 1]*, like, it can even make a difference which software that you use. And, like, you have basically exactly the same code but different software packages produce different results and...and surprisingly that, like, I mean, I am not a *[discipline 1]* scientist but it seems like even people in the *[discipline 1]* community, some of them are not aware that...that, like, the same code applied...different software packages can produce different results and then claim fraud or... (laughs) #00:10:36-8#

Q: Yeah (laughs). Yeah, it's everywhere, I have to say. Okay. And, last question from this first block. What kind of data are you generally using for the different purposes? #00:10:58-3#

R: Yeah, my love is mostly behavioral data. I mean, I started doing some psychophysiology stuff but that's, like, more the exception. #00:11:01-6#

Q: Okay, mhm. #00:11:03-7#

R: Yeah, like, endocrine data in one paper. #00:11:09-2#

Q: Mm. #00:11:11-1#

R: Erm, I haven't done anything with video lately. I, like, started some psychophysiology, EGM, facial...facial EMG, GSR. I would say 99% behavioral. Like, more than 99% behavioral. #00:11:29-4#

Q: Okay, and for these behavioral data, are there any differences in the documentation quality? #00:11:32-7#

R: Erm, I'm not sure (reflecting on the question). Like, you mean, like, between behavioral and physiological or...? #00:11:39-8#

Q: Mm, yeah, as you did not use so much physiological data but mainly behavioral data, are there also differences for different behavioral data? So, for instance, for mainly RT data or something like that, for response data and questionnaire data, for instance. #00:12:00-3#

R: Yeah. Yeah, for...it's...it's hard to say. I mean, like, I...cause I say, I document the raw data with the code. I, like, at least for what I post, I wouldn't see a difference. Yeah, and since I have, like, limited experience looking at other people's raw data but I think it's...mmh...like, simple self-report measures, obviously, that's, like, the easiest. But, like, the more complex the data aggregation is, like, I don't know, with, like, response-time data from evaluative priming (? ...) etc. That's where things start to...to get messy, just with the complexity. And then...like, the amount of data that would have to be aggregated. #00:12:48-2#

Q: Mhm (agreeing). So this would also be the reason...or one reason for you why there are differences, complexity of the data... #00:12:52-0#

R: Yeah. Exactly, complexity of the data, yeah. #00:12:58-3#

Q: Mhm. #00:12:55-4#

R: And the complexity of the raw data that go into, like, the...the use for the aggregated variables. #00:13:04-4#

Q: Yeah. Okay. Good. So then we can switch to the perspective of a data provider. What sorts of metadata do you generally provide when you upload your datasets, for instance to OSF? #00:13:20-2#

R: Yeah. I mean, in some sense, that's, like (...). We typically provide only the raw data just to avoid confusion. I don't know, like, one, one thing...again, it's kind of related to your question but it's, like, a little bit, probably more information that's like...erm...one of my greatest concerns in, kind of, like, trying to make sense of other people's data is that they often, like, upload too much. It's like, it's all for the sake of transparency, but in the end, that transparency is undermined because people think that they need to upload everything and then some, like, an external person is basically pretty much impossible to find your way through. Like, even with...even with appropriate metadata. But it's just, like, the level of complexity is...is...is increased by uploading a ton of irrelevant stuff which basically undermines transparency instead of, like, providing, like, ensuring transparency. So, my strategy has been, like, just...just raw data files and in most cases, this is a single raw data file that includes all the data. Sometimes, if they come from, like, different programs, I don't know, if we used, like, [*software 1*] and [*software 2*] (...) raw data files. And then, mmh, I, like, syntax file that basically specifies each step of the data aggregation. In using the...the data file for the actual analyses so that if people would rerun that syntax, like, they could basically recreate our data file with which we worked. And...and also then all of the following analysis steps. #00:15:13-1#

Q: Mm, do you have any concrete examples for cases where you would say: Okay, this is just too much and no one can understand this dataset? Did you have such an experience in the past? #00:15:26-6#

R: It's usually not the complexity of the dataset. It's basically, people upload, like, like, I don't know, for one study 25 data files, and then just, like, finding the right one is like: Oh, this is the raw data file, and this...and this is when I did this and this is when I did that and...and...just finding your way through, like, which is the right data file, which one should I use...even if they provide metadata, it's just, like, it's very time and resource-intensive because people kind

of, like, want to err on the wrong side or want to err on the side of providing too much than too little. But that, in the end, undermines transparency in my view, so. #00:16:13-7#

Q: Okay. But there is then also no correct versioning of these datasets, for instance? So that you can see: Okay, here is version 1 of the dataset and then comes version n. #00:16:26-6#

R: I mean, usually...usually, that's...usually, that's the case but it's, by having, like, so many different datasets and then one for this and one for that analysis, it's just, like, very difficult to find your way through. And it's also...makes it very difficult to...to…how can I say, like, okay, how did you get from this data file...how did you generate the other data file etc. What I...what I personally prefer is just, like...just upload the raw data as you get them from your study in one file, label everything properly, what refers to what in the raw data file, and then provide your analysis code, erm, not analysis code! Your aggregation code that specifies, like, what...which raw data did you use in which way to create which variable. What does (...) variable, what is it supposed to mean and how has that new variable been analyzed. #00:17:22-9#

