DC09\_transcript\_deidentified

**SUMMARY KEYWORDS**

data, people, published, transcripts, thinking, informed consent, context, project, irb, qualitative data, qualitative, research, big, questions, challenges, sharing, researchers, terms, depositor, identified

**SPEAKERS**

Sara Mannheimer, DC09

**Sara Mannheimer** 00:01

Um, so I'll do a quick overview of my research. I am, I've identified a similarity between qualitative data reuse and social media data use, thinking of both types of data as like qualitative data that were created for one reason, and then are being used for another reason. And so I I've also seen that both of the, these types of data have overlapping issues and challenges. And so the sort of idea behind this project is that by connecting the communities of practice, for both research and data curation for qualitative data and big social data I've been calling it that we can improve data curation practices, and improve responsible practice for both and maybe even potentially, like help qualitative researchers begin to scale up some of their research with qualitative data reefs. So, um, yeah, so the main question is, how can data curators best handle qualitative and big social data to support ethical, epistemological and legal data sharing practices, or epistemologically sound? So the interview is structured around these six issues that I've identified, which are context, data quality data comparability, informed consent, privacy and intellectual property? So we'll have six questions that addressed each of those issues, then I have a, like, warm up question and wrap up questions. So it should take about an hour, up to 75 minutes. Any questions before we start? Okay. All right. And so when I asked you to, or first, can you tell me briefly, I know a little bit about this, but the types of data you usually curate, and then what your interests are regarding data curation?

**DC09** 02:18

Yeah, [I focus on] qualitative data. And so it's, while it comes in different forms, I would say the vast majority of what I handle are interview transcripts, slash focus group transcripts, with focus groups being a much smaller portion, but because they're all kinds of transcripts. I think of them as the same approaches that we apply in terms of curation. So transcripts, basically, primarily interviews. In terms of social media. Sorry, you call it big... What do you call it?

**Sara Mannheimer** 02:49

"Big social data." You can call it social media data too.

**DC09** 02:54

We have only had a handful of inquiries, and I believe I mean, I tried to look quickly, I think we only have one such project published. And I don't I mean, probably for various reasons that we may not yet fully realize. But we have been thinking about why that is. So I can tell you more later why you know why I think these data don't even get the or the researchers who do these kinds of projects don't even get to us. But the vast majority of what [I curate] is just straightforward, qualitative data. And really, primarily the transcript kind. We have done some videos, some audio, a lot of archival materials, which I know is not what you are interested in. But that's another type of qualitative data that people deal with a lot. And some have started thinking about, you know, making, making archive based research more reproducible, and so coming to us with that. So that's, that's really how I would describe the vast majority of what I handle. Yeah.

**Sara Mannheimer** 03:48

Cool. Will you tell me more, we may...

**DC09** 03:51

Oh sorry, you asked about my personal kind of interest or focus or something like that. That's interesting. I mean, again, we just as a repository, we serve people as they come to us. But my personal interest is in trying to figure out how to get the conversation started with researchers early enough in their research process, so that this aspect of the sharing is not something that is an afterthought. It's not something that's imposed on them, imposed on them after the research project is over. And then it creates—again, as you I'm sure know—creates so many more difficulties. And even if it becomes possible to share the data, it's always more time consuming, more expensive, just more, you know, labor intensive, if you will, to do it. Right, if you if they haven't been thinking about it from the beginning. So my main interest right now and really, it's been almost since the beginning of my work at the repository is how do we insert ourselves in these conversations early enough? How do we like enter these pipelines where, you know, by the end of it, things are flowing smoothly and not... that sharing is not presenting itself as another hurdle for researchers. That's kind of my main main interest. Yeah.

**Sara Mannheimer** 05:00

That is very interesting. Um, do you write about that? Because though I'd like to know more, as you know more.

**DC09** 05:10

Yeah, I mean, absolutely, especially since we're focused on how to make human participant data, you know, that particular type, regardless of what form it comes in, and as I said, kind of, technically, or technologically, we're dealing most with transcripts. But in the past, that doesn't matter, even a transcript, they're still problematic, because as you said, it's difficult to deidentify qualitative data completely. And if you do manage to deidentify it a lot—it's still not completely—but a lot... how much use do you leave in those, you know, how much value do you leave for other people. So it's a very tricky balance. So it's something that we discovered very early on, we have to just do more work on and we have, yeah, we have a whole research project on working with transcript research data with... try to reach out to IRB, for example, to get them to because they actually are an institution that meets the researcher very early in the in their research process, if they're going to be talking, you know, if they're going to be dealing with people. So I always meet the researchers very early on, and if they're aware of the data sharing possibilities, slash requirements, and they make sure that the consent process—consent form—addresses it again, this is what we want. If the IRB is promote that or if they don't discourage it, you know, if they do the things that align with the path to data sharing, things become much easier. So that's kind of one channel through which we're trying to promote this conversation to an earlier point in the research process. That's, that's something we've been working on.

**Sara Mannheimer** 06:38

Nice. Cool. Um, tell me more, just since we're on it now, let's just go through it. Why you think that big social data researchers don't publish in [the repository where you work]?

**DC09** 06:53

Well, I, again, this is completely speculation. I have not talked to a single big social science researcher. Well, yeah, even the person whose stuff we published, I am sure [they] would not describe [themself] that way. And that was kind of an incidental part of his of his research. So I have not talked to a single big social science researcher directly about this. But my suspicion is that [big social researchers]... First, they don't think about their data the same. They think, they think of it as kind of found data versus data that straightforward qualitative researcher creates in tandem with their with their participants. So I think that's one big thing. It's just they think about what they are collecting and analyzing in completely different terms, and then they probably, again, speculation, but they probably think of it as it's out there anyway, like, What do I have to do to make it more reproducible, or findable, or whatever, it's just there, if I found it, somebody else can find it. So that's, that's my kind of hunch about why they don't even they just the prospect of ditching I don't think occurs to them in the same way. Now, and again, I'm going to mention this, I don't know to what degree it's going to be useful for your, for your, you know, structured interview protocol. But we have decided, actually, again, as part of [a project investigating sensitive research data], [colleagues and I are hosting a workshop at a computational social science conference], which is basically, you've probably encountered the term, it's the it's the kind of people who do social science with the types of data that allow big, you know, big, large scale computations, which in many, it's not the only type of data they they work with, but in many cases, that means social, social media data, like big, you know, corporate, that they find. So we decided, and we, you know, as we're preparing [for the workshop], as we're preparing kind of the the workshop agenda, I want to talk to them about... we're discovering that it's going to be, I'm sure, it's gonna be a very different conversation from having... in comparison with the same conversation that we've had with social science disciplines where we kind of can imagine what the research process looks like. So we are mostly going to go there and kind of ask them questions and try to listen and find out. So after that workshop, I'll have maybe more to share with you about why they don't you know, whether my hunch is correct, why they don't share, you know, so that's on the more kind of substantive reasons why probably they don't think about the need for data sharing. I also think, to the degree that some of them might be presented with kind of a mandate, you know, you should think about data sharing, whether from a funder or a journal, I think they might then encounter the more technological impediments such as, you know, Terms of Use that make that actually impossible, like, you know, yeah, you know, people don't, don't often read the terms of use, but if they do, or if they're asked to do it, they'll discover that, depending on how they've collected the data, and in most cases, they've collected them in whatever is the easiest way technologically, they can't then do anything further with it. They're not allowed to. So I think that's kind of another hurdle for the very few who might have even been thinking about or might be considering the insurance so was Yeah. They tell us if they if they agree or if there's completely different reasons they share.

