DC04\_transcript\_deidentified

**SUMMARY KEYWORDS**

data, people, survey, qualitative data, big, project, interview, pieces, metadata, archiving, social, researchers, collected, context, published, research, crowdsourcing, irb, informed consent

**SPEAKERS**

Sara Mannheimer, DC04

**Sara Mannheimer** 00:00

So you received the questions in advance. I'm seeing that qualitative data reuse presents some similar issues and challenges to big social data research and data curation. And I'm trying to sort of bring these two communities together to see what they can learn from one another and what data curation practices we can use to support both types of research and data sharing. So that's that whole first page. And so I've done a literature review and identified six key issues that are in common between big social data and qualitative data. And I've structured the interview around those six key issues. So we have six main questions. And I think the interview should take about an hour up to 75 minutes.

**DC04** 01:10

Okay, sounds good.

**Sara Mannheimer** 01:11

So first question, tell me about, about you the types of data you usually curate, and then just what your interests are, and your background is regarding data curation?

**DC04** 01:22

Sure. Um, so big step back... background. My training is primarily in sociology. I did a PhD in sociology at [university], focused on quantitative data analysis and social theory, networks. From there, I came to work at [university library] as [specific title redacted; the position consists of two roles relating to data librarianship and data education]. Those are two very big hats, but people in libraries tend to fill a lot. And so the two pieces of my job in general, I help people when they have questions about data [having] to do with humans, whether that's survey data, or interview data, or digital humanities, or Facebook, or whatever it may be. And that's anywhere along the research cycle from planning data collection, all the way through to analysis and visualization and publishing, curating the data and so on. The other half, a lot of my time is spent both teaching workshops and guest lectures, but also helping build a program throughout the libraries that sort of covers, not necessarily everything in data science, but the pieces that are missing, that we have a need for in our university that, say, the statistics department isn't covering or the Teaching and Learning with online tools people aren't covering. Um, as far as data curation, specifically, I intersect with [curation] a lot because of my social science role. And because a lot of people are collecting and/or creatively reusing human subjects data in one form or another. Surveys and qualitative interviews and things being the most common. So I am our representative to three social science [data] repositories, [repository names redacted]. [My duties as a representative] include helping people find and use data, but also helping people with deposit process and especially with [data repository] because it's brand new. We've really been trying to promote people thinking ahead when they're doing a data management plan or whether when they're coming up with their thesis or dissertation research, if it's not funded, they don't have a DMP than thinking ahead to building an informed consent that allows them to share the data, helping them sort of feel comfortable stepping into the process. That you know, there's people who can help you make sure that it's de-identified properly that you can do this and that there's value in it besides besides just having done a thing.

**Sara Mannheimer** 05:00

Great.

**DC04** 05:02

So that's my biggest role. There's a couple, I will say, besides, there is a big social piece too. It's less common for people who are working with that data to be archiving it because of issues with like, what do you do with Facebook data? If you share other people's Facebook data, you know, that whole thing.

**Sara Mannheimer** 05:28

Yeah.

**DC04** 05:29

Becomes complex. Um, so I do more helping people collect and clean and work with that data, then archiving it, but it comes up occasionally. And then there's one big grant project from one of our history professors called [project name]. That is, a series of surveys were done of [number] of U.S. service members during [historical time period]. The survey data was published in a crappy difficult to use form through [repository]. The but there was like an open question for like anything else you want to tell us at the end of the survey, and all that data was preserved in [an archive], but nothing was ever done with it. And it's just this treasure trove. So now, the project is to clean up and improve the survey data while extracting through crowdsourcing all the qualitative data that was attached to that. We don't know which [survey respondent] it was attached to necessarily, we can't connect the survey with the free response generally. But now we have this database with all of that text now available that can be searched, that can be downloaded. And so, or will be so it can be searched and downloaded. And all of that we have all the data now. It's just a matter of preparing it. And getting it both, it'll probably be both on our website and in our repository.

**Sara Mannheimer** 07:04

Very interesting. That's cool to be able to extract good data afterwards. And make it more usable.

**DC04** 07:11

Yeah, it was it. I mean, it took a lot of steps. Like they went into [the archive] and scanned things, and then they put it on a crowdsourcing site. So we had volunteer transcribers go through each thing three times, and then somebody else wrote a script, clean that up and somewhere down the line, I come along and help.

**Sara Mannheimer** 07:32

Wow. Very cool. Okay, I really appreciate you joining me. So for the rest of the interview, I asked you to identify a specific example when you supported curation for either big social data, or qualitative data. But since you've done a little of both, I think we might sort of talk about both kinds if we have time. Did you come up with something that you were kind of thinking you might talk about today?

