DC06\_transcript\_deidentified

**SUMMARY KEYWORDS**

data, project, curate, data curation, question, survey, coding, context, responses, sharing, curator, funder, lab, study, people, respondents, phase, collected

**SPEAKERS**

Sara Mannheimer, DC06

**DC06** 00:00

I have here interview guide up. And now I feel like I have enough a background.

**Sara Mannheimer** 00:09

Okay, perfect. Well, let me first things first, let me send you the consent form. And we can review it together, you should receive it in your nd.edu email in a moment. Have you received it?

**DC06** 00:51

Got it.

**Sara Mannheimer** 00:53

So basically, the consent form says that this, it describes the research as we just discussed, and then it says you can stop at any time. And that we're just going to talk through your experience curating qualitative or big social data. And there aren't any foreseen risks, but we are going to talk about your workflows and your work and your processes. So if you, that is potentially sensitive. There's no declared funding, you can decline to participate, participate, or you can stop at any time. I'm going to record the interviews, I'll store the data securely during collections at, I will publish excerpts from interviews and, if you consent to data sharing, I'm going to try to publish full interview transcripts in a data repository. So I'll work with probably qualitative data repository to publish the transcript.

**DC06** 02:08

Cool. Today is April 27. Trying to make sure I'm doing the DocuSign thing right.

**Sara Mannheimer** 03:27

Are you having problems?

**DC06** 04:04

I think I got it.

**Sara Mannheimer** 04:06

Okay.

**DC06** 04:23

There we go.

**Sara Mannheimer** 04:30

Nice.

**DC06** 04:33

Okay, hooray. All right. I think that's done.

**Sara Mannheimer** 04:37

Perfect. Thank you. Alright, so we can get going. I, so we've discussed my research, and we signed the consent form. So I'm going to have, I've identified through lit review six overlapping issues that I see in both qualitative data reuse and big social data. So the interview is structured around these six issues. And so we'll have one introductory question six questions, one on each of the issues and then a wrap up question. So the first question is, tell me about the types of data you usually curate, and then what your interests are regarding data curation.

**DC06** 05:27

The types of data I usually curate and range from simple survey data that we collect in the context of grant funded projects, typically, as described in our data management plans and IRB documentation, in such a way that the data curator knows what the guiding principles will be for the elected data's sharing and reuse. Okay, and typically guided by the funder and the research project's timeline, or iterative release or iterative sharing with their constituents, respondents, research communities and through their publications and data share or data release. The other kind of data curation I do, which is more common for me, is data curation in the context of [a specific disciplinary field] research projects and their aims often data that has a computational component, either data that's curated for the purposes of use in the context of input files for computational models. These might be geospatial input files, these might be demographic input files, these might be survey input files, often [specific topic] survey files, which include both qualitative and quantitative information regarding [topic redacted], demographics, and other coded values. And the other form of data curation that takes place is when we look at the output from those models. So typically, when we do computationally enabled research, we will run the model many, many times. First, for fitting and later for analysis. And in those contexts, sometimes the data we use for fitting is treated in a different way than the data that is output for analysis. So we often have two phases of data curation treatment for output data of computational projects. In a nutshell, those are the three lines of data written broadly that I would curate data files for computational input, data, data that informs a model. This is often multiple files from multiple different sources used to parameterize different aspects of models. Then the model runs, data is output. If that data and those runs are relevant to the fitting phase of the model development they're created in a different way than if they are the files that are output for analysis during a later phase of the research. Umm, those are some of the most common ways I find myself curating data for my own projects and in the context, of my own work as a researcher rather than a curator, some of my responses might be a bit different. But in general within my role is as [job title] at [university unit]. When I consult with people on their research data management needs, it's typically at those phases of work proposal development, data management plan offering input on documentation for IRB then collection, via the research project's stated goals, and then data curation to either run models, or support data analysis. So I would say, um within the context of my work, that's pretty typical. A word for the wise about the way I curate data that may differ from some of the others that you interview. One of the things that funders pressure data curators to do, that PIs pressure data curators to do, and that projects with big goals pressure curators to do, is to groom the output of data models, so that it can be reused as input for other models in other contexts, sometimes, this is used for forecasting. And sometimes this is used to because there's no way to generate the data faster than the computational method, it would take too long to collect it on the ground. So we use the model to generate the data to get to the next step, rather than doing a survey or doing field observation. Um, so we could think of those as simulated data sets that take the place of observed or collected data sets. That's sort of the the oddball sort of data that I curate and the way people use it.