**Sara Mannheimer** 09:58

This is a great example of the two communities connecting. So I'll be here, I'll just keep an eye.

**DC09** 10:08

And there's, I mean, we discovered again, even just figuring out where to [host the workshop] was a big challenge for us. Because we just don't know that field. I mean, none of us are computational social scientists. So we tried to ask people, you know, ask around among friends and friendlies in the, in what we thought was that community. So several people converged on recommending that we go to [this conference]. And actually, we were supposed to do it last year. Our initial application was for an in-person one. And when they said what was going virtual will still be interested in doing it, we were not prepared. We were like, the whole thing is about, you know, very close discussion and people interacting. We don't know how to do this virtually. So we we kind of deferred last year. But given how things are going, we just said, Okay, we'll have to figure out how to do it virtually this year, because it is an important conversation to have.

**Sara Mannheimer** 11:04

Great. All right. Let's move on to the specific example. In my email, I asked you to think of a recent time when you either curated qualitative data for sharing. Or... Yeah, that's probably the one that you'll talk about. But do you did you come up with a specific, recent example. And, you know, we can talk about other examples as we go. But sometimes, I found that having one thing in mind can help us come up with more specifics as we go through the questions.

**DC09** 11:33

Yeah, I mean, I, I have several examples in mind, as you said, I think in terms of the social media data, we literally have one. So that's the one I'm going to be thinking about. The other one, I know kind of the most recent one, I don't know that that will be a very illustrative one for the kind of more substantive things you're asking about. But let's go... I'll first think about the most recent one. And then if it's just too... because it's a little bit of a, you know, not the best example, I get for data sharing, but a good example of the challenges of data sharing. It ended up being a very, very scaled-down version of otherwise very rich and super interesting transcripts, because they had not thought about, you know, getting proper informed consent. So we kind of had to work with them and suggest what we call, what do we even call it like abstracted illustrations of their analysis, like of their coding and of their... yeah.

**Sara Mannheimer** 12:24

Huh, okay.

**DC09** 12:25

So, anyway, I'll think about that. If it turns out not to be a good illustration for whatever you're asking about, then I'll think about a more maybe a more... a richer example that was published recently. And I have some of those.

**Sara Mannheimer** 12:35

Perfect. Okay. Um, some of the, I also am wondering about, like, whether these projects that you have in mind, had specific data management plans that required them to be published? Is that usually who comes to you?

**DC09** 12:52

No, no, that is very rare, and extremely rewarding when that happens. So much smoother, again, when people have thought about this, but now the, the, the typical project we work with, is not that still.

**Sara Mannheimer** 13:08

Okay. All right. Um, let's see. All right, so let's move straight to the first issue of context. So I have a little quote, to prep for both qualitative data and big social data.

**DC09** 13:24

Sorry to interrupt, I did read the interview guide. So if you want to just, you know, trust me that I read it. And if I don't remember something, I'll ask you exactly what you had there. But I remember all of that, because I read it.

**Sara Mannheimer** 13:35

Oh, perfect. Okay, cool. That's just sort of like give background because context is a little like, murky, the idea can be a little murky. So during your example, what challenges did you encounter, if any, when you were trying to capture the context in which the data was collected?

**DC09** 13:52

Yeah, for us context is very big for the reasons that you and this quote kind of lay out for qualitative material, it's really, it's key, it's, you know, the data really cannot be understood. You know, even just as support for any arguments that the original researchers are doing, let alone for future secondary reuse and all that they cannot be understood without at least some of the context. So for us, we take our role to be very big in that in that respect, that we guide people, guide depositors to provide that context in the form of various pieces of documentation. And so for example, the concrete data project that I'm thinking about that I told you, you know, ended up publishing not that much data ultimately, because they just couldn't. What we thought was the value added was, you know, encouraging them and kind of leading them with with prompts to give more context about how the data were collected and who the... you know, who the people were from whom the interviews were in focus groups were collected.

**Sara Mannheimer** 14:49

Can you give me just a touch more information about the study?