**DC04** 08:07

I mean,

**Sara Mannheimer** 08:12

The goal is just to get this...

**DC04** 08:14

But yeah, I think I mean, I think the [survey of U.S. service members] is a good example. On the qualitative side, the big social is a little harder, just as I said, I don't do a lot of helping people curate and archive that, because they're usually not enthusiastic about curating and archiving and all the pieces that belongs.

**Sara Mannheimer** 08:34

When you work with them, though, do you like talk about privacy? And, you know, some of the bigger issues around big social data? Or is it more of a technical...?

**DC04** 08:48

No, there's definitely there's privacy. And actually, usually the start is either a technical "How the heck do I do this?" Or "Am I allowed to do this?"

**Sara Mannheimer** 09:00

Right.

**DC04** 09:01

And often, it's the first leading into the second. Just yesterday, we had an inquiry, like, "I want to do a sentiment analysis on [thousands of newspaper] articles from the [name redacted] database. I see they have an API, can you help me?" Well, no, I can't, because we're not legally allowed to do that with our agreement. But if you have a few thousand dollars and would like to share it with them, I'm sure they'll help you. You know, but when they can get that data within the permissions and things and we can set that up. There's definitely discussion about that. And part of it, what's publicly available usually, unless you're like, scraping from the logged in account, or, you know, is usually fairly public. You know, Facebook's API only allows public pages, not personal pages. Umm, Twitter everything's public already anyway. So we want to encourage people not to try and connect the dots, you know, we want, you know, we're trying to look at big picture stuff. We're not trying to out people or whatever. But yeah, and a lot of times, a lot of the people that come to me working on this are very sort of keyed into those issues already, which is great. Um, maybe I did, maybe it's because I don't get the computer science tests, they go to our engineering data consultant. He has to deal with them.

**Sara Mannheimer** 10:40

Great. Okay. So let's talk through your example, the main point of using one example is to use this strategy called critical incident technique. W want to use a recent example so you can call it to mind and you have some specifics in mind. So we don't always have to stick exactly to your example. But it's a good frame for the interview. Okay, so is, you said that the example is part of a grant funded project, the [survey of U.S. service members]?

**DC04** 11:13

Yes.

**Sara Mannheimer** 11:15

And did they have like a data management plan or some kind of requirements regarding their data with the grant, do you know?

**DC04** 11:22

Yeah, hold on one sec. Yeah, so I did not come in at the very beginning of first grant, this is the second cycle of [funder] grants that's funded this project. So we're like, four, four years into a six year project-ish. But, um, part of the goal of the project was always to make the data available, and it was already part of [the archive] and in the public domain. So there weren't privacy issues there. But the, you know, the what we were going to do with the data to make sure that we were using this money to add value and make it public, especially since part of the data we're working with, is already published in [data repository]. So you know, subscribers can get to that. You know, so what's the value add? How are we making this accessible and useful? Is the other piece that maybe, I'm not sure that would fall under the data management plan piece of the application. But it became a big part of it, things like creating learning objects, and workshop plans and things like that, where people from different public and researcher groups can see how to take advantage of this. And the website piece will be very much sort of public-oriented, play around and explore and see, you know, what did U.S. soldiers say about race, or whatever. As opposed to, or the food, you know, stuff that anybody, like a high school class could easily come in and do cool stuff with. But then we're also doing the archiving the data so that it's useful for researchers as well. From a more social science perspective. Yeah.

**Sara Mannheimer** 13:34

Great. Okay.

**DC04** 13:39

Feel free to cut me off anytime I'm like...

**Sara Mannheimer** 13:41

No, no. This is all really, really useful. Okay, I'm going to skip down to context. So for qualitative data, just to give you some background, I have this little paragraph there. "Qualitative research is a process that may include deep and prolonged contact and connection with research subjects attempting to understand the research subjects within their own context. And qualitative data is therefore highly context dependent. Context is a source of data meaning and understanding ignoring context, under using it or not recognizing one's own context driven perspective will result in incomplete or missed meaning, and a misunderstanding of human phenomena." So that's sort of like where I'm coming from in terms of context when I ask these questions.

**DC04** 14:29

Yeah.

**Sara Mannheimer** 14:29

So during this example, what challenges did you encounter, if any, when you were trying to capture the context in which the data was collected?