Q: Mhm, yeah. #00:17:25-8#

R: And, yeah, so in some sense, I'm all in favor of transparency but I think in that regard, often less is more because it increases transparency and, like, I think the problem that at least I have been facing with, like, looking at other people's datasets is, like, well, it's too much and, like, 80% to 90% stuff that is irrelevant but people just, like, I mean, there's this uncertainty and that's why I think, actually, I really appreciate what you guys are doing with that kind of, like, standards, like, for, like, what...which data should you provide, how should you provide them. And just, like, follow these guidelines and it's, like, easier for other people to find their way through instead of just, like, this uncertainty: Oh, what...what is expected to be...that I should upload? And what is...it's...but I think, particularly because it's...it's a relatively new development and it's because (...) particularly among younger scholars. They just want to be as transparent as possible and then, just like, everything that they have on their hard drive is on there, and it's just...that makes it just (...) #00:18:38-3#

Q: Yeah, there's...there is no standard at the moment, that's the problem. Yeah. We also see that, so that is why we have the project (laughs). #00:18:48-0#

R: Yeah. #00:18:49-2#

Q: Okay. Mm. #00:18:51-9#

R: And don't get me wrong, I don't think that, like, most people have bad intentions but in some sense, it's kind of like, if I...if I wanted, like, if I had bad intentions and I wanted to conceal something, that would be a very easy way by, like, appearing transparent but in the end, no one would be able to, like, identify what went wrong. #00:19:13-3#

Q: Yeah. And the question is also: Do I really need all this information for my work, right? Because if I'm doing some kind of secondary analysis, do I really need all this stuff which mainly comes from, yeah, bibliographic origins? Yeah, so, yeah, okay. Next one. Do you think that the metadata that you provide are sufficient for reusing your datasets? Do you have any feedback from other scientists? #00:19:49-4#

R: I mean, I think one thing that...that I could do better, and I've seen this now, like, mmh, working with my...my grad students and some junior researchers in/and other groups. I think something that I should probably, like, adopt, is, like, I have done this for (...) general readme file, it's just, like, okay, here, this, I don't know...I've actually done it for some set. Like, this refers to that, this refers to that, this refers to that, that is basically, like...what I'm trying to do is usually provide some guidance in the structure of how I upload things. But I think, like, a specification of, like, okay, here is this, here is this, here is this, and here are the basic steps...kind of, like...like, basically, when you go to OSF, it's kind of, like, first document that you open and...shouldn't...I think, in an ideal case, it shouldn't be longer than, like, half a page...when you read that half of a page, then you basically know what is what and where you can find things and... #00:20:57-0#

Q: Okay. #00:20:58-1#

R: Hope that it makes sense. #00:20:59-6#

Q: Yeah. Makes sense. Good. Have you ever used any metadata standards for annotating your data? #00:21:14-8#

R: Not really. I mean, to be honest, I...I'm not aware that there are, like, metadata standards. #00:21:19-2#

Q: Yeah, there's actually...at the moment, if we talk about discipline-specific standards, there's only one for fMRi data, this is the BIDS standard. But for any other sub-disciplines in Psychology, there is basically no standard (...) other data (...) behavioral data. #00:21:47-9#

R: Yeah. #00:21:49-0#

Q: Yeah. Okay. Then the next question, we can skip this one. And, yeah. 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? So perhaps you can think about this question in...in...yeah, in the context of the JARS? You are aware of the JARS, right? #00:22:18-5# #00:22:18-0#

R: What, sorry? #00:22:17-1#

Q: You know the JARS, right? So the Journal Article Reporting Standards from the APA? Yeah. #00:22:24-5#

R: Yeah, yeah, yeah. #00:22:20-4#

Q: You know perhaps not the name but... (laughs). #00:22:25-4#

R: Yeah. #00:22:27-2#

Q: Yeah, because I've talked with other researchers and they just asked: What is the JARS? #00:22:32-5#

R: Yeah, I think it's...I'm not sure that I, yeah...I think in some sense I would probably repeat myself, like, it's like, raw data instead of aggregated data, like, only raw data. And instead of aggregated data files, just aggregate code with...with proper specification of...of each...of each step in the...in the coding. #00:23:07-1#