**DC09** 14:53

Yeah, I can I mean, I can. Yeah, why don't I just send you the even the link, but I'll read you the description. Yeah, sure. I'll give you some information. It's from public health, from the field of public health. And as I said, it's like a very, very interesting research program that the main PI conducts in general. And it's this particular project was called, it's a supplementary data set. Or let's call it a data collection for an article coming out on [health conditions in a non-U.S. country]. So here's if you just want to look it up later. That's the full information. So what we did with them was kind of guide, they basically had submitted no, what we call documentation, but no pieces of context, it was just like, here are the interviews, okay, we looked at them. And we said, No, we looked at the informed consent forms, but you cannot do anything with these interviews. Let's think what else we can do instead, because they, and they explained to us that their journal would not publish the paper if they didn't have some form of the data shared. And so we had them answer a bunch of questions basically, that that mean, kind of very simple, metadata questions, but put them in narrative form, so that they write a [narrative description] that explains kind of the process of collecting the data. And I think that's where a lot of the the context can be found. Again, for this project, but that's what we do, typically. And then the last step for these people was, in addition to explaining the context of reducing the data, right, reducing the rich transcripts that they actually were working with, to the minimal, you know, illustrative excerpts that we're actually publishing because of the informed consent constraints. And even that's, you know, that's a piece of documentation that normally, or sorry, a piece of context, you know, including the documentation that normally people wouldn't know about, like secondary users would not guess or not know why, for example, didn't you publish all the transcripts that you were telling me that you've collected? So what I mean, I think we just tried to guide people to provide to provide that context in the form of the [narrative description]. And then just again, even more simply, and it looks super decontextualized. But it actually provides context in a, in an implicit way, with the metadata elements, because I think they do, they do provide some context, not as richly as a narrative document. But still, they tell you, you know, the timeline of the data collection, and the—oh, in this case was very interesting. Actually, that's kind of unusual for us, but apparently not for public health. There were like, I don't know, more than 10 contributors to the data collection part. And I didn't, I didn't at first think to ask about that. Because they had, you know, they had two names and contributors. And that's often what we see, like one or two people who have worked on the data. But then I looked at the pending article, and it had, you know, more than 10 coauthors. So I said to the person I'm communicating with, you know, were their roles in any way related to the data? And [they say], Oh, yeah, I should, I should absolutely give them credit. So if you look at the list of contributors, it's huge. I mean, for a typical project, that's a very large list. So that provides some of the context as well about how actually, again, labor intensive it is to collect this type of this type of materials. So, you know that, does that answer enough about kind of what we do in the repository and to try to provide some of the context? Now it can't be it's not like it's not the methodology chapter of a book. No, it doesn't give that kind of, you know? What's the word that qualitative researchers often use? Positionality discussion and all of that, like, it doesn't go to that level, because I think that would become too burdensome. I think, to do that on a regular basis. In some projects, people choose to include that more, because they feel that it's important pieces of context. But it's not. It's not in all cases. And so it's it's kind of a balance of providing at least a level of context that's necessary to understand where the data are coming from how they emerged. And all of that.

**Sara Mannheimer** 19:08

And for this study was, what was the problem with the informed consent? They hadn't suggested that they would publish the data later?

**DC09** 19:17

We actually we did publish the informed consent, because they know often, we always review it, and sometimes, and we always offer to publish it, and these people actually agreed, so I forgot exactly. I think it explicitly said, No, it wasn't even that the informed consent was silent on the topic, which is more—not more often—but often is the challenge. In this case, I'm pretty sure it said something like the research team are the only ones who are going to see your material, I think. So it was very, very kind of, you know, unassailable in a way like you couldn't you just couldn't go around in any way. Now, the only thing you could offer in such cases is reconsent. And be you know, as an option for what the PI can pursue. And in this case, they could you know, it wasn't the again it took them [a few] years and [over 10] people to collect the data. So there was no way they could they could reach the exact same people and explain all that and reconsent. we've had very, maybe two cases of successful reconsent. And it's only for very specific populations, like, you know, in both cases, I'm pretty sure it was at least that, you know, people that you know where they are, you can easily reach them not, you know, the populations that were interviewed and collection focus groups here just would not be reachable for that kind of process. So yeah, so it was fairly, I don't remember exactly. And, again, we can look at the informed consent because it was published. But I think it was fairly stark in saying that only the research team will review the transcript.

**Sara Mannheimer** 20:41

And so you put this metadata, you create the [narrative description of the data]? Like, have you heard from users of qualitative data in the repository about challenges that they have encountered when trying to work with data that has been sort of decontextualized? A little, you know, in this way?

21:06

You mean, secondary users? Yeah, no, the answer is no. But that's, it's not even that we haven't heard of them talking about that topic. We just, that's one big, really lacuna that we're not sure we can, we're not even sure some of us are not even sure that we should, but we certainly are not sure that we can address in any way. So we don't, you know, in the interest of making it as easy as possible and as seamless as possible to access the data. Unless they are behind a specific restriction, which, you know, we have a lot of those projects, but not the majority of them. So unless they are behind a specific access control restriction, we do not track who exactly uses the data and what they do with them. So we have no idea ultimately. Nobody has come out of their way, like people could email us and ask something or say like, I'm having trouble with understanding blah blah blah, that has not happened, no.

**Sara Mannheimer** 21:57

Okay. Great. Um, so what about the social media data that you had in mind. Do you know much about like, context, challenges are strategies you use to sort of try and preserve context? In that case?

**DC09** 22:12

I'm thinking. So I wasn't the one the one directly leading the curation team on that. But we did talk about it a lot. Because again, it was in the end, so far, the only successful... successfully published one with with social media data. It was... Yeah, I think the context there was pretty straightforward. But okay, so here, I will just tell you a little bit about that project. And to be honest, I think, if you've talked to [a colleague, they] probably told you about it, because it's the only one. But it wasn't... It's not the typical social media data that you probably again, are thinking about. So it wasn't personal. It was Twitter feeds, but it wasn't personal ones. It was institutional ones of... even in the cases where there were like individual people's accounts, they were [writing publicly]. So it was either [public] organizations' tweets, or individual [public figures], but still referencing their published work, not so much like personal opinions or anything like that. And so in a sense, it didn't have some of... it still had some of the challenges like making sure that it's, you know, that we publish, republish the data in accordance with the terms of use was one challenge. The other one was making sure again, as, as you probably have thought about this, the links are, you know, terrible, terrible type of data to to publish. So we always do perma cc hoping that, you know, that will be around longer. So those those challenges were there. But other than that, there were no, like really human participant challenges. It was not somebody's personal tweet that this person had scraped. So in a sense, it was, you know, easier, maybe. And that's probably why it's so far we only successfully published one. But I don't in terms of context, it was because it was on a concrete topic that [published work was] about. The context was kind of there. And it was not the only type of data in this project, it was kind of one of three different components. So the other component really provided the context as well. It was about events in [specific locations]. But it was, you know, if you access the whole—as a secondary user, if you access the whole project, you would not be lost for context.

**Sara Mannheimer** 24:21

Nice. Okay. Yeah. And with news, it's like, the context is there if you're tweeting about [a published work]. Yeah. Assuming you can access the link.