**DC04** 14:40

So we definitely did. They may be a little different than the usual ones when you're doing primary collection, right, because we have existing secondary data, but we don't have, except in a very limited number of cases, we don't have any way to link the qualitative data to the context of the survey, which gives us a lot more information about the people—everything from their rank and age and, you know, state of origin, what branch of the service they're in, all of that unless they put it in the text. Some pages, we have maybe one question, that's a closed-ended question to connect to it. And that's it. So we did, we've tried to do what we can to, you know, use a robust set of tagging, and searchability, within the transcripts so that when people are looking at it online, you know, we might be able to say, look at all the ones that concern race, and things like that. And in the couple surveys where we have a better shot, one of the other people in the project has tried to connect back to the survey as much as we can. So there's something there that a researcher could dig into, it's not going to be an immediate part, because it's not universal, it's not going to be front and center in the website. But it's really an interesting, so all we really have to go on is the text for the most part. And then also tracking. What we do have is we know when the surveys were taken, and sometimes like a specific branch that this one was given to, there's like a hundred different surveys or something ridiculous. This one was given to people in a specific branch in this time period, or this one concerns something specific. And so we try to keep that up, since I'm the one working with sort of building the metadata scheme for the qualitative data deposits, at least, um, try and keep that as much as we can have that attached together so that people can see what this is from.

**Sara Mannheimer** 17:04

Nice. Um, have you talked to anyone who has tried to use this data in the past from earlier stages of the research? Or is this kind of, the project still in progress? Like it hasn't gone live yet?

**DC04** 17:18

One of the people on the project has used the quantitative pieces of this data before. So you know, he's sort of brought a user perspective into a lot of how do we deal with survey data. As far as I know, I mean, other than sort of the initial, like, you know, we we've had, like, I think that the lead the PI did publish something on the process that brought out some of the early findings of like, interesting themes they saw coming up, just qualitatively, looking at what was, what was transcribed at the time. But there's no structure. Anything so far, this is more backwards from a traditional project in the sense that the front end of the project is preparing and publishing the data so that in the future, people can use it, as opposed to the front end being collecting it so that I can use it and then publishing it as an afterthought.

**Sara Mannheimer** 18:24

Right. Okay. So just to review, you're trying to link back to the survey, when you can, you're putting in a lot of metadata to help with people understanding the difference.

**DC04** 18:43

Yeah, whatever we have on particular...

**Sara Mannheimer** 18:47

And is there any demographic information on the the qualitative responses? Are there just sort of, if you can...

**DC04** 18:54

No, not really, I mean, the, it's not unusual for people to volunteer that, right? Like, a lot of times, you'll see people complaining about being mistreated because of their race, as an example. Or being mistreated by the officers, or whatever. So you can get bits and pieces, but there's nothing consistent attached to it.

**Sara Mannheimer** 19:19

And when you make the metadata do you pull, would you pull that [demographic information] out? Like, do you have specific types of metadata that you're looking for?

**DC04** 19:27

All that we would pull out for that would... that would show up in the tags. Probably.

**Sara Mannheimer** 19:33

And those are crowd sourced tags, you say?

**DC04** 19:36

Um, the tags are not crowd sourced. They were more, um, I think there were some TAs or researchers connected to the project that did those. The actual transcription was crowd sourced.

**Sara Mannheimer** 19:54

Okay. Got it. Great. I guess I do want to ask you this question 1a, like, when you're helping researchers with big social research. Do you talk about context, when, with them, or have you ever?

**DC04** 20:20

Yeah. So there are some projects that there has been some talk about curation that are sort of borderline on big social. So for example, social book reviews on Goodreads, or, or beer and wine reviews on websites. We, our, me and our engineering data consultant have worked with people from food science on like, food and wine and beer reviews and sentiment analysis stuff and things like that. So there's definitely conversation about, you know, how do you preserve? How do you not only preserve but notice in your research, the particular context, so the social book reviews, the grad student that was working on that, the, the fact that some of the books re-, released much earlier than others, and had more time to be reviewed, impacts how you treat the data, if you're turning that qualitative data into quantitative things, or these were nominated for awards or something, and there were certain, you know, if you're looking at things over time, then there's trends that would happen because of the award nomination itself, for example, and thinking about or like, it came up when one person responds to another review, like, "I've seen, blah, blah, blah." How do you deal with that? So I'm not sure how much that has impacted the thoughts about that curation, other than to say that, you know, like, we would definitely encourage them to put, you know, when were these data from, if possible, attached each post rather than the whole big picture, but also when were they collected? How were they collected? So we know, you know, is it likely to be a full sample? Or is it....

**Sara Mannheimer** 22:25

Right.