**Sara Mannheimer** 12:24

Okay. Ah, very interesting. And so you're mostly, you're embedded... you're working alongside these researchers as they create the models, and it's kind of like curation at each stage?

**DC06** 12:43

Yeah, quite frequently. It depends on the project and its context. But I do think that embedded data curation is for the projects that bake it into their budgets, and bake it into the way that they put forward their approach to data curation and management over the life of their projects. They get much more replicable research much more reusable data, and much more rapidly publishable data.

**Sara Mannheimer** 13:18

Yeah.

**DC06** 13:18

Out of their projects, it's easier for them to present their work to their communities whether that be at a disciplinary conference, or in a paper or a poster... than it is for those who have to do those curation steps ex post facto.

**Sara Mannheimer** 13:38

Alright. Yeah. Okay. Great. Okay, next step is to identify a specific example that we'll use pretty much for the rest of the interview. And you can draw from other examples too. But it's helpful to have one specific time so that we have more details and specifics, so can you describe a recent time when you either curated big social data, so that's like social media data? Or qualitative data for publication and sharing, or if you observed someone else firsthand, but it seems like you'll have experience yourself.

**DC06** 14:27

Um, I have not curated what you describe as big social media data, for any project. That doesn't mean that others [at my institution], particularly or [a librarian] who works with our social science researchers and economics graduate students more might do or [social science researchers] might do. But let me explain why. Part of it has to do with the licensing for social science datasets, and the way our university negotiates those and the way our researchers get access to that data. Right now, our researchers, for example, to use the Twitter data set, have to pay to play and because it's essentially a vended data set accessed via API to, again, write it broadly. We don't curate it for them, the vendor does. The researcher typically in those cases, for example, the Twitter case, and I have done several consults on this. They need our advice on how to get the part they want out of the bigger data set.

**Sara Mannheimer** 16:22

Right.

**DC06** 16:22

And then what to do with that. So we do not curate the data so to speak, it's more like we teach our researchers how to use programmatic tools to isolate the portions of the data set they need for their research. So they're almost always doing analysis and text processing from the get go. Yeah, that's an example. But I just wanted to put that forward. And then to give you an example of another kind of data set that I have consulted on extensively, this is an interesting project, and I'll share a couple of others. This one is [a lab that studies health and human behavior]. And this project has been iterative over many years, including both qualitative and quantitative data set accumulation, and coding. So when we support this lab to curate its data, part of what we're doing is supporting them to curate the data in order to fulfill their grant funded mandates, who in this case [a federal funder], we're also helping them to curate their data, so it supports the kind of publications they need and the next phases of their work. The most famous portion of this data is [a longitudinal study]. This study covered [about a decade], with survey data collected in [a number of] waves once every [number of] years beginning in [year]. And then participants were reassessed using similar measures [periodically]. In addition, there was another data collection effort of qualitative diaries that were collected. They were concurrent data collections with the other studies but managed separately. And in total, this project took in data from [hundreds of] people in [an area of the United States] and was iterative in nature. And it's one of the most extraordinary studies of its kind that I have been consulted on and one I have a lot of admiration for. The way the project continues to evolve is wonderful. And the way they do their data collections at [periodic] intervals which you can see on their current study data collection page describes a little bit the challenges of data curation and the kind of individuals over time who participate in these bigger projects. For this one, the team in the lab add to data curation and retention mandates, and data curation tasks related to putting forward the information collected for coding and a lot of curation tasks related to confirmational coding. So when more than one person would code so that you know that the data is coded accurately, think of those as task oriented curation activities related to a large data set, you might think that [hundreds of] respondents doesn't amount to a lot of data. But if you do it over [decades] with this many questions, and this many repeats, the data set grows in extraordinary fashion, and many, many of these collection efforts also coincided with all kinds of other concurrent studies. So it's a really fascinating one that myself and [my colleagues] have been privileged to consult with over the years. And [the PI] is terrific to work with because [they themself have] served on our IRB. So [they are] a senior researcher not only within this project, but also within the university context of [their] service roles to the needs of others doing similar research that made [them] also extraordinary to work with so I like to use this project of [theirs], as, uh, if you can do it the best way for the longest time, how might you do it?