**DC09** 24:29

Yeah, I'm thinking Actually, we do have one in process and I don't know that... I just remembered it because you know, I was only looking at the published projects, but we do have one in process that's—who knows what will happen with it. I don't remember exactly right now, what the... why it hasn't been published yet. But it involves... it doesn't it again, it doesn't involve just open social media postings out there. So in a sense, it's still different from the kind of typical thing that you're imagining, maybe. But it involves users of a, of a new app. Like a [type of app redacted], that's I think, if I'm remembering this correctly, that's supposed to [function of app redacted]. And so the, but I don't know, actually, nevermind. It's it's like a discussion of the app and all that, but I don't believe any of it is out there on like, any kind of open social media. So in a sense, it's very different. So nevermind, scratch that. Good that I hadn't thought of this little example before. But I don't know why I brought it up now, sorry!

**Sara Mannheimer** 25:34

Okay, let me ask you quickly—I don't want to take too much of your time, and when we're talking about both qualitative and big social data, and then what's different or what's the same between them... it can take a lot of time. But did you see similarities? What similarities and differences did you see between the context issues for qualitative data and the context issues you encountered with the Twitter data?

**DC09** 25:57

Yeah, in the case of this example, I don't think that there's much similarity. But I would say again, more typically, like thinking about these things, and what it might look like to do, to deposit—to publish a project that does use, you know, more traditional, let's say, tweets, or Facebook posts or whatever, something like that. Just trying to imagine, again, because we have been talking about it even not in the context of a specific project. I think the in terms of context, the challenges would probably be greater, I think, for, for social media items. Because as you say, I think in your, in your introduction to this section of the questions, because whatever is grabbed from somebody's stream of postings is necessarily, you know, quite disembodied from their whole social media persona, which not to even mention, is like a social media persona—is not that person. And I think in a sense, the context is so much more removed to begin with. So I think if I'm thinking about this, the way your question has made me think it's, I think it's more challenging. Yeah.

**Sara Mannheimer** 27:02

Okay, great. Let's move to data quality. So during your example, what challenges did you encounter, if any, when you're trying to document data quality, so like missing data, or quality of method or bias? And it's like, not really that, like, you're worried about the quality of the data, but rather, how do you communicate to reusers that this data is high quality and trustworthy? You know? And so if there were any issues, how did you communicate those so that future users would know?

**DC09** 27:33

Yeah, I don't... so first to say that I personally, and I don't think as an institution, we don't really think about data quality, because I think that implies kind of a, an assessment that's not ours to make. But we do think about data completeness. And I think, you know, the examples that you gave of what you think is quality, I think, has to do with that to some degree. So we do absolutely, again, in that in trying to get depositors to write this [narrative description of the data], we ask them very explicitly, can you explain, you know, first your sampling technique. Which again, for qualitative materials very often is, you know, some form of snowballing or some form of references, and so on. So please explain that. Then explain. You know, what was your full set of, let's say, transcripts, in this case? What was your full set that you had? and were able to collect and analyze that many? And then what portion of those did you, you know, are you trying to deposit because, you know, again, in the case of the specific one that we're talking about, where they did not deposit any full transcripts, it doesn't, you know, there's no, like, the drop off is down to zero, basically. But in some other cases, people might like, say, Okay, these people gave me permission, so I can do these, and I can deposit these, but then there's these other people that didn't give me permission. And so those I cannot at all under any constraints. And so we would have them describe this kind of process of really quantitative diminishment of the materials to to give a sense, like how incomplete really, in the end, the shared product is. But to just give a sense. So you know, that oftentimes, that would happen. So you know, you would have, let's say that many transcripts, and you can only deposit some subset of them, and then maybe even have that subset, you may have restrictions on some that only some secondary users might satisfy. So we will, we will just ask them very explicitly to describe that. That's kind of one our one of our approaches. The other thing we look at, and that's become, I don't know if it's become more of a problem, but it's been, lately, we've had some cases where the transcripts are kind of awful. And again, as somebody who works with qualitative and shows both you and I know that they can be especially if you're using automated software to do that, and without without additional manual labor to fix them, which is obviously much less time consuming, but they're just awful. So I have one depositor right now that I'm talking to, who's doing the process of, you know, preparing [their] data, but they're useless. Like [their] transcripts are completely useless. Now [they] had... And I told [them] that, I mean, that's, again, as as the repository we read through these things, for a number of reasons. But one reason is you try to imagine yourself as the secondary user, like, could you make any use of this? And in this case, the answer is no. Now [they have] full audio transcripts—sorry, full audio recordings, which is what [they] probably used for [their] own analysis, because even I'm sure the transcripts are even useless for [them], but [they] probably never had to look at them, because [they] had the whole audio. Now, she can't share the audio, but potentially could share the transcripts and we were discussing, like, what could you do and, you know, just purely for the purposes of data sharing, it's a lot to ask somebody to sit back down and re-transcribe, or at least even fix the automated stuff. It just, it's, I know how long it's gonna take. So I completely understand [their] concern with that. So trying to figure out what to do about that. But the point is, you know, we would not publish transcripts like that, like that just not, you know, in terms of speaking about data quality, that's just there's no quality, that there's nothing useful in them for a completely technological reason. So we are very concerned with that. And in terms of, and I don't know how that particular negotiation will end up, I don't think we've had such a such a stark problem on that front before. So that's why I don't know exactly how this will end. And sometimes we've had, you know, ones that needed some fixing. And we, you know, we really, we recommend what we think are the best practices, not just in terms of the content of the transcription, that's one aspect of it. But then, you know, people often come to us with transcripts without speaker tags, without timestamps, like those kinds of things are easier to fix. And we will lead them and guide them to do that as much as possible. So those are the kinds of things in terms of data quality that we try to apply. And that's part of the curation service and the curation value that we that we think we provide, so that they can be more useful and kind of a better, more accessible, and again, I really don't want to use the word quality. But...

**Sara Mannheimer** 31:58

I guess I could rethink that word, because I'm thinking of it more about like, trust, I guess, like how trust, like, how trustworthy is the data? And how do you like communicate to users that they can trust the data—that it's complete, that it's well documented? All of that. So? Okay, great. And then for quality, trustworthiness, completeness between qualitative and big social data? Do you have thoughts on the similarities or differences between the two?

**DC09** 32:31

Yeah, again, I mean, I was worried already, as I was reading your questions that I would have much more to say about the straightforward qualitative, than the comparison, so probably not.

