DC10\_transcript\_deidentified

**SUMMARY KEYWORDS**

data, people, curator, consent form, context, qualitative data, consent, datasets, authors, curation, set, curate, metadata, identifying, research, field, big, publish, challenges

**SPEAKERS**

Sara Mannheimer, DC10

**Sara Mannheimer** 00:03

So I'm Sara, I am data librarian at Montana State University. I am working on this research that I'm trying to, as I explained in the consent form, connect these communities of practice around qualitative data, with big social data, where researchers tend to be more in the computer science realm. So the goal is that we will be able to identify data curation practices that will help both communities be more responsible when they're conducting research. So I'm talking to the qualitative and big social data research communities, but it's more about what we as librarians and curators can do to support the research. I did a lit review where I identified six key challenges—consent, data quality, data comparability, context, privacy, and intellectual property. And those six issues overlap between big social data and qualitative data. And so the interview has six key questions, one for each of those. And then some beginning questions and follow up questions. It generally takes 75 minutes or less to have these conversations. So yeah, any questions for me before we begin?

**Sara Mannheimer** 03:12

All right. So let's get started. Tell me about the types of data you usually curate and then what your interests are regarding data curation?

**DC10** 03:21

Hmm, yeah, So I work at [a data repository] as a data curator, where I specialize in human data like human research data, Human Subjects Information (HSI) specifically. So those are the types of datasets—if they're more challenging, which they usually are, people sent tend to send my way in the group.

**Sara Mannheimer** 04:44

I asked you to identify an example that we can use for more specifics. And you can talk about other examples too, but this just helps structure the interview. But maybe you'll have some insight about big social data as well. Okay, okay, cool. So, did you have an example in mind already or an example or two?

**DC10** 05:14

Yeah. So the first the specific example for qualitative data. I see a lot of qualitative data sets that come through [the data repository]. They're usually interview datasets, transcript datasets, and I find that curators are afraid of them. So there are some data sets that tend to wait to be curated longer than they should, because newer curators don't know how to handle them or know that, Oh, this is going to be something that will take me longer to properly curate. So I try to be, given my HSI background, a curator who is willing to dive in on those and take one for the team, even though I find them more interesting than a lot of the datasets we receive. So it's a win-win, I guess. I don't know how specific of an example you need here.

**Sara Mannheimer** 06:22

Maybe like an interview transcript that you've curated recently, so that you just have it top of mind.

**DC10** 06:29

Sure. Yeah, there was a really interesting one that I worked on recently, that was actually a combination of field notes from students that were in a science class. And interviews, short interviews that they had conducted before and after the class to track how much they'd gotten out of the experience. And the field notes were particularly tough to curate because they were photocopies. And there was just a lot there, people scribbling little notes in the margins, and people not following one singular format. So since everyone had their own style for their individual lab notebooks, I had to go through those really carefully and work very closely with the authors to make sure that we weren't revealing any identifying information.

**Sara Mannheimer** 07:38

Nice. Okay. Especially with students, right? Were they under 18?

**DC10** 07:45

These were college students.

**Sara Mannheimer** 10:05

Okay, cool. Great. For the qualitative data, did they have a data management plan that required certain treatment of the data? Do you know?

**DC10** 10:22

The only data, the only type of data that [the data repository] has its own special workflow that curators are supposed to follow is human subjects information, like we need to make sure that things are de-identified properly. But there is not a workflow or a guide for qualitative data that I know of.

**Sara Mannheimer** 10:49

All right. Um, okay, I think we'll get started on the issues, then. The first one is context. So I have these two quotes, just to give you an idea of what I mean by context. So first, for qualitative data.

Since qualitative research is a process that includes deep and prolonged contact and connection with research subjects, attempting to understand subjects within their own context. That means that qualitative data are highly context dependent. So Heinz writes, context is a source of data meaning and understanding in itself and ignoring context, underusing it or not recognizing one's own context driven perspective, could result in missed meaning, or misunderstanding of human phenomena. So this idea that data is part of a much broader context, especially in qualitative research.

And then for big social data.

When we collect data from social media platforms, just like in traditional spaces, the context is important. But the context of a social media post might be absent or difficult to understand, since social media posts are short pieces of text, or an image or a video that are taken from the larger context of someone's personal and public life. And you might not know that context when you're doing social media research. So the out-of-context effect can be compounded when data are collected at a large scale.

So during this example, what challenges did you encounter, if any, when you're trying to capture the context in which the data was collected? So just start with the qualitative and then we can move from there?

**DC10** 12:37

Yeah, I think the largest struggle in my role is that, you know, most qualitative data sets, you're dealing with human subjects. You're concerned with potentially identifying them, and you have to follow certain guidelines. And in doing so, you remove a lot of the context that exists in those data sets to begin with, and that's after they've gone through and done their best to protect their research population. And I have mixed feelings about that, because the scientific community has a lot to gain from having at least the fullest picture that they can take away from qualitative data sets. But we also have to keep in mind that we have to protect people, we have to present information in a way that aligns with what we've, as researchers, said in our consent form—what people have actually like consented to going into a study. And most people are not comfortable with having the broader context revealed about them when they're being interviewed or included in the study.