**Sara Mannheimer** 16:48

Yeah.

**DC06** 17:02

With the caveat that this project also shows us because it's taken place over such a long period of time how funder mandates can change over time and therefore really impact what your mandate for sharing is and your readiness to share. When [the PI] first conceived this project, [they] did not have a sharing obligation, in so much as [they] had a publication obligation to share the results of [their] research for the betterment and knowledge of all, not necessarily to share [their] data for reuse.

**Sara Mannheimer** 23:15

Right.

**DC06** 23:16

So the way that [they release] the data for each cohort now is optimized against today's data sharing mandates, retrospectively sharing as much as possible at the right release moment. But the project also illustrates the tension and the unfunded burden for researchers when data they've collected in the past was not funded or planned for today's release conventions.

**Sara Mannheimer** 23:50

Yeah. Very interesting. Okay. Um, and do you have, was there, is there qualitative...Like, can we narrow this to just the diary study, maybe like, did you help to curate that where...? Because I do, I'm like specifically wanting to talk about qualitative data.

**DC06** 24:16

The short answer is no, the lab curates it, we help with it from the point of view of sharing and we share the coded anonymized data. Only think of us as part of the pipeline.

**Sara Mannheimer** 24:36

Okay.

**DC06** 24:37

Where for us, the data is essentially blinded. I mean, we know what cabinets the other information is stored in and how to hook them back up to de-anonymize them, but we never would have access to do so.

**Sara Mannheimer** 24:54

Okay.

**DC06** 24:54

If that makes sense. So we understand each phase's purpose within the data collection workflow, is the best way to put that I think.

**Sara Mannheimer** 25:08

Okay. And where has has the data... So do they have new data management plans as the funder mandates change over time? Yes. Yeah.

**DC06** 25:22

Yeah. They have to.

**Sara Mannheimer** 25:23

Yeah. And then, uh...

**DC06** 25:26

During their reviews, too.

**Sara Mannheimer** 25:28

Okay. And then is the data pub-, published somewhere like in a data repository?

**DC06** 25:36

Yes, iteratively.

**Sara Mannheimer** 25:39

Okay.

**DC06** 25:40

And the other thing I would point out related to this, the lab, is now well they've stayed connected to their survey group, or their respondents group and how those respondents have sharing information back to them is part of how they communicate their results. I can't go into that in detail. [The PI] does that better than I could. But just to say or add that, I think it is really important when we collect qualitative information. And I give you an example from my own research in a second that the value comes back to the respondents in that they know when to expect that value.

**Sara Mannheimer** 26:45

Great. Right. Yeah. Great. Okay. So I'm going to start going through using this example. And I sent you, I just reupped it in your email that interview guide in case you want to follow along.

**DC06** 27:10

I have it open.

**Sara Mannheimer** 27:13

So we have these six areas. And I do see we're kind of low on time. Do you have to leave right at 11:50?

**DC06** 27:19

Let me double check. No, I don't. I'm good until... I'm good until the top of the hour.

**Sara Mannheimer** 27:28

Okay, cool, that I think that will be helpful. So our first question has to do with context. So I have this quote at the beginning of my question. "Qualitative research is a process that may include deep and prolonged contact and connection with research subjects, attempting to understand the subjects within their own context. And so qualitative data are highly context dependent. And context is a source of data meaning and understanding nor in context, under using it or not recognizing one's own context or event perspective will result in incomplete or missed meaning, and a misunderstanding of human phenomena." So that's where this question sort of is based. And so during this example, with the [longitudinal] study, what challenges did you encounter, if any, when you're trying to communicate or capture the context in which the data was collected during your data curation activities? And then what strategies did you use to address those challenges?