**Sara Mannheimer** 32:38

If you don't, it's totally fine. Yeah,

**DC09** 32:40

I think what would happen and based on the one example that I told you, I have in mind, but also hypothetically, thinking about how these such projects with social media data might look, essentially two projects might look, I think the the process of asking people to document their own process, like how did you select these particular users or institutions or whatever? How did you select these particular tweets, or Facebook postings or whatever? We would ask them in the same way to just describe it's I think, in that sense, again, to the degree that I can picture those projects, it will be similar. Yeah. Okay. Just believe that describing the process, that procedural transparency is, you know, an important way, probably not the only one, but an important way and one that we can make public because there's other things probably people can do, but they're not easy to necessarily put as part of the published version of their data. But this is one that can be included in the documentation. And so that description of the process, I think, should enhance trust in secondary users. Like they, you know, they know what happened, whether they agree that it was a good process or you know, methodologically sound or whatever, then it's up to them. That's, I think, who should judge quality. But yeah, but the process description is fully there, and you can kind of follow it.

**Sara Mannheimer** 33:49

Yep. Okay, nice. That's really helpful. Alright, data comparability. So during your example, what challenges did you encounter relating to comparability or interoperability for your data set? So I guess thinking about metadata interoperability, and how future users might like combine two data sets, looking at different methods or with different research questions. So do you, as you go through your curation practice, do you think about this and what challenges do you encounter?

**DC09** 34:27

Yeah, I honestly, I personally don't think about this a lot. I think partly because I mean, really, for two reasons. Partly because I think in qualitative, the kind of qualitative social science that we have seen and that we, you know, the scholars that we serve, that that kind of social science. I don't think that that's a—I don't see people making that kind of effort. Like I don't think that's a goal, or research goal that people come to us trying to achieve. So in a sense, I just don't know that we, you know, need to fulfill it, because so far, at least, you know. But again, with a big caveat, we really don't know what people do with that data. Maybe everybody comes in like, is trying to do this, these comparable studies and are disappointed, and then they leave—that totally possible. So far, we have not made that a big priority. And I mean, I'm talking about kind of content comparability, if we're talking about, you know, otherwise, we do very much care about having our metadata be interoperable, being, you know, our catalog being part of all the big catalogs where people search for data. So in that sense, that's a big priority. So I don't I don't know, really, if I'm addressing what you were thinking about.

**Sara Mannheimer** 35:41

Yeah. For sure, I think part of it. And for, you know, this issue, like for big social data, metadata, interoperability is more relevant to the research itself. Whereas for qualitative, it's more about the repository, and like discovery, and finding and connecting with bigger catalogs, but yeah, thinking about trying to scale up qualitative research and like use, reuse published data for like longitudinal studies, you know, or something like that. I feel like that is some sort of ideal, ideal, use, you know.

**DC09** 36:18

I think maybe more I can see actually there being more like, current and future qualitative social scientists trying to do that, just because they've grown up in the in the, you know, in the digital age, I think the people that we still serve the, you know, the everybody that I've personally dealt with, and the people that I imagine, typically as our as our typical you is our typical depositor, I should say, That's not the kind of research that they're interested in it just again, I just not something that I think that they have, I mean, I'll put myself in that group that we grew up with. And so in like, if you're a qualitative researcher, you're not looking for that kind of huge scalability, I think of, of research questions. Again, that may be just my own particular take on what qualitative social science looks like. But I can also see what you are describing becoming more prevalent. And I don't know if those people just identify themselves as computational social scientists. And again, I'm hoping to discover some of that. It just may be, you know, that they, they identify themselves differently or brand themselves differently, maybe so I don't know. But one thing, and I don't know, again, you tell me if that or not, but we do care about and we have been working on another front. And that's again, more [the expertise of a colleague], because it's more technological. But we do think about interoperability of stuff that's, you know, qualitative features that have been analyzed with, with software, you know, analysis, packages, like NVivo, and Atlas, and all of those, because we again, we discovered that that's a that is a harder for people. So if somebody deposits to us, you know, doesn't deposit their raw materials, let's say for one reason or another, but does deposit analysis output from some package? You know, that's good, that's better than nothing. But what if nobody else, or very few other people have access to that same package? So there's actually been work, ongoing work that again, we inserted ourselves and then kind of tried to catalyze it further, to create this... an interoperable exchange mechanism among QDAS — that's what they call it, QDAS—qualitative data analysis software—among QDAS programs and that the programmers for these and others, because in many cases, you know, those are private companies, but the people behind the companies came together, and were willing to work on that. And I thought that was kind of big. So that's, you know, one small specific angle for achieving interoperability in qualitative data. It's, it's, you know, again, not probably what you're thinking about, but I guess we do think about some aspects, maybe not the ones that that you had in mind.

**Sara Mannheimer** 38:52

Yeah, that's really interesting. Yeah, because when you think about in the quantitative world using R and Python and other like, programs that are open source that then you can publish your code and people can easily reproduce. Yeah, I did look at what's the open source one, it's called, like...

**DC09** 39:10

I want to say QDex, but I'm not sure that's true.

**Sara Mannheimer** 39:14

Yeah, that could be it.

**DC09** 39:24

Can, I mean, again, the one that that I'm talking about, which may or may not be the one that you're talking about, well, I can send you the link afterwards because again, it's not directly but we are we QDR are involved in those efforts. And again, in fact, tried to, at one point, they had kind of died down. Other people had come together trying to do this. And for one reason or another, it just kind of fizzled out from what I understand. And so we've felt that it's it's worthwhile pursuing and tried to re energize them. And I think there is a working version of this now.

**Sara Mannheimer** 39:54

Mm hmm. Well, maybe I'll use that. I haven't started my data analysis, yet.

**DC09** 40:00

What it... what it does is it doesn't—you don't use it as your own software package, you use whatever software package you have access to work in use. And as long as it's one of the I think it's five or six, but they're really the main the main ones that people use, but as long as your whatever you are using is one of those, you it can then be reformatted easily. So all the things are not lost for another user who may not have the same package, so that's what this thing.

**Sara Mannheimer** 40:25

Oh, okay, very cool. Yeah, I'd love to see that. All right. Um, let's move to informed consent. So we talked a little bit about consent forms and how those factor in but what challenges did you encounter in this example, relating to informed consent for participants, particularly consent for future use of the data? And then what strategies do you use to address those challenges?

