DC03\_transcript\_deidentified

**SUMMARY KEYWORDS**

data, deposit, qualitative, qualitative data, study, curation, curator, researcher, respondents, included, transcripts, curated, metadata, worked, context, terms, questions

**SPEAKERS**

DC03, Sara Mannheimer

**Sara Mannheimer** 00:02

Okay, so to start us out, can you tell me about the types of data you usually curate? And then what your interests are regarding data curation?

**DC03** 00:13

Yeah, so I've been [a curator] for over, just over three years, and I've worked on a lot of different types of data, mostly quantitative, we get a lot more quantitative than qualitative. But that I have, I feel like I have seen an increase recently, and the amount of qualitative that we get, whether it's like strictly interview transcripts are just long string variables in quantitative data sets, we get quite a bit of that as well. Um, the kind of social science areas that I often work on are [redacted].

**Sara Mannheimer** 01:38

[Redacted -- brief discussion of the details of DC03's workplace]

**Sara Mannheimer** 02:17

Okay, so next, we'll identify a specific example. I'm using a strategy called critical incident technique, where we try and look at one specific example. And it'll hopefully try and help us get some more concrete information as we go through here. So either when you support a curation for big social data, or qualitative data for publication, or sharing or observed firsthand, but you will have done it yourself.

**DC03** 02:48

Yeah, qualitative, specifically, I have worked on really large data sets, but nothing that I would consider big data in that sense.

**Sara Mannheimer** 02:57

So tell me about your example. Have you thought of one already, or?

**DC03** 03:09

Yeah, I went back and kind of looked at the, the one that sticks out the most, and also the most recent that I actually curated. The one that I curated most recently was actually released in [a couple of years before this interview]. So it has been a couple years in that sense, but um, that one was pretty much strictly qualitative transcripts. There were [a few dozen] transcripts. And it was, the study focused on [a sensitive health topic]. So quite a sensitive topic as well. And it was in a very narrow geographic, it was, it was within [a small population in a specific geographic area]. So in terms of like, just sensitivity, a nature of the data, it was very, very tough to read through, to read through people's personal experiences with that, because there were different, different respondent sub-samples to, so there were like [lists several different respondent types], so yeah, so that's it sticks out just in terms of like, the one I did curate personally, most recently, and just like, was very impactful while reading and reviewing it for disclosure concerns. So.

**Sara Mannheimer** 05:17

Yeah, I can imagine. I look forward to hearing more about it. Was this part of a grant project like that had a data management plan, or required specific treatment for the data?

**DC03** 05:30

Um, I don't partake in those types of conversations. That's [a different part of the organization]. So they built a relationship with that researcher and said, "Hey, this is really interesting data, we think that it could benefit from [more in-depth] curation."

**Sara Mannheimer** 06:13

Okay. So let's just move down to context. So I have a little quote, to give you an idea of what I mean by context here for qualitative data. "Since qualitative research is a process that may include deep and prolonged contact and connection with subjects attempting to understand subjects within their own context. Qualitative data are highly context dependent." So Hines et al. write that "context is a source of data, meaning and understanding and ignoring context are under using it or not recognizing one's own context driven perspective could result in incomplete or missing meaning, and a misunderstanding of human phenomena." So thinking of that, and I'll skip the big social data definition, since we're not talking about that. But thinking about that, like thinking of context in that way, during your data set curation process, what challenges did you encounter, if any, when you were trying to capture the context in which the data was collected?

**DC03** 07:24

So within curation we're pretty limited to creating that like metadata, based off of whatever information, the PII or the researcher provides us. So usually, that's through, you know, documentation, like, if they had if they gave us informed consent [forms], or how thoroughly they provided a summary and methodology description... is really limited to what they provide us. Or if they provide us like a website that we can go and glean additional information from. I don't feel like in a lot of my experiences with qualitative data I've ever received, like, the researcher's personal contacts, it's more so just very broad, like study summary and study, like what you would find on a research paper, and not so much that the more those types of like little small caveats that might have. The I don't know that the researcher would would be only really be able to provide that context on.

**Sara Mannheimer** 08:29

Are you able, like, do you ever reach back out to researchers to ask for a, you know, if they don't provide consent forms or other info that you think would be helpful? Do you ever reach back out? Or is it just whatever they provide at the time?