**DC06** 28:29

Um, the great news is for the [longitudinal] study, and not to trivialize the question, but the project is big enough that our role as [data curators], is entirely separate from the data capture, and the communication with the respondents. So the project's big enough, and with permanent staff and data collections phase-specific staff in such a way that the communication across the longitudinal arms or efforts of the study at their yearly and even less frequent intervals, and is always done by the PI and a member of the staff. But as I'm sure you are familiar with in academia, especially with qualitative work like this, the chance that the staff, lab person is the same at year one and year seven is low.

**Sara Mannheimer** 29:40

Yeah.

**DC06** 29:41

Let alone at year nine and so on. And so because this study has been repeated so many times with the same respondents, the person in that role will change. So I'd say one of the challenges is staff continuity, which is the this case is the burden of the PI, not the data curator.

**Sara Mannheimer** 30:06

What about context in terms of like, who the people are where, you know, when you're collecting data from these people like, what kind of contextual information do you include? Like as metadata? You know, you said they're from [a region in the U.S.]. So maybe that's an important contextual clue. We know that they're of certain age groups and talking about [the topic of the study]. Is there anything else that you put in during your curation process to sort of support this idea of like, keeping the data in context as much as possible?

**DC06** 30:40

That's a good question. And there are a couple of things. We might think of the survey questions. And the diary metadata. Even if we focus just on the diaries. One of the things we know to keep those in context is what other arms or response efforts of the project they have participated in the past. So we know how many of those surveys they've answered before, we know how complete their responses were. That's the whole point of this survey.

**Sara Mannheimer** 31:36

Right, right.

**DC06** 31:37

We know [details about them related to the research topic]. So there is a lot of metadata associated with their context within the project as a whole. And the purpose of doing the diary data collection.

**Sara Mannheimer** 32:15

So I'm sorry, once again. So I guess the longitudinal nature of the study has provided a lot of context in itself, you're getting a lot of info from these?

**DC06** 32:31

Yeah, and it's important not to detach each effort's input from the rest, otherwise, it has no value.

**Sara Mannheimer** 32:42

Right.

**DC06** 32:43

Um, and the qualitative responses, and the kind of things people record and share about their experience, provide what you could think of as additional context in and of themselves.

**Sara Mannheimer** 33:02

Right.

**DC06** 33:02

Related to the more formulaic questions they might encounter in the annual surveys.

**Sara Mannheimer** 33:09

That's really good.

**DC06** 33:10

Yeah, this is really, yeah, a really beautifully constructed project. I love it. And I can take no credit for it.

**Sara Mannheimer** 33:19

I'm going to skip to this data comparability question number three, because I feel like this is related. If you're doing surveys, you're doing diary studies, like, how do you, what kind of thinking have you done about comparability or interoperability of these data, different data sets you have with others, like either within the project or moving to other types of [related] studies that you might want to connect with this project?

**DC06** 33:47

That's a great question and I'll leave the connection with other studies response firmly in [the PI's] court and say that that's a question for the PI rather than the curator. But it does help the curator to know out of the gate so to speak and over time, what those ambitions are, and what the events are. I'd like to have those conversations related to event specific data releases. For example, when will you be publishing this? When will you first give a presentation about this? What's your funder deadline for sharing? We might think of those as very pragmatic questions to ask relative to this kind of data curation. But what you'll find is that they hired PIs to do iterative releases that help them understand which audiences experience their analysis or their lensed view of the data at which points because everybody's really getting incremental views of the data collected over time. And everybody's getting just snapshots of that, slices if you will.

**Sara Mannheimer** 35:09

Yeah.

**DC06** 35:11

And that means that you have to pay attention to what we might think of as best practices for social science data collection, broadly. Are our date formats, the same? Are our, is our blinding mechanism the same? Is our blinding good enough? Is our, do we have confidence in our coding? Did we keep the data dictionary the same for the coding? Or has it changed over time? If it's changed over time, why was that to help the coders or to help those who would interpret the data later during analysis? And so on.