**Sara Mannheimer** 14:11

So what strategies would you use to try to include as much context as you can while still protecting their identities?

**DC10** 14:20

I think it's a matter of distilling what aspects of the study are most important? And the if there's associated data, or if there's associated code that's used to run coded data that's pulled out of qualitative information, like if they're coding interviews, and then using the code from those interviews to run an R code or a MATLAB code or whatever that looks like.

**Sara Mannheimer** 14:53

That happened in this example? Was there also a code?

**DC10** 14:58

They did like simple Excel dataset for that, like it was just they distilled the information down into like a very simple Excel sheet. There was not a code that they were running for that, but I have seen that where you're then running things through the script. So that is often a great guide for me, to see, Okay, what do they think are the most important aspects of this overall data set? And what are the things that I should if I can, as a curator, keep intact? Because clearly they believe that those elements have a lot to... that they can service the community that they're trying to service, right? Like, they can be additive. And I want to trust them as experts in their field. But then, of course, if what they're trying to publish is like people's individual IP addresses, or people's individual birth dates, or names or things that are very directly identifying, I can't justify that.

**Sara Mannheimer** 16:12

Yeah. What about for big social data, like thinking about context there? I mean, I guess you're familiar with this idea if, even if you don't do it, but if you have specific examples, that's helpful, too. But like, how do you think about what context means when you're conducting research with big social data?

**DC10** 16:47

Yeah, I think in general, big social data, it's complicated. Because when you're when your people have, without necessarily knowing it, when they check the box for like, whatever the terms and conditions state, which are usually pages and pages and pages, and people don't read them. There's usually stuff in there that says, like, we have the right to collect information about you. And that doesn't necessarily mean that they have the right to share it externally. But at least internally, they have a right to data about you as a user. And there's an interesting thing that occurs when when user data that people have consented to being collected, is made public in a way that de-identifies people, but exposes the fact that the data is being collected in the first place, if that makes sense, that will often elicit this fearful or shocked response from the general community when they're like, wait, we didn't know that you were doing that.

**Sara Mannheimer** 19:50

Okay. Back to context. When you're working on curating qualitative data, do you use any strategies to try and sort of highlight context? You know, I'm thinking just like any type of documentation or you know, broad demographic information, like what strategies might you use to support context?

**DC10** 20:43

Like I was saying, I think a lot of it comes down to like, what the authors really want out there—strike a balance there. But also, metadata. Really reinforcing the importance of metadata. I think with a lot of these repositories, people think like, Oh, I can just like, plop my data here. And the only person who will ever look at it or care about it is somebody who's doing a deep dive on my related manuscript and has read every word of the manuscript, and therefore, they have everything they need to contextualize the dataset. And that just isn't the case. So as a curator, I feel responsible for reinforcing Hey, this should be interpretable and reusable on its own, and the more context you can provide the better. So adding a really good README file, adding back, you know, better metadata on the main page. That sort of thing is where it becomes really important.

**Sara Mannheimer** 22:08

Yeah. Yeah, that's good. A link to related publication really helps with that context part. Okay, cool. Let's move on. Let's do privacy next. And so we were just talking about that. So during your qualitative data example, what challenges did you encounter relating to privacy? You talked a little about that. And then what strategies did you use to address the challenges?

**DC10** 22:43

So we have, I was saying, an internal guideline for how to deal with human subjects information. But no guideline will ever be comprehensive, because humans are complicated, and things that you wouldn't necessarily think are identifying can become identifying, particularly with small study populations. So, for example, you might be asking for a neurological study whether someone is right or left handed. And then you have one person who says, Well, I'm ambidextrous. And that suddenly becomes this potentially identifying piece of information. So you really have to be careful with how you approach datasets of that nature. And in the example of the field notes and the students, I can't I came across that quite a bit because it was like, is a doodle an identifier? Is handwriting an identifier? Probably, you know, if you if you have a small cohort of students, and there are only so many of them that write in cursive or one person has a very specific style of writing. Or one person always labels their pages in a certain way. And their classmates know that. How do I protect these students not only from like, the larger community, but from one another? Like, do I want them to be able to identify one another when looking at this study? And according to the consent form, No, they shouldn't be able to. Which means I had to like go through page by page and ask myself like, what do I need to worry about this handwriting? Do I need to worry about this doodle? Do I need to worry about the way that this person writes their dates, things like that? Which is really hard.

**Sara Mannheimer** 24:57

What did you do like if you did see a doodle or handwriting that felt identifiable. Did you like try to redact it somehow?

**DC10** 24:59

Yeah, those had to be redacted.

**Sara Mannheimer** 25:05

And there isn’t any data that's restricted access in [the data repository]?

**DC10**

No, it's all open access, CC0.

**Sara Mannheimer**

Okay, so I have this question, what similarities and differences do you see regarding privacy for qualitative data and big social data? And like, what do you do you encounter different challenges when trying to protect privacy for qualitative data or big social data? If you have thoughts on that.