**DC03** 08:44

Um, it's really project specific. So like, [some researchers are required to include consent forms, depending on their funders and workplaces], so that's an easy like, Okay, if it's not there, I ask for it. Um, but other ones, it really depends on what kind of relationship we have with the data producers, or [the type of deposit], we, it's hard, they don't always respond to our inquiries when we have questions and stuff like that. So that's another challenge is just not being able to contact them or get that reply from them. Yeah. This particular researcher did provide consent forms, which was really super helpful and also provided their interview scripts as well so that that was really helpful to use you know, what kind of questions were asked. They were semi-structured so, so obviously, you know, they asked little side questions or ask a respondent, you know, to elaborate. And those types of things are often... those probing questions are often included in the interview transcript and I can't quite remember off the top my head if they were in this one. And the only way, [if a user wanted] to know, would be to actually apply to get that data and actually look through the transcript and see where they ask those questions.

**Sara Mannheimer** 10:09

Okay, so the data was restricted, is restricted.

**DC03** 10:15

Yeah.

**Sara Mannheimer** 10:15

Okay. And we'll talk more about this. But did... I feel like another way that context could potentially be compromised is during the de-identification process. So was there, did you use de-identification? Or did the person who created the data, the data producer?

**DC03** 10:38

Yes, they had de-identified names and like direct identifiers. Within the consent form too it said that their names would be de-identified and made anonymous. There are also some things listed for like [one type of respondent], that references to their [location] would be anonymized as well. And that's, that was one of the things that I was finding, in my review, while we review these data to be released, that some of them were missed, or acronyms were often missed. So, so yeah, we [employ a strategy to reduce the risk of reidentification], to make sure that those things are, are de-identified.

**Sara Mannheimer** 11:24

Tell me more about [this strategy]. What does that mean?

**DC03** 11:29

Um, so that's something [the organization], I think, really, we really pride ourselves on doing a very thorough [review]. So anytime we identify something as a risk, we'll discuss it, like I'll discuss it with my supervisor. And we will develop then a plan on how we're going to [address] it, usually, it's just masking. So removing the word, we use a standard descriptor, [details of the process redacted], so that the user knows that it's us that removed the text, and not the, the, the data producer, they'll use whatever method they want to use, you know, usually, they'll just say like name or something like that. Everyone uses kind of a different method, our method is to use [this descriptor]. And that descriptor is something broad, but still provides at least some type of context to the user. So that, you know, they know it was a name, [but it has been redacted]. So yeah, we do try to give them at least enough context that they'll be able to follow along with the, the transcripts, but still protecting the respondent's identity.

**Sara Mannheimer** 12:46

Nice. Okay. Great. So I guess to review, your strategies are mostly about documentation and bringing in as much information from the data producer as possible about the way the study was done?

**DC03** 13:05

Yeah.

**Sara Mannheimer** 13:06

Sounds good. And we'll skip all these 1a, 2a, etc [referring to subquestions in the interview guide], because we're just looking at the one type of data set. So we'll move on to data quality. So in this example, what challenges did you encounter, if any, when you were trying to understand and document data quality? So missing data and bots are not applicable, but maybe bias, quality of method? And then what strategies did you use to describe, clarify, and communicate these issues?

**DC03** 13:42

Um, well, I think again, kind of how we [conduct redaction and masking] still helps in terms of like trying to maintain as much quality of the data, we also we really want to make sure things are accessible [in the repository]. So if a transcript is provided to us in a text format, like in Word, or PDF, or a plain text file, we provide it in in all of those different formats— we provide it in PDF, Rich Text Format, and plain text format, just so that users dependent... you know, they might be throwing this into some other type of program, it's accessible to them and whatever, you know, whatever they have paid for on their own computers to be able to access the data. That's kind of something that we do at [the organization].

**Sara Mannheimer** 14:43

Yeah, I'm trying to like, I guess with quality, maybe it might not even be like your... maybe you might not even think that it is your responsibility to, or we you know, we might not find it our responsibility to address quality because It does seem to be more of like, you know if it's about the method and but I guess I'm thinking....

**DC03** 15:09

We try not to change anything. I'm like, we don't change typos. We don't, you know, fix grammar errors or anything like we really... we're archiving, we're archiving what was given to us. So, um, so yeah, so we, when it comes to me, at least, I'm, I'm really, especially with qualitative, I'm looking at, you know, the, the respondents confidentiality and protecting that, more so than, like, with when we work with on quantitative data, that's, that's something different where we are, you know, trying to make sure all everything's labeled correctly, and, you know, documented within a codebook, and whatnot, there, there's a lot more in terms of that type of value added stuff that we can do with quantitative data sets, that are, I think, a little trickier to do with qualitative. Because a lot of you know, a lot of times are just as a Word document, and everything's just, there's not too much formatting or anything.