**Sara Mannheimer** 35:50

Yeah.

**DC06** 35:52

So codebooks become very important. And I consider them an element of the data we curate. And an essential element is the curation of codebooks, evolving codebooks.

**Sara Mannheimer** 36:04

Okay.

**DC06** 36:04

Version codebooks.

**Sara Mannheimer** 36:06

I feel like this is related to question two about data quality, too. It's like how high... you know, what is your confidence in the coding? Do you have other questions that you asked them to get at data quality for these data in order to help them be reusable in the future?

**DC06** 36:33

That's a good question. And sometimes for us, that comes from understanding whether the thing they want to make reusable is their coded data, or their analysis of it, and who the audience for that reuse is and what data format they want to receive the data in. That's challenging for a study like this, because those analysis programs change over time in the formats they will accept. So keeping those formats forward migratable. And the clients of data fields that are, were data types that are embedded in those shared sets, also becomes important. At the lab, those are handled with a lot of due diligence. So although we could think of this project, ironically, a little bit like satellite data, you send it up with the instrument on board, you know, you can't change the format it's going to send down because the instruments already up there.

**Sara Mannheimer** 37:57

Yeah.

**DC06** 37:58

Um, this is a little bit like that, because the project's so long, but on the flip side of that is that we know how to forward migrate and work with the data formats, going back through the many years worth of data, and the coding is so consistent and so tightly managed, that it's, there're no, in software, we would call it breaking releases, there are no breaking releases in the lab's data collection. Maturity, we could call it a maturity schedule, almost for data collection, I think is a nice way to look at it.

**Sara Mannheimer** 38:43

Great. Let's move on to question four about informed consent. So in this example, what challenges did you encounter, if any, relating to informed consent for participants, particularly consenting to future use of the data? And then what strategies did you use to address those challenges?

**DC06** 39:06

That's a good question. Particularly in this context, and it's one I've helped other researchers who've had a harder time with it than this lab did. What we would call going back for secondary consent, or reconfirming consent. If you have to reconsent someone it's time consuming.

**Sara Mannheimer** 39:33

Yeah.

**DC06** 39:34

And with a group like this in [this] study, it's better for consistencies sake if you never have to reconsent. They do have the wonderful anchors in the calendar for each new interaction with the respondent. You could think of that as a renewed consent, rather than a reconsent. So each phase they're renewing consent. And that's pretty awesome. So that makes this cohort of people in this study an amazing cohort.

**Sara Mannheimer** 40:21

Um, okay. Let's see...All right. So there's... Okay, that makes sense. Great. Let's move on to privacy and confidentiality. Have you encountered any challenges relating to privacy for the people in the data set? And then what strategies do you use to address those challenges?

**DC06** 40:49

There is the challenge of protecting respondent privacy, there is the challenge of protecting or blinding one coder to another when we're doing confirmational coding. And there is the challenge of restricted access to files we have a burden to keep over time. In this project, there's actually a substantial concern for and budgetary need for restricted access to many of the actual survey materials. And the recordings and the transcripts and the survey response forms and the diaries and so on. And this is part of the reason why the project has had a need for permanent staffing during its phases of active data collection and coding. And a reason why staff turnover is really important on a project of this length. So the lab manager for this project, matured into a career with bigger opportunities. So when [they left for a new job], [they] literally had to groom the team to take over at the right phase of the project. And to do that in such a way that the restricted access and privacy protections for the data were maintained during the onboarding of new staff. Yeah. So that continuity of access for the PI has maintained over time within the project. It's a testament to [their] professionalism and to [the PI's] that that transition was successful. So again, it's in this case, it was a lab manager's role to do that part well, rather than the PI's or the data curators, which is both a benefit and as I've remarked on a rare maturity challenge. Yeah, because of course, I would like to keep the same person who does a great job in that role the whole time.

**Sara Mannheimer** 43:50

Yeah.

**DC06** 43:51

The person in that role can grow and change and develop.