Q: Mhm. #00:23:04-6#

R: Like, aggregation, and then…then the analysis, it's kind of, like, erm...and, I don't know, like (...) this is...all of this stuff, I actually learned many years ago from [*person 3*] as a grad student (laughs). (...) ahead of time back then. (...) like, it was not actually for transparency, like, for...for others but it's mostly kind of, like, that you understand what...what you have done. Like, when you go back to your old data file a year later, it's basic...and that's...that's how...what I've been doing is...like, is, yeah, document each data step, so it's like, have a brief description before each step or, like, I don't know, like, a series of steps, what...what is this code doing and what...what is basically the relevant output here, like, e.g., like, an aggregated variable. Then for analyses, it's like, okay, like, what...what are these analyses doing (...) post-hoc analyses, what are doing the regressions etc. and... So I'm thinking more along...along these lines. It would probably be a bit different than Journal Article Reporting Standards but I think it's...it's probably getting in...in that direction. I...I mean, it might be my personal preference but I do not want to see any, like, gazillions of...of aggregate...different aggregate data files. I...I would prefer to see raw data and the code, so that when I...when I download the data file and run the code, basically I get what is reported in the paper. And then I can include my modifications, aggregate the data differently as I want, aggregate them in the same way but analyze them differently etc. #00:24:59-2#

Q: Mhm. And would you say that it makes sense just to upload, for instance, also the hypothesis with the dataset? So, for instance, in the readme file you mentioned before? #00:25:12-8#

R: Erm (thinking about the question). I think those are different things. I know that there's been some debate between, like, Brian Nosek, OSF, Centre (…) in Science and then, like, Alison Ledgerwood, on the other hand, has been very outspoken and kind of, like (...) that, like, hypothesis derivation is...should be separated from data analysis and data treatment. These are, like, different...different things, so I think it's...it's okay to provide reporting standards for those as well, but I would consider those as separate issues. Like, for example, one of the biggest concerns that I have that I think is, like, only some people have pointed out (...) I mean, yes, preregistration is...is great...is a great tool to, like, create transparency and basically gonna, like, like, like, providing some prior commitment to your data analysis procedure. But, like, one issue is...is...is that...in some sense that, yeah, so you preregister your hypotheses but...and that's an entirely different thing. #00:26:29-8#

Q: Yeah. #00:26:27-1#

R: It's like, yes, you can...you can have...but it's like, and I think there's also confusion in the field about, like, hypotheses and theories, and...like, just like, or like, there are theoretically-derived hypotheses, and just like, personalized predictions: I predict that. It's like...that's great...but it's technically not a theoretically derived prediction, it's, like, it's a personalized prediction. And (...) I...I don't care what you personally believe. What I care is...is, like, which set of theoretical assumptions have which logical implications? And so, erm...and that's, I think it's...it's...it's a bit of a gray area because technically, right now, all of this stuff is in the preregistration but I consider that part of, like, your, statistical way of, like, analyzing data, how you...the hypotheses are a bit of a...a different can of worms. And because there are no constraints, and then people...like, it's so common, it's like, it's like: Oh, (...) I predicted that, and it's like...it's...it's basically just, like, a set of ad-hoc assumptions, and these ad-hoc assumptions then, depending on the outcome, can be changed. But that's, I think, where things become flexible and where people feel like we need some preregistration, and...I know that, like, Liz Page-Gould, e.g., is like: Oh, we should have preregistration for, like, the generation of hypotheses, it's like, yes, it's great, but in some sense, it's like, if...if you derive a certain prediction from a set of theoretical assumptions but the derivation is logically incorrect, it doesn't matter if (laughs) you derived it, and that was your prediction, what matters is the logical relation of your theoretical assumptions. And if someone else comes and is like: Ey, you made a logical error in this...this set of assumptions implies this, not that. And that issue increases with the complexity of the theoretical assumptions. And, mmh...so anyways, that's what I'm...where I see, like, on the one hand, I see why people like Nosek think it's like, it should be combined because (...) want to reduce, like, flexibility in ad-hoc assumptions. But on the other hand, I'm...I'm also on the same page as Alison Ledgerwood, it's like, these are different things, one is basically the...the analysis and the statistic treatment of your data, which is a different issue than the derivation of your hypotheses. (...) Like, there has been, like, very little discussion of the hypothesis and generational aspect yet. I mean, some people have brought it up, like, Liz Page-Gould and with some suggestions that I think is...erm...don't really get at the problem. To be honest, I don't have an answer to that...that problem, like, other than better...better training and logical reasoning. (laughs) #00:29:25-7#

Q: Okay. Yeah, that would be good. (laughs) #00:29:28-4#

R: And...and in theory development. (laughs) #00:29:30-5#

Q: Yeah, yeah. That's also a great debate, yeah. Okay. Yeah, then I thank you for your time and... #00:29:38-7#

R: Sorry, (...) #00:29:42-9#

Q: I'm sorry, I didn't get it. #00:29:43-9#

R: You're welcome. Yeah, I'm back, yeah. #00:29:44-3#

Q: Okay (laughs), wonderful. Good. Yeah, then I wish you all the best for the new year. #00:29:52-2#

R: Thank you, you too. #00:29:53-2#

Q: Thank you. #00:29:52-0#

R: Good luck for your project! #00:29:56-5#

Q: Yeah, thanks. I hope we come to a standard. #00:30:01-6#

R: Yeah. Alright. #00:30:03-2#

Q: Then see you! Bye! #00:30:03-1#

R: Yeah. See you, bye! #00:30:07-7#