**DC09** 40:51

So that I mean, that's our like, of all the various questions that I read on your questionnaire that's the one I can probably talk the most about. That's a big, that's one of our biggest challenges when it comes to the human participant type of qualitative data. [Brief interruption]. So yeah, so people not thinking about the data sharing aspect, when they're conducting their informed consent process is, by far the biggest challenge. The strategies that we do, I'm going to tell them to you kind of in the same order that we offer them to people, the best thing that's kind of second best if you didn't address this in your informed consent, is to do reconsent. Basically, you're then able to go through all the details of what data sharing means what kind of data sharing you're planning to do, be super specific about, you know, here's their approach. And here's what their Terms of Use are, here's who accesses them, you know, give really, really detailed description of what's going to happen to these data and get, which means get truly informed consent, right? If people say yes to all of that, then they really understand what's going to happen. That's second best, but very, very rarely applicable, because people just don't, you know, have a good way of contacting the participants that they've talked to, you know, some years back in another part of the world records. And then the next one is okay, if you can't reconcile, and then we really very closely read what exactly the informed consent form says. And in some cases, so there's several kinds of categories of things that informed consent scripts, say, in some cases, it's so clear cut, like it says very explicitly, me or my research team are the only ones we're ever going to see these data, identified or deidentified. And in those cases, we really just can't process the data such. We then offer various creative suggestions of providing some transparency. Again, it's not really data publication at that point, but it's like one thing, the example that I told you with this, like giving examples of every... so these people had, you know, done, kind of straightforward, qualitative coding, thematic coding of their transcripts—of their numerous and super rich transcripts. Again, I'm gonna say like that, so that you understand what a loss it is. So they have done all this. And basically, all we were able to publish is, for every code for every theme that they've identified and kind of sub themes or whatever they had to give the definition that they use. So like the code book part, and then several illustrations. So that, unfortunately, we've published many, many projects like that. So that kind of qualitative diminishment, you know, I was talking previously about the quantitative where people can share all the transcripts, but some of them they can, blah blah blah. Here, we're talking about qualitative diminishment, that you can't do anything with the full texts, the full words that people told you, but then you can just give like little snippets, little examples, little illustrations. So that's one approach. That's an approach that often we find the depositors take if they've been required to share data by journals, because again, their article's about to be published, like that's all they really care about is making sure that their article gets published, and they accept that as the reasonable, you know, in terms of time, time, commitment, reasonable time commitment to produce that from their otherwise rich data.

**Sara Mannheimer** 44:06

I guess, for that, like for the reason that potentially the journal wants, like, so that people can go and like, check, and, you know, like, verify what you've done, but at least has allowed for that.

**DC09** 44:18

Exactly. And with again, that's why we offer it again, as the repository that's not our preferred solution. But anything else is much more time consuming. And so we also tried to understand the position of the of the person who's not at all thrilled about doing this, and are really just doing it as a, as a hurdle that they have to jump. What we do in all those conversations is kind of trying to not be too discouraged about the specifics of this process of this broad subject that, let's say has to do that. But to talk to them about the future and hoping that this message will stay in their mind for the next time when they're planning such a project that then they will do it correctly. So you know, who knows how long it... Who knows if it will happen, but also who knows how long it will take for us to even see the fruits of those conversations. But we always make sure that's a big part of the discussion that we have with them. Then going back to other ways of approaching this. So that's kind of the most the most minimal solution that we have come up with. And people often take it, I told you, in this scenario. In other scenarios, we may, again, depending exactly on what the informed consent says, oftentimes, it's just silent. It's like, that's kind of it's, it's problematic in its own way, but at least opens the door. So in those cases, we have said, we're not gonna we're never gonna publish identifiable data, if people have not explicitly consented to it. But if the data are deidentified in some form, and or potentially, you know, behind access controls, and what exactly we recommend to depositors really depends on the specifics of the case, and especially how, how sensitive the data are. So we even if the data not at all sensitive, we will not publish them if they're identifiable, if the person has not consented. Specifically, if they're deidentified, and the data are not sensitive, we may publish them, again, depending on exactly what was said. Sometimes there's conflicting things in informed consent. So one part of it says, You know, I will publish this draft or I will make... you know it never says published, but it says something like I will share, I may share the deidentified transcripts with the with the research community or with other scholars or something like that. And to us that's like enough of a license to, to go to a repository like ours that does serve just scholars, and that does have some kind of limitations on who can access the data, and so on. So. Oh, in one part, so, one part of the informed consent might say that and the other part might say, you know, but they're going to be kept under lock and key in my office, and then we're like, just, okay, let's, let's figure out, you know, we basically will have like, an interview with the depositor and ask them: What would you say your participants really expected? Do you think that you emphasized that everything would be under lock and key? Did you make it clear that, you know, you will be sharing with at least other scholars and like, it just, it's, you know, we do these interviews on the informed consent form, and have different, you know, branching trees about where it ends up in a given project? We've actually systematized it for ourselves. Like, that's not a public document. But essentially, what I related to you, written out so that we can follow it consistently every time.

**Sara Mannheimer** 47:30

Very interesting.

**DC09** 47:33

I think. Yeah, before that, we were always doing it. But it was, you know, more... I mean we always did it but was not like written out as a procedure. Whereas now whoever exactly is leading the curation team follows the exact same steps like if, if the informed consent is of this type, follow this path, if the informed consent is of this conflicting, self contradictory type, follow this path, and so on, and so on. And to be honest, a lot of the endpoints of these branching trees, say: do not publish the data. So we ended up having to just discourage people. But I gave you a sense of some of the strategies to maybe maybe make it make some portion of it publishable or some version of it publishable.

**Sara Mannheimer** 48:11

That's great, thank you. Okay, I'm going to skip these ones about big social data, because I'm also interviewing people who, unless you have some things that you want to bring up there,

**DC09** 48:23

I just wanted to say that, again, this is more like us trying to imagine both the specific conversation we're gonna have at this conference I talked about but also potentially eventually serving such scholars, I think the challenge of the informed consent there, again, probably will be greater—certainly will be in a different form, but it will probably be greater. And I think I've actually seen discussions amongst IRB people, because I told you, we're trying to be in touch with the IRB community, and I'm on a kind of a professional forum that IRB personnel are part of, they are discussing, sometimes they're discussing that same topic and what they require. And so for example, they also do something like, you know, if somebody is proposing a project where they're going to be collecting social media data, in the discussions that I've seen, IRB personnel really want the researchers to identify themselves. So nothing like you know, post-fact scraping of stuff. To identify themselves to say like, this is what I'm doing this is... basically to do some version of informed consent online. Now, again, only some studies I'm sure can follow that model, and others just don't. So that's that's kind of the big dilemma, how one will handle the ones where that's not possible. That's just not the research approach. So I have no answers, but I'm just saying I think that that's that's the that's probably a big part of the conversation. And what we'll be asking these computational social scientists. Have they even been bothered by that. Have they even thought about that. And then if they have, what have they even done, like just to see in practice, what do people do?