**Sara Mannheimer** 16:13

Right. Do you include, like, information about how they analyze the data, you know, like, you know, how they coded themes or anything, in terms of analysis?

**DC03** 16:34

If they provide that, yes. I know, I've seen some [datasets] where, PIs have given their coding schemes for things. It's pretty inconsistent, though, in my experience with qualitative data as to when data producers give us that information or not.

**Sara Mannheimer** 16:58

And so for this example, we're talking about with the [example we are using] did it... was it just the transcripts? Was that the full data set?

**DC03** 17:07

Yeah, the the raw transcripts. And then there was a demographic, quantitative data set. Okay, yeah. So really kind of leaving the user to build those coding schemes themselves from a very kind of blank canvas, I think. Yeah, I guess I use it, we also, through [a part of the organization focusing on metadata]. They go and try to attach publications related to studies. So that would, I guess, be another way that a user could they could go and see what studies have been released that are associated with this data and look through their methodology section, together, that kind of information as well.

**Sara Mannheimer** 18:01

Great. Okay. Alright, let's move on to number three: data comparability. During your example, what challenges did you encounter, if any, relating to comparability or interoperability of the data set? So different research questions, different methods, metadata interoperability? Did you think about any of that? And did you use strategies to address challenges there?

**DC03** 18:37

Um, I can't think of anything that I you know, I could see that sometimes questions were asked out of order from their, the interviewer's, you know, transcript that the or the, from looking at the transcript compared to their interview script, you can see when the PI asked things out of order, but I think that sometimes the the nature of semi-structured interviews that they you know, jump around sometimes or you, you start talking before you start recording, you know, and but there's again, there's nothing really, we have, my role as a curator, I don't really have too much say in how a researcher often prepares what they're going to deposit. So.

**Sara Mannheimer** 19:25

In terms of [your organization], what strategies do you use to for like metadata interoperability? Do you use like the Data Documentation Initiative metadata?

**DC03** 19:37

Yes, yeah. We don't release... like when we have quantitative data sets, our DDI is automatically generated through internal processes that we have, but we don't have any type of automation, for creating that for qualitative transcripts like that. So we try to do some of that, in more table-level information. And using like README files, that kind of, they'll just list like, you know how many transcripts were deposited, and are being included. We're trying to get a little... Because it's been, it's been a little inconsistent in the past. And we're trying to build better consistency, because we're seeing an influx of qualitative data getting deposited, we're trying to, you know, get a, get a jump on, on how we need to make sure we're preserving that type of information.

**Sara Mannheimer** 20:42

Cool. I'll be curious to see, like, what strategies you use to standardize the qualitative data more. So for README files, it's just sort of open format, you just write whatever needs to be communicated? [DC03 affirms]. Okay. Great. All right. On to question four, we're halfway there. In your example, did you encounter any challenges relating to informed consent for participants? Particularly consent for future use of the data? And if so, what strategies did you use to address the challenges?

**DC03** 21:22

Yeah, so so this PI did give us an informed consent [form], that it was tailored for each of the different sub-samples of respondents. And there was a question and then the answ-, you know, like, by consenting to this, you agree to this data being deposited. I think it said, in a public archive? And the study actually was initially set to be [publicly available], so that pretty much anybody could download it. But through my [curation process] and communication with my supervisor, and then with the PI, we decided no, this is just too sensitive. And too, you're able to re-identify participants too easily just as it is, to be able to do public. So it was changed to restricted access. But just having that information was really helpful. And again, the same like with the, [locations] were supposed to be masked, and I had found some instances where they weren't. So that was like a really clear, no doubt, okay, I need to remove this information from the transcripts, because the respondent was told that that that one, that you put in there, the other things that are more broad, like, I feel like sometimes they'll say, yeah, your your information will be de-identified or you'll be made anonymous, but especially with with qualitative interview transcripts, people really like to talk and go into really deep detail about themselves, especially with something that was so personal. Just a couple of Google keywords, and I could find these respondents really easily, which was part of the reason of why we decided to make it restricted. And so what does that mean for the respondent? You know, they they're agreeing to be anonymized, but they're also providing all of this, you know, kind of extra detail. What were their expectations? It's sometimes hard to, especially if they're not coming from a research oriented background where you know, it, are our expectations the same? And so we lean towards protecting the respondents', you know, identity. So even if it's an indirect identifier, when you can combine all of these, what can we remove to make sure that they're still getting as much context as they need for analysis, but still protecting them?