**DC04** 22:26

Is it going to be cut off because of the limitations of the platform and so on? And some of those are fairly similar, the big, the biggest difference is that with the qualitative data, except for this one project, with the qualitative data, a lot more of the conversation revolves around sort of setting things up for success in the very beginning. Right, working with the IRB and the informed consent to make sure that data sharing this allowed for and getting over the hesitancy about that, you know people, I found people are actually more, people are more hesitant to share other people's interviews or even surveys than, than putting their life story out on Facebook or something. Whether that's because they feel like "Well, you already, you should assume the internet owns it."

**Sara Mannheimer** 23:30

Yeah, interesting. Do you think that has to do with our overall culture or like, would you...?

**DC04** 23:40

I almost feel like people are afraid, a lot of people have this apprehension about IRB and human subjects stuff. Like it's like this Boogeyman, which is extra funny because our IRB here has been smooth sailing compared to past ones I've worked with, but, um, and I think that people often sort of de-, dehumanize the big social data. So even though it's not aggregate, it's all this individual stuff that's often very personal, right? Um, it doesn't feel as personal as actually asking someone to take a survey or do an interview for you. Um, so I think those two pieces maybe contribute to the feeling that way, along with who the users often are. The people we work with with the big data are usually data scientists, computer scientists, engineers, people who think in big boxes and mechanisms and are, um, taught less to be attuned to, um, to active--, to proactively think about the human consequences. Right? It's sort of this emerging field of ethical data science beyond just like, what do you do to comply with things? But actually, like, how does the way we work with big data impact different people groups disproportionately?

**Sara Mannheimer** 25:31

Yeah.

**DC04** 25:32

As opposed to not just the individuals.

**Sara Mannheimer** 25:34

Right. Mm hmm.

**DC04** 25:36

But that's still not a big part of it. You know, it's not the big conversation that's going on in computer science, that's conversations going on with the, Philosophy of Thought people.

**Sara Mannheimer** 25:51

Right. Yeah. I am hoping that's part of what will come out of this, is like connecting those computer scientists, with people who have been thinking about this for some time and helping the two sort of come together. So um, okay, we might have to, like, step it up a little to finish on time, but we'll see. We're on question two out of six. So during your example, what challenges did you encounter if any, when you were trying to document data quality? Like was there missing data or any issues with the quality of the method that you identified that you needed to document for future users? And if so how did you do it?

**DC04** 26:35

Yeah, so a couple things. One, we had to make a decision fairly early on about how to deal with if there's disagreement in the transcriptions. So three people transcribed each page. Um, and so I think, honestly, I don't remember where the group settled on that, in the end, I wasn't handling that piece. But there was discussion over, you know, data preservation, getting the most possible things versus making it cleaner and easier to use. I know that for the website, at least, you know, we needed to settle on one. But whether that was done by hand, or by going back and looking more or what, I don't know. The other, as far as data quality, there are plenty of space, there's there's things that are not legible. And the qualitative data, there's pieces of the data files from the surveys that are just not, clearly not accurate. And we tried to identify those and tag them. And I think we sort of settled on a side of keeping intact as much as we could, but adding annotations, that probably will show up, like if you try and access it on the website, if you're looking at that particular one, you'll see that, or if you were to download that particular survey data file, say, or whatever, where it would say, you know, some responses were skipped on this because they didn't fall in the categories given or whatever.

**Sara Mannheimer**

Okay, and you have those annotations on each one?

**DC04**

On each transcript? I wish I could... So the transcripts I'm trying to think, so on the transcripts, I think the most we're going to get there probably is just there may be chunks where it's like in square brackets like illegible or something. I'm just leaving that and we didn't, we didn't correct spelling. And we didn't you know, we tried we tried to leave the original stuff intact even if it was words that would be offensive, or even if there's spelling [issues]... when we could.

**Sara Mannheimer** 29:21

Okay, great. Um, and I guess I also I'm, do you see similarities between the these strategies you used to like support, document and quality for the [survey of U.S. service members], and some of the big data stuff?

**DC04** 29:44

I mean, the big data stuff. There are some times just because you're at scale. There are some you know, sort of automated or crowdsourced methods you can use as well. I'm, um, I've done research using Mechanical Turk that, you know, that's a big piece of like, how do you ensure that you're getting good quality there. But a lot of it at the end of the day too, is just digging into the data, getting to know it, looking for things that are, looking for things where it's not immediately clear what you should do. And at least trying to note them.

**Sara Mannheimer** 30:30

I feel like these two examples you have are similar. Other people I've talked to actually have a researcher, like, who conducted the interviews available to them, you know. Whereas because you're in this historical mode, you can't speak to like the soldiers, or the...