**Sara Mannheimer** 49:56

Yeah, some things I've read about are like bringing together focus groups. If you have scraped data from a particular type of community, you could bring together a focus group with a few people that might be representative of that community and ask them what they would think. Or you can have like a community advisory board who can help you when you're planning the project. But that only works if you have one specific community. I mean, if your community is like all Facebook users, then yeah. Okay, cool. Let's move to privacy and confidentiality. So during your example, what challenges did you encounter relating to privacy for the people represented in the data? And then how did you address those challenges?

**DC09** 50:39

Yeah, again, in this case, because the informed consent was so contradictory to data sharing, in the sense that removed the privacy and confidentiality challenge. I think that becomes a challenge only in the cases where the informed consent even gives some clue to people or had given some clue to people that their data may be shared. And then we start thinking about, okay, what is the you know... but but it says, it will be deidentified, let's say, or that your identities will be kept confidential, whatever that means, because I've discovered it means all kinds of things to different people. But then we start thinking about, you know, to what degree is manual, substantive deidentification possible in this in this context. And then if not, or even if it is possible, but insufficient, do we then need to apply access controls those, those are basic solutions, I mean, there's not much beyond that, that we've come up with, we have one really, really good example of what we think is properly / sufficiently deidentified data that we made that remains still very rich and useful for secondary users. But it was done as a pilot project, it was done like to see how one can do this. So they had the time and resources and paid to do it, and so on. So it's really not translatable in most in most practical other cases. So basically, what we learned and how I would summarize our experience with that particular pilot was: it can be done. It can only be done by the original researchers, like there's no way we could do this, we gave a lot of suggestions, and we read through with them back and forth, you know, to make sure we we suggest new things, other things that they may need to mask and kind of, they never removed the stuff completely. But we figured out kind of aggregate versions of descriptions and attributes, qualitatively how to do that. So we don't think that we at the repository could do that, or anybody else for that matter that's not the primary researcher who really knows their context and their participants. So that was one thing we learned, which again, immediately limits the possibility of it being done widely. So only the primary researcher and research team can really do proper qualitative identification where the data still remain useful. It's very time consuming. And there's just currently no practical way for people to do this. It was not, you know, not not an optimistic result. But on the other hand, it can be done. So I don't know where that leaves us.

**Sara Mannheimer** 53:11

What percentage would you say of projects get restricted access? You said it's not that many.

**DC09** 53:18

Yeah, we should probably actually come up with our statistics for that. I don't... I would be making a guess right now. It's not that many. It's probably do you want me to like, I could probably give you a more precise figure after if that's okay.

**Sara Mannheimer** 53:34

That's okay. And then I'm also thinking about, well, I guess we'll talk about this in intellectual property a little bit... like your terms of service. You were saying that [the repository where you work] is mostly for academic researchers. Is there like a some kind of form that people sign to say, I am an academic researcher. And here's my IRB approval number for this. What research I'm doing now or something like that to use the data in [the repository].

**DC09** 53:59

So there's both of those but in separate places and for separate users. So the very basic, you know, [form, in which] you say that you're gonna—you don't say necessarily that you're a researcher, but you say that you're only going to use the data for scholarly and academic purposes. Okay, that's our just basic, you know, terms of terms of conditions and use. Then for projects where the depositor wants to apply restrictions, one of the possible restrictions and people do take advantage of that one, is to say, "beyond that, I want you to submit, let's say a research plan." That's actually the most common combination that people—researchers are satisfied with for conditional secondary access is if [the repository] can confirm the academic affiliation. So we actually then make sure that somebody is writing us from an academic email—is really what we typically do. Because they could register with a different email, just click that, yes I am only going to use this for scholarly and academic purposes but if the depositor wants more than that, then we would actually ask the secondary user to write us from an academic email. So we'll do that to confirm. That's our way of confirming academic affiliation. And then the depositor might ask for a research plan. And we ask the secondary user to submit a research plan that basically describes roughly what they're, you know, looking for in the data. So that combination: confirm the academic affiliation, brief research plan, in most cases of restricted use satisfies the depositors. In some fewer cases, they have also wanted an IRB, a secondary use IRB application, which, again, from our conversations with the IRBs, they hate that. They don't think that's their place to do. Especially for deidentified data, it really honestly isn't, if you read the Common Rule that they follow, they shouldn't be asked for that. But yeah, it makes some depositors more comfortable. So you know, many times with the IRB, the IRB is kind of reluctant, they can't even really truly review a protocol, because for them, it's like, not human subjects research. I cannot do much more with your request here. But they also understand, you know, they understand the whole dilemma. So what they will do is like write a letter saying, This person is okay to access this data, but it's not a formal, you know, approval from their point of view. So anyway, in some cases, that's really very few. One comes to mind. I think there's more, but just one comes to mind where somebody has that and the reason why that particular person—that particular depositor had it as part of their access restriction conditions was because their original IRB actually had said that they say that the data can only be used if another IRB approves the other project. That, yeah, that that's that's that case, I'm trying to remember other things people have asked for the in the same vein, but not the same thing as actually secondary IRB approval was somebody who wanted. So academic affiliation, confirmed research plan, and human participant protection training of some sort, like CITI, or something like that. To them that was kind of just just for a person to signal that they're serious about these things they've thought about them. So that's several different, you know, conditional things that go in the vein of trying to make sure that the data not just out there for anybody.

**Sara Mannheimer** 57:16

This is, this is one of the really interesting part because IRBs think of social media data as data reuse, as existing data. So they're like, oh, we're not really involved. And I feel like I don't know, it still is human subjects data, like, I guess the definition of human subjects data is a little murky now. It's like, if it's like existing online, does that mean that it doesn't require the same types of ethical review?