**Sara Mannheimer** 23:55

Yeah, that's a really interesting question. Like, what were the respondents' expectations? Because Yeah, it's easy to say, Oh, yeah, I believe in open data, I want this to be useful, please make it available, but not knowing really, what the de-identification process looks like, or you know if it's possible. So I guess, I'm thinking about like, I don't know. I'm just, trying to think of like, how to dig deeper into this. But like, so basically, you're, when you are working with your team at [the organization] and with the data producer, you decided that even though these respondents had consented to their interviews being published, they didn't exactly understand all of the implications of that. And so then you decided that it couldn't be open?

**DC03** 24:59

Yeah, I think that's kind of, that's ultimately what we decided. That it was just, some of the information was just so so specific that, like, if I were just having a conversation with someone, you know, that I was meeting in a coffee shop or whatever, like, maybe it might be something I might disclose. But when you're talking about something, in relation to this really sensitive topic, you know, we just wanted to make sure that ultimately, they were being protected. Yeah, yeah.

**Sara Mannheimer** 25:47

And what did, who did you restrict the data to? What are the terms of use?

**DC03** 25:56

Um, so I believe [researchers] can apply for it. But usually, with our restricted access, you have to have applied for an IRB, have IRB approval to to download the data. And then just kind of, you know, agree to the terms of use, you know, can't try to, can't use the data to try to re-identify people and stuff like that. We include some some language in [the data record] on you know, what kind of restrictions are there. And then when we include like processing notes, we say, hey, this, this data is restricted, you'll have to follow the terms of the agreement. But generally, for restricted, secure download. It's an IRB approval.

**Sara Mannheimer** 26:51

Okay. And so that probably means that users would be academic researchers?

**DC03** 26:59

Yeah, yeah. We also have, we also have [data accessible only through a data enclave]. So that's where you have to you, you can't just download the data to your own personal laptop, you actually have to have access to our [secure network] to use the data. So that's like a higher level of restriction. And then we also have a physical data enclave where you actually have to physically be in [the organization's location] to access those data sets.

**Sara Mannheimer** 27:26

Oh, okay.

**DC03** 27:27

So we have different levels of restricted access.

**Sara Mannheimer** 27:32

Nice. Can you direct me to like a place on your website where all of that is described if there is, or like your Terms of Use are outlined?

**DC03** 27:59

Let me find something quick... [directs Sara where to look].

**Sara Mannheimer** 28:49

Awesome. This is perfect. All right. Well, I already talked through five. Is there anything else we missed about privacy and confidentiality? You talked about de-identification and [the process of redacting and masking datasets to reduce risk of reidentification]. What other strategies are we missing here?

**DC03** 29:17

Let's see, we'll [redact and mask] locations, anything less than whatever the study level geographic areas supposed to be. Sometimes the respondent will save the name of like a church or a denomination or like a coffee shop and we'll Google search and find out like, okay, is this a local place? Is it a state chain? Is a regional chain? And so those kind of like, even though it's a coffee shop, it's not like super directly identifying a respondent if it's specific to their town and state is supposed to the highest geographic area, you know we'll, again, take those steps to de-identify the data that way. So we take that pretty seriously. And that for qualitative data specifically, is what we spend most of our time on in the curation side of things.

**Sara Mannheimer** 30:24

Yeah. Do you read through each line individually?

**DC03** 30:28

Yeah. Yeah, that, we we will not release data that we have not had our eyes on. So this, this study took... between my review, and we also have two rounds of quality check on this type of an intensive level study. So it, it was there was roughly [between 10 and 20] weeks of time logged on this study, from [the PI depositing the data] to [the data being available], which is pretty typical of our qualitative, our qualitative. We have different levels of curation at [the organization]. And anything that is like that has a qualitative component like transcripts will automatically be set to [the highest level], just because we know it takes a long time to read through the data. Yeah.

**Sara Mannheimer** 31:28

Okay, I think that covers it. Well, this is our last question. During your example, what challenges did you encounter regarding intellectual property concerns of archiving or publishing the data? So like, participant intellectual property, or intellectual property of the data creator's institution or anything like that, and did you use strategies to address it?