**Sara Mannheimer** 43:57

Sure. Yeah. Um, have you done any work relating-- So you said that de-identification happens in the project group as well, rather than the curator phase? And do you like check for that before you help them publish or? And then when you do publish, is there any like public data that has restricted access? So there's like access points that people have to request?

**DC06** 44:23

That's a good point. Um, yes, and yes, and yes. Starting from the last most question, asked access to restricted fields is through PI only and only within the latitude that the IRB documentation allows, access to the--I'm losing my train of thought-- access to the....

**Sara Mannheimer** 45:05

Oh, de-identified?

**DC06** 45:05

De-identified data is released according to the funder's schedule unless a publication or presentation encourages or amplifies the need for earlier release or for partial release. And the identified, the identifying steps, um, are separately managed from the recruitment and renewal steps because these are respondents we do want to keep in touch with. So from that perspective, they're managed in identifiable way by the lab for future communications. That information is just kept separate from their responses, you might think of it as a customer relationship management approach to maintaining the relationship with the subject over time being a separate but concurrent activity alongside the data curation for their responses.

**Sara Mannheimer** 46:19

Okay.

**DC06** 46:21

So since we do revisit these people and re-engage with them on a periodic and anticipated basis, it's important that that from the perspective of data curation that both those activities can happen well, both the blinded coding and the communication renewals.

**Sara Mannheimer** 46:43

Right.

**DC06** 46:44

And continued data collection. So that makes it also I think, a really nifty project for consideration in studies like your own. [The PI] might be a fun person for you to interview too.

**Sara Mannheimer** 46:55

Yeah, I will... I have these two links. So that's a good point. I'll take a look at [them]. That's actually my last question for you. Who else should I... But -- intellectual property. Let's get through this one. And then it's our last main question. So during your example, did you encounter any challenges relating to intellectual property concerns when you're publishing and sharing the data? So like, participant intellectual property or intellectual property in other ways?

**DC06** 47:27

Um, short answer is no. Um, in the case of this project, those responsibilities lie more firmly with the PI than the data curator. Um, does this data sharing fall under Fair Use? Again, no. Um, and I'm actually going to have to check in see what [the PI's] most recent release is because part of our challenge as [data curation team] and supporting many researchers across the campus is deciding when to use [a data repository] and for what releases and when to use other means and how to keep active data management separate from sharing. And when... so I have to double check how [the PI] did [their] last release before I could comment on how her data sharing modality encourages what we might call Open Science rather than Fair Use.

**Sara Mannheimer** 48:36

Yeah, okay.

**DC06** 48:37

Because these responses are not, um, and the analysis of these responses would be [the PI] and [their] lab's intellectual property, but only used under the release of the participants.

**Sara Mannheimer** 48:58

Right. Right. Okay. Great. All right. Well, that is, that is all of our questions. Well, we really ran to the finish line. Were there any other issues that arose that we haven't talked about that you think are really important, in these last three minutes that we have?

**DC06** 49:22

I would say lab and staff, lab manager and staff reticence in the early years to use [a data repository]. Um, not because of any mistrust, um, simply because of the need to decide when and the perception that it would be time consuming, and the notion of whether the older data had been properly funded for that level of attention at the lab side of the coin when it comes time to prepare data for sharing. And again, I'd have to check some of the most recent releases to understand better how this has evolved. But I think for most labs, most of the time when determining how to share this sort of data, you're trying to determine how to share according to the milepost or milestone you've set for sharing, whether that be to meet a funder mandate or to have something ready for publication, or ready for peer review. And whether you have the lab resources to do that to a gold standard, a silver standard or bronze standard, those are just made up gradations for the purposes of us to communicate. But I think those are the challenges of basically where to set the bar. And they're often determined not by money, so much is by time.

**Sara Mannheimer** 51:12

Yeah.

**DC06** 51:13

How labored are staff, who has the time to do this, and to what level of quality can they do it in the time available? And what is the platform they're using for sharing demand of them when it comes up time?

**Sara Mannheimer** 51:26

Great. Awesome. Thank you so much, this was really great.

**DC06** 51:32

Oh, I'm glad you enjoyed it, I sure did.