**DC09** 57:48

What we have learned from our interactions with IRBs, and we did also some interviews and focus groups, we didn't specifically ask about that, but it had come up in those as a, you know, kind of a thorny scenario, because we generally ask them about thorny scenarios. What they, what I've learned is that they really go by the I mean, most of them are just very closely looking at the letter of the regulations that they have to follow. And so maybe personally, they would agree with, you know, there's a need for some kind of ethical review, and most people would think IRB is the place for that. But that's not what their regulations say. So they really, that's where the struggle is for them is, and they're, you know, like everybody, they're overworked. And they have enough of the projects that do fall under their purview that they need to pay attention to. So they're like, why are you bothering us with this other stuff that we really are not responsible for. Find another place, you know, do it the way you understand it, but don't bother us—I mean, I'm exaggerating, honestly. But, you know, that's, I think that's kind of their position, but at the same time, and it a lot of times comes down to also just the personalities that staff these positions, to be honest as well. And so some of them are not willing to just leave it aside, because the regulations tell them. Yeah, so it's, it's when it comes to these kind of new types of research work. It's the I mean, the reality is just and that's, I think, probably always has been the case—that regulations take a long time to reform and revise and the Common Rule was just, practically just revised. But it was the result of like, 10 years of work prior to that where research online looked very different from how it looks now. And so you just can't catch up with it. And if your institutional role and position is bounded by these regulations, you know, yep, that's just what you go by. So that's that's, that's the crux, I think of that dilemma. Many of them think about is it just that again, because of these other constraints on time and resources, they would rather not be bothered until there's clearer expectations and clearer guidance.

**Sara Mannheimer** 1:00:03

Yeah, for sure. Okay, let's move to our last question intellectual property. During your example, what challenges did you encounter around intellectual property concerns? Like participant IP maybe? Or sometimes data can be construed as belonging to the institution where the researcher works? Is this something that you'll think about when you're curating data?

**DC09** 1:00:30

Yeah, I mean, again, it only comes into play, when there's some kind of specific issue. So in the examples that I gave you now, it will, you know, that was not a concern. And I don't know exactly, you know, when you... So the second specific question about the fact that the data technically always belong to the institution, even if researchers don't realize that, that would think about, and I'll tell you an example that does come to mind, it's not one of the ones that we've been discussing, it's a much, much older one, where we were not able to publish the data, because of these constraints. I'll tell you something that... but it really only comes up, who owns the data -- if that's what you mean by intellectual property -- only comes up when there's some, you know, unclear scenario. Now, in most cases, it does not come up, let's say that. Yeah, um, but in that one case, it was it was both the IRB was involved. And then the fact that by the end of the whole process of trying to... basically the person that was trying to deposit the data had not gotten informed consent from the beginning, informed consent for sharing. I mean, [they] had gotten informed consent, but not for sharing. And [they] really wanted to share the data, and [they] wanted to do it, right. They were also very sensitive from a very kind of particular, identifiable population, so [they] actually could go and reconsent. And that was, that was what [the researcher was] trying to do, which would have worked, you know, if it could have worked out would have been great. But [the researcher's] IRB basically was on [the researcher], because if you're going to be reconsenting, then you have a whole new, you know, you have to have an amendment of protocol, and the new informed consent to be reviewed and all that. [The researcher] went back to the IRB, I think they were kind of on board, I don't remember any great resistance, they had some more questions about what the repository does and what our protections are. So we went back and forth for a few months. And then by the end of all of that, [they] graduated, the depositor. And so [they were] no longer really affiliated with this institution. So the IRB, like when it was time for them to make the final decision, they're like, well, [they're] not, you know... we're not authorized to say anything about it, but [the researcher] could not take the data with [them]. So that's where you know it... Yeah, it was very disappointing. But the point is that that's the kinds of scenarios in which the discussion of who owns the data and where do the data stay, and especially like, in this case, again, a student not not like a faculty moving on to a new position, which I think would the same problem might present itself, but the solution might be slightly different. I don't know. In this case, it was a student. And so [they] just couldn't, I mean, technically, [they] still had the data, like they were on [their] computer. But technically, [they were] not supposed to even have that there. And so [the researcher] certainly didn't get formal permission to go reconsent and publish in a repository that like, became out of the question, because because [they were] no longer affiliated with the institution that claims to own the data.

**Sara Mannheimer** 1:03:14

Hm.

**DC09** 1:03:16

And again this is the other part of what I'm going to say about this is not something from previous experience, but just from monitoring that IRB, you know, discussion forum. In the case of faculty members, that's often apparently a problem that IRB is called to resolve. And they're much more willing to come up with creative solutions. And again, even though the faculty members don't technically own their data, they will often allow them as long as they can provide the same kind of security conditions or whatever. They would often allow them to, you know, take a copy or whatever. So. Okay, I know, again, I know that I already discussed this as well. But that's just from secondhand knowledge.

**Sara Mannheimer** 1:03:56

Very interesting. All right, let's go to number seven. Are there any issues or challenges that arose during your example that I haven't asked you about that you think, are important for me to know?

**DC09** 1:04:12

Yes, but not like... I don't know if it's in any way relevant to you other than just as a common, it's actually a common concern that affects some of these things that you are interested in. So the example where we published a very diminished, you know, excerpts type of data versus the full transcripts. The person that I was communicating with, was not the primary investigator but their project manager. So that that's only relevant to some types of research, but still to many. And so apparently in public health, that's very common. And so this project manager, you know, I understand the need for this position, and especially if you have [more than 10] people collecting data, you know, over multiple countries and all of that it's it's necessary and the PI is trying to move several projects at the same time, so they do need point people to deal with, but it was the very rare case where the PI didn't want to even—I'm assuming, didn't want to because I suggested it, and the project manager kind of ignored it and never did it, but—didn't even want to be copied on the correspondence. So like, it was a lot more. You know, again, it's secondhand in a sense as well. So that's the kind of a, I don't know, it's a challenge. I don't know that I think it impacts some of the other things, but maybe not in such a drastic degree just makes everything even slower. And even like, [they] always, you know, the project manager always has to go back and ask the PI, and then, you know, whatever [the project manager] relates, I go by, but is that really what [the PI] said? So that's kind of that was a challenge. I don't know to what degree it fits in your kind of interests. But..

**Sara Mannheimer** 1:05:39

Yeah, that is really interesting. That's like, yeah, a level removed from the, from the participants, and then even one more level removed from the PI and so yeah. Great. Okay. Last last question I have for you is, who else should I interview.

[Discussion of who else Sara should interview]

**Sara Mannheimer** 1:07:17

Yeah. Cool. All right. Thank you so much. This has just been really fun.

**DC09** 1:07:22

You're welcome.