**DC03** 31:59

Not in this one, I... there was nothing regarding intellectual property with this one, I have had other studies I've curated, where the [PI] included a copyrighted instrument that it's, you know, when you read through it, as it says that it's it's copyrighted, but they've included that data within the data set and within their, like, you know, full questionnaire. And so that was just me like going back to the PI and being like, "Hey, is it was this supposed to be released? I'm not sure. You know, it says it's, it's copyrighted instruments. Did you have permission to to include this with, with your deposited data?" In that particular case they did not. So they had to resupply redacted materials. And I think like I couldn't, the data was restricted access. And they had permission to release it as restricted, but like I couldn't do, I couldn't put the the DDI the variable level metadata for that, [I couldn't make that study publicly available]. So. So there are a couple of different things that we had to do in terms of that type of intellectual property. That's the more common thing that I've seen within curation where we have a copyrighted instrument that wasn't supposed to be included somehow.

**Sara Mannheimer** 33:20

And do you put licenses on data [at your organization]?

**DC03** 33:26

I don't, I don't think so. That doesn't sound familiar. No.

**Sara Mannheimer** 33:33

Okay. So how do people like when? When people are coming to use data [from your organization], what are the like potential routes in? How do you sort of like regulate use of data? Oh, I'm not hearing you for a second. It might be..... I still can't hear you. Shoot. It might be my internet connection. Hold on. It's good. No, I can't. Okay, let me try taking my headphones on. Okay. Yes. All right. Bluetooth off. Okay, sorry about that. Go ahead. Sorry.

**DC03** 35:06

Um, what was that? I can't, trying to remember what what are we in?

**Sara Mannheimer** 35:12

Terms of, or use, like how you regulate use of the data sets.

**DC03** 35:18

It's very project specific. So depending on who the funders are, there might be different, you know, different policies on that. There are some things that come in through that might be just just are able to be disseminated to the public. But other things that are [only available to certain users]. So that's all very spot project specific and study specific and us on the curation end, like our [supervisors] will give us that information. So they work with with with the funders and with the [researchers], and depositors to decide what that needs to be. And then we just make sure it's included in the metadata, we click the right buttons for the metadata so that on the user's end, they're able to access it, however it's been deemed appropriate.

**Sara Mannheimer** 36:16

Okay, sounds good. Well, that's it is, are there any other issues or challenges that arose, especially during your example, or anything else that you're thinking of, that we haven't touched on?

**DC03** 36:34

Um, it was just a really tough subject to read. So that I mean, just, just from a mental health standpoint, it was, it was very heavy. I, my... So I mentioned that [many weeks] worth of time was logged. But it took from, I was assigned the study [Month Year], and it didn't get released [for several months]. So I had other assignments that my supervisor gave me, just so that like, I could split my time throughout, you know, throughout the week, so that I wasn't constantly, like, having this heavy subject material. So that was, I think, just part of it was the review process was very heavy. Yeah.

**Sara Mannheimer** 37:24

Did you put any kind of like content warning on it or something? I mean, I guess there's a description of the project. But...

**DC03** 37:31

Right there is there is a description, we are actually talking about that internally a little bit about what kind of like trigger warning should we be providing on any data, not just qualitative data, but any data internally. So [some data we review is] related to crime, violent crime. [Additional details about internal processes redacted]. So if I do find something that could be triggering, I am supposed to alert [person in charge of the deposit so that they can add that warning for future curators]. Just as a like a heads up like, hey, this deposit, the codebook, included graphic images or something like that, like, just to kind of internally keep a check out for each other. Um, but yeah, I feel like we have been having some discussions about what we might need to do externally as well, if that's appropriate for us and, and what would that look like?

**Sara Mannheimer** 38:40

Yeah. Interesting. All right. Well, anything else?

**DC03** 38:51

I can't....

**Sara Mannheimer** 38:52

You've been super helpful so far. So thank you so much.

**DC03** 42:08

I'm glad that I participated and talked with you and it wasn't, you know, it wasn't painful or anything. I'm so used to being on the other end of data. It's not too often that I'm, you know, being asked the questions, I'm usually looking at answers to questions. So.

**Sara Mannheimer** 42:24

Well, it was really fun to hear all of this because you know, working like every day, as a curator, you have a lot of deep knowledge. So it's been really cool to hear that. Thank you.

**DC03** 42:35

Yeah, I'm glad it's been helpful.

**Sara Mannheimer** 42:38

Okay, great. Well, that's all I have. So thank you so much for taking the time. I really appreciate it.

**DC03** 42:45

Yeah. Oh, no, no problem. And if you think of any other questions, feel free to email me or whatever, whatever you need. So then good luck with the rest of your your study.

**Sara Mannheimer** 42:54

Thank you so much. Have a great day.

**DC03** 42:56

Thank you. Bye.