**DC04** 30:47

Right. You can't...

**Sara Mannheimer** 30:48

Researcher who originally conducted the survey.

**DC04** 30:50

Yeah, you can't go back and do that. Yeah, that'd be fantastic. That would be a great study.

**Sara Mannheimer** 31:00

Okay, on to data comparability. And so what challenges did you encounter during the [project looking at surveys of U.S. service members] where you were trying to... um, relating to comparability or interoperability of your data set?

**DC04** 31:14

I mean, so besides internal interoperability issues of matching the surveys and the qualitative data, there's, it's really hard. And I'll just say, up front, right now, this holds for big data too, to choose formats for sharing data, that the whole range of potential users are going to be familiar with that work well, that can't scale to just, you know, being archived and downloaded and things like that. [The data repository], had the [survey] data in ASCII format, originally, and it was in like two separate ASCII files for each survey that had to be matched up. So it's this huge ordeal, but like, we're still working out with the qualitative data, say, for the [project] on the archiving side, what that will look like, we'll probably have a separate file for each verbatim, which will be fairly small, and then they'll be archived in a big package or whatever, or a number of big packages. But sort of, and document in the metadata along with those. So I think we settled sort of on like, a separate... So each file, each verbatim would also have, for each set of verbatims, from a survey maybe would have a metadata file that describes the survey and what was in it, how many files were in it, and so on.

**Sara Mannheimer** 32:56

Can you tell me what a verbatim is?

**DC04** 32:58

Like, so just like one trans-, the transcription from one person's free text responses. So it's either one or two pages of free text responses. Sometimes they they continued on. But, you know, we're not talking a 30 minute interview, we're talking, you know, a half page of scroll.

**Sara Mannheimer** 33:23

Yeah, yeah. How about with metadata? Did you encounter, so like as thinking about potentially, your these transcripts being used in the future with like, other surveys of soldiers, you know, like, what have you have you thought about, like, how to make sure that like how to encourage interoperability or comparability in the future?

**DC04** 33:49

I mean, we're keeping things in plain text readable formats, whether it's a whether it's a TXT file, or like a structured XML, or JSON file or something like that. But not something locked down to a statistical package. And if we do, if we do put pieces in a particular format, and providing tools to shim, so you can get to something else. So like, if we have them all saved as JSON files, we're gonna have to do some training and maybe have a little Python script or something that can self execute that you can run to take all of those that are in a directory and turn them into text documents that Joe Schmoe on his computer can read, without having to be a computer scientist.

**Sara Mannheimer** 34:52

Right. Okay, cool. So you said that you see similar issues with big social data, and comparability? I think that's good. All right, on to informed consent. So in your example, did you encounter any challenges related to informed consent for the participants, particularly consenting to future use? And then what strategies did you use to address those challenges?

**DC04** 35:27

So this is easy. In my example, we can't get informed consent. Many of them are dead. The ones that aren't, we have absolutely no way to connect anything. And it was already in the public domain, in the [archive], so that piece wasn't an issue. Um...

**Sara Mannheimer** 35:49

What... do you know what they consented to at the beginning, it was just like a normal survey consent?

**DC04** 35:55

I don't know. That piece was all managed before I got there, I don't think I ever saw any forms.

**Sara Mannheimer** 36:36

Okay, so you're just sort of making, it's in the public domain. So you're saying there's...

**DC04** 36:42

Yeah.

**Sara Mannheimer** 36:43

This isn't like an issue for your project?

**DC04** 36:45

Yeah, it's not an issue for this project. With the big social stuff, you know, it's basically just that the limits are basically following the terms and conditions and using publicly available data, data that people have put intentionally on the internet in a way that anyone can find it without a login without credentials, anything. And that's about as far as those projects, I've been involved in have gotten. It does get a little gnarlier. Like there was some discussion, somebody was doing research on, comm-, looking at the comment threads on posts about [topic] on [television network]'s Facebook page. Right. So the page is public, therefore, everything that's posted on it is public, but not all of those people necessarily, were quite as intentional about what public means compared to if they put it on their own profile page, and they choose [to make the post visible to either] friends or the public, or something like that.

**Sara Mannheimer** 37:58

So what what did you come up with when you were discussing with those researchers?

**DC04** 38:05

I mean, I think they settled on, I don't think that they ended up publishing that... re-publishing that data, they just collected it and used it and acknowledged, you know, the complication of it a little bit when they were putting their paper together. And and so from that decision perspective, it was still guided by sort of the legal, whatever, legal status where, yes, public, we're allowed to do this. But I think that they would have even if they were allowed, I think they would have been hesitant to share that data. For those reasons.

**Sara Mannheimer** 38:50

Yeah. Huh. Yes, so then that affects like the reproducibility or, you know, the ability of the research to be extended?

**DC04** 39:05

Yeah, it's hard to extend and scale it. Yeah. And I mean, people aren't doing that a lot yet. But we would like to.

**Sara Mannheimer** 39:11

Yeah. Okay, great. This is great. Okay, we're on question five. So actually, we're doing fine on time. Privacy and confidentiality. During the [survey of U.S. service members] example, what challenges did you encounter, if any, related to privacy of the people represented in the data? And then what strategies did you use to address the challenges?

**DC04** 39:43

Yeah, so again, because of the nature of that data, we did not implement a lot to explicitly protect privacy, right. The pieces... not only was everything available in one way or another, but the survey data that would limit stuff down much, was already was already there in its form. And we sort of had, it would have taken a lot to, to make the decision that it was preferable to change the data, other than just cleaning it up. So it's more usable and better labeled and things. Given that this has already been used for so many projects, it's part of the public record, survey data has been used in the past. I imagine there's qualitative data in there that you could identify someone with, probably with probably some less than things they might not want shared, if somebody were to track that down, and I guess all you can do sort of from an ethical perspective, setting aside whether it's legally permissible is say, you know, this was a long time ago, and, um, hopefully nobody goes to that effort for something that happened that while a while ago, and actually don't use it, think about stuff. Yeah.

**Sara Mannheimer** 41:26

Did you have like a terms of-- You know, because you're sort of changing the format of the data, even though it's been in the public domain for so long, you're like changing the access of... the ability to access?

**DC04** 41:38

I mean, it would be a lot more likely that someone would de-identify the qualitative information when it's masked and digital and searchable and all of that, than in a folder in the [archive]. Yeah, no doubt. Um.

**Sara Mannheimer** 41:56

Have you thought about...

**DC04** 41:57

Oh, go ahead.

**Sara Mannheimer** 41:59

I guess I was gonna ask like, have you thought about a Terms of Use sort of thing? Is that something that you've discussed, because I know some data repositories do that, have special Terms of Use that says you can't try and identify the data.

**DC04** 42:13

When we put what we put it in [a data repository] we'll undoubtedly have something like that attached to it. Um, we haven't had that discussion about the web page, that's going to be for more public consumption. And that's probably when we should have because, you know, there will be modules for high schoolers and middle schoolers, and but also people from all over and people will probably do, you know, try to do like, see if they can spot somebody they know, or whatever. I don't know what, it's still a small percent of the armed forces. But there's a lot of people in there.

**Sara Mannheimer** 42:54

Yeah.

**DC04** 42:55

We're at thousands and thousands of soldiers that are in this data.

**Sara Mannheimer** 43:03

What an amazing data set. That is really cool.

**DC04** 43:05

It's crazy. It's a fun project to be involved in, besides, despite the complications.

**Sara Mannheimer** 43:12

Yeah. So I guess, let's talk about big data, the challenges of privacy for for big, big data. You've talked about like, just using public data, and then the challenges of what is public and private, really? So maybe this is kind of related to your answer just before about informed consent?

**DC04** 43:43

Yeah. Yeah, you basically run into the same things here, which is that the data are technically already public. And we don't have a mandate to change that. But there could be issues. I mean, they're they're they're... It's certainly worth thinking about how it might be used, what could be protected. You know, um hopefully, say that a lot of information in the qualitative stuff about race relations, will be used to think about, you know, how the military was similar and different from the rest of the country in terms of how people connected to each other across races, both on and off the clock. But you could also probably, I'm sure a resourceful person could use it for some sinister ends, you know, even without trying to re-identify people. But just to sort of paint a picture of one group or another as less moral, less deserving, less something.

**Sara Mannheimer** 45:12

Yeah. I mean, I feel like that's just how, when you put data out there, you just don't know how it will be used. And yeah, we assume it's for good. And probably it mostly is.

**DC04** 45:27

Yeah, it becomes more of an issue I think when instead of putting data in an [academic research data repository], you're putting it somewhere that's meant for meant to attract any old set of eyes. And it's going to have press releases about it. And it's going to be, you know, it's already been like, you know, little radio interviews and whatever, but there will be lots of public stuff. And our hope is that lots of people interact with it. So there's a lot more. Um, it's not necessarily more technically vulnerable, but practically, it's much more likely to be at risk of something like that, if there's ways to use it in that way, I guess.

**Sara Mannheimer** 46:15

Yeah.

**DC04** 46:18

This isn't a place where I have the answers, it's just, I've been thinking about it more over this kind of thing more over time. Like, you know how in Flint, the people really basically gotten tired of researchers coming in, and essentially collecting data on them and then leaving, and getting famous publishing data about it, and not really ended up giving them any benefits a lot of this. And they sort of took matters into their own hands and said, "We, you know, we want to basically a governance thing, where if you want to come and study us, you come and ask us, tell us what we're going to get out of it. We'll say okay, or no."

**Sara Mannheimer** 46:58

Yeah.

**DC04** 47:00

And I really like that sort of idea of community driven data governance. And the same thing, you know, when we've had even when we have community speakers, and their speeches might be archived on our [systems], or, or they're working on projects, we want to make sure that, to the extent possible, people have control. You can't do that, in the case of this [survey of U.S. service members] data. Or really, either in the case of the big data, because it's so, so, so disconnected already. But when you're working with new qualitative data, when you're, you know, talking to people to do try and find those ways to let people have a say, not only informed consent, but later. Like, do you think this represents you?

**Sara Mannheimer** 48:04

Yeah.

**DC04** 48:05

You know, yeah, even if it's just like an internal evaluation study, we're studying [data-related teaching and learning practices] in our university. So, you know, when we get the report ready, we'll run it by them and say, "Hey, does this make sense? You're the ones we talked to. We used your words."

**Sara Mannheimer** 48:24

Yeah. Yeah, I talked to someone who did, a qualitative researcher, who sent the transcripts back to the people who had been interviewed, for it to be checked. It was more of a sensitive topic. So it was very important that that happened. I feel like that's a lot to do in a normal qualitative study. But that's one way.

**DC04** 48:49

Yeah.

**Sara Mannheimer** 48:50

And I've also read about the idea of doing focus groups for big social data, like, could bring together a representative sample of the community you're studying in a big in a social media context, talk with them about like, what their expectations would be and what they think the community might want. It's, you know.

**DC04** 49:11

There are some really great opportunities that are difficult from a human subjects management perspective, but really powerful that are going on with right now, with sort of pushing the boundaries of mixing big social data and qualitative data in some ways, like, this is not mine, but I was fascinated by a presentation I heard the other day. A researcher asked families for permission to collect data from their group chats, in the whole family and then they basically had like a focus group with the family. So it's like a family interview almost like family therapy or something this was in that a that, that the therapy department where they would talk about the dynamics, right? Like, "Oh, well, we have a separate chain, we have a separate group email thread without Uncle Bob on it. Where we talk about things that, you know, that he's crazy about, or whatever." Hmm. Interesting. They said that it was a real pain to negotiate the whole thing with the IRB and with the participants, right? Because you're asking this, you basically have to get permission of however many 5, 10, 20 different people to use this data. And I'm sure they can't, if they publish it, it would have to be restricted. But it's a work, those kinds of things are, I think, really powerful ways to faithfully represent people's own experience.

**Sara Mannheimer** 51:07

Yeah. Yeah. It's so interesting. I mean, this is such an interesting topic, because no one knows exactly what to do. You know. I'm hoping that at least we'll come up with some new sort of recommendation. Alright, last question. Intellectual property. So during your example, what challenges did you encounter, if any, about intellectual property concerns of archiving or publishing the data? And then what strategies did you use if you did have challenges?

**DC04** 51:47

Yeah. Um, so in the [survey of U.S. service members], we did have IP issues, because, um, [the data repository], essentially claimed to have added value to the original data, by sort of archiving it in their format and making the code books and metadata and whatever. So there had actually had to be a negotiation process about the survey data in particular, basically, so that we could have sign off from the [archive], that the data we were using, we were allowed to use and we didn't have to worry about like, we didn't have to put it on a particular website, or pay fees or anything like that. Um, and, and it worked out. It hasn't been a deal, a major thing for the qualitative data, um, associated with that project. Um, but in working with big social data, social media, text mining, all of those things, it's always an issue. And I could certainly see it being bigger in other kinds of qualitative settings too. But, you know, like, the first, I would say, I would say, in 50%, or more of the big data related projects that come across my desk, the first thing I have to do is have a discussion about what the terms of service are, whether they're allowed to get the data in the way they think they want to get it, whether they're allowed to use the data and the way they think they want to use it, whether they're going to be able to share their data.

**Sara Mannheimer** 53:55

Yeah.

**DC04** 53:56

And so on, especially with the databases, but the social data, too.

**Sara Mannheimer** 54:02

Yeah. That's interesting, because in the databases, there's like, similarity to be drawn between that and the social media companies where they like, claim intellectual property.

**DC04** 54:17

Yeah, yeah, it's all about control. And like, you know, like with the library databases, the bigger ones, usually you can work with them, to get access to programmatic access to say, 2000 articles the way you want. But you can't just get it from their website, you have to pay them an absurd amount of money to basically buy, pay them to pull the data out and give it to you. And hope it's what you need and in the right format. And we fought as a library to, we've had discussions about when those data buys do happen, whether it's a one off from a database we already subscribe to or like, say buying access to, like these small data vendors. Examples I'm familiar with, like voting--somebody does county level voting data. And we have tried as a library to find ways to funnel more people into talking to our collections people before they hit buy, rather than just telling their department I need $300 for this, because we found that some of these data sources, you know, have been, not only have we paid for them more than once, but they've had to go through legal and purchasing and all that, you know, all that effort more than once. And we're trying to move toward pushing the publishers for models where we can buy it. And if we're paying for it, we can let at least our users use it on our website.

**Sara Mannheimer** 56:11

All right, great. Well, my seventh question is, are there any issues or challenges that arose during your example that we haven't talked about yet?

**DC04** 56:24

I mean, in both cases, the biggest challenge is getting people to buy in that it's necessary and important to try and share their data responsibly. I mean, sometimes with big social, they'll do it anyway because they want to show off and you know, have their code repository and whatever. But when I encounter researchers, they always agree in principle, that safe data sharing is good. They agree in principle that transparency is good. But they don't get it, either they get stuck at the but I'm worried about, I'm worried about my participants phase. And, and, you know, like, when we tell them, you know, you, you can't guarantee anonymity ever, they're like "Okay, bye." Or they're like, "That sounds good. But unless somebody is making me do it, that's too much work."

**Sara Mannheimer** 58:03

Totally.

**DC04** 58:04

Yeah, and we put in an open access policy where people are supposed to share things they publish. Um, but that is really, which is really good. That is very good. But it's gonna mean even one more thing that people feel like they have to do when they're already done with the part they care about. You know, like, your paper's out the door, you got published, you got your CV. You don't think everyone, anyone's ever gonna need to use that, again. Especially qualitative people are very much like, "Well, yeah, but why would anybody want to use that?" I'm like, well, if I had a database of all these studies, I could do a pilot for my study and say, you know, what, have other people found that's similar? Like, what have other people in similar situations, you know, had to say. And I can start there when I'm developing my interview questions or figuring out how to interview and I might even be able to actually use that for real conclusions. But even if I can't, even if I have to go out and collect my own interviews, then I'm not starting fresh.

**Sara Mannheimer** 59:30

Yeah.

**DC04** 59:31

And a lot more context.

**Sara Mannheimer** 59:34

Hopefully, the more data we have published, like, and the more people who do that, then the more that will take, take hold, you know, the more examples we have.

**DC04** 59:47

Yeah, my big pushes on this have been one getting data published, which is just painfully slow, because you have to get you know, it's two years from the time you catch somebody to the time it might get published.

**Sara Mannheimer** 59:58

Right.

**DC04** 59:59

So I I've been in my position [for a few years], I think we have yet to have any one, no, we had one deposit completed at [data repository], I think. I'm guilty too, I have not finished my first deposit. But also my other emphasis is just on on teaching people how to think about the data. So giving workshops where... and an online learning objects and things, where people can see the benefits of using secondary qualitative data that you don't always have to ask. Don't always have to interview somebody new, always have to survey people figure this out. So we've got, you know, the 20 big surveys that are used by thousands and thousands and thousands of researchers to do cool stuff.

**Sara Mannheimer** 1:00:57

Yeah.

**DC04** 1:01:04

A lot of qualitative researchers have epistemological reasons to be hesitant about that. Right. Everything's a different context, and you just can't get deep enough.

**Sara Mannheimer** 1:01:12

Totally.

**DC04** 1:01:13

But I think that's a little artificial. I think that's taking it too far to the one side, I don't think you should just treat qualitative research, like Big Data research that you didn't get off the internet.

**Sara Mannheimer** 1:01:29

Right.

**DC04** 1:01:30

But you also shouldn't treat qualitative research as this like, pristine thing that "Oh, you weren't there. You wouldn't know." We can still gain value from it. Otherwise, why are you doing the research and publishing it? Because even your published, you know, like, that's subject to the same things.

**Sara Mannheimer** 1:01:49

Totally. Yeah. Well, thank you so much. I really, this has been such a fun conversation. I really appreciate you taking an hour out of your day.