**DC10** 25:51

I've seen both at [the data repository].

**Sara Mannheimer** 26:17

Okay. I think that'll be helpful to us. Yeah. Cool. Yeah.

**DC10** 26:22

Okay, can you repeat the question for me now?

**Sara Mannheimer** 26:26

What similarities or differences do you see regarding privacy for qualitative data and big social data? And that's kind of like, similarities and differences in the nature of the challenges that you encounter?

**DC10** 26:40

Yeah. So I think people who do qualitative research tend to be a little more thoughtful about how it's presented, or how the data themselves are presented. So I find that those data sets, you tend to see better metadata, you tend to see like a really comprehensive README, you tend to see more context to support those data. And with like, the big social data, it's usually just like, the data set will include, like 10s of 1000s of users, and they don't treat them like real people. It's just kind of like, here's a point on my chart. And like, I get why that is. But there tends to be more thought and more care behind qualitative work. And that is reflected in the way that datasets are prepared.

**Sara Mannheimer** 28:00

Okay. Cool. Let's talk about data quality. During the qualitative example, what challenges did you encounter, if any, when you were trying to document data quality, so not so much like you examining whether the data is good quality, but just thinking about any issues that could arise like bias or method or missing data that you needed to communicate to a potential future user or that the authors needed to? And then what strategies did you use to address those issues? If any, it's okay to say that you didn't encounter that as well.

**DC10** 28:44

No, I think that's a, it's a bit complicated, because I have different expectations for— and maybe this is because I was a scientist myself, like I have different expectations for people at different points in their careers in terms of like, the quality of their data. So we might store we might publish a data set from a high schooler. Like a really advanced high schooler who's like, why wouldn't I publish my work if it's interesting and additive. And it might be very rudimentary, and it might not have a ton of context provided, but I'm cheering for a burgeoning scientist. And then, there's a different standard that I might have when looking at a postdoc's data set or somebody who is very advanced in their field because people look at their work and see it as like, an example of what the field should be doing. And that's where I think quality plays in more. So it's this weighing of like, How much is this serving the community? How much better could it serve the community, just given how, how its presented, how its organized, how its described how it's contextualized? And can I push the authors in that direction? And, to the extent that it's not like, gonna take me a month back and forth with with, like, to the extent that I can within the realm of reality for my job, I will go back and forth with the authors to say, hey, like this could be improved, this could be more reusable, this could be formatted in a way that makes more sense.

**Sara Mannheimer** 31:18

Yeah. Okay. Cool. All right. And are there different data curation strategies you would use to address quality issues for qualitative data versus big social data?

**DC10** 31:38

different strategies?

**Sara Mannheimer** 31:40

Yeah. Or is it all just sort of like talking to authors and providing good documentation?

**DC10** 31:46

I think that, again, like qualitative data sets tend to require more of a deep dive, like you have to pay more attention to the details. And that's okay. It usually just means like, there's more of a wealth of knowledge present there, more context present there. So I do tend to go into curation of a qualitative data set, knowing Hey, this might take me an hour and like preparing myself for more effort, more of a careful review. And it doesn't necessarily mean that I'm more likely to reach out to the authors and request something because I would say, they're probably less likely to need something changed about their data set. But as a curator, there's more responsibility to really carefully review qualitative data. With the big social data, it's more about like really carefully reading through the metadata, making sure that things—how do I say this—like that there's no like no major ethical or privacy issues that I need to be concerned with. Usually, at that point, like, it's already gone through multiple ethical approvals before I as a curator ever see it, but it's still something to be thinking about. And it's normally just like, I was saying, like a million rows have the same sort of data. So it's more just like, does this file load properly? Does it run through the related code properly? And are there any major issues in the metadata that I need to be concerned with? But I would say I reach out to... I don't like, see, one is being way more problematic than the other. I think they're, they're equally like, I have to reach out to authors from time to time, but for the most part, they're okay.

**Sara Mannheimer** 34:07

And you said you have guidance for de-identification for qualitative data? Do you have any guidance for social media data or big social data?

**DC10** 34:17

No, I think it's it depends on like the form that that takes, like, if it's somebody doing work on like, virtual reality imaging, and it's not actually... there's no human subjects information involved, that it'd be treated just like any other data set where human subjects aren't involved. Same guidelines.

**Sara Mannheimer** 34:47

Alright, cool. Let's talk data comparability. During your example. What challenges did you encounter relating to interoperability or comparability of your data set. This is thinking about like, part of the idea behind depositing data is that you'll be able to use them for future studies, maybe combine them with other research to do more longitudinal analysis or something like that. So was this a challenge that you considered when you were curating this qualitative data? And if so, what strategies did you use to address that challenge? And metadata interoperability is part of it, too.

**DC10** 35:40

And that's the biggest struggle, I think, for the researchers themselves, right? Because they see the value in qualitative data. But they also want to make sure that they're adding to the overall field as much as possible. And how do you strike that balance? And I think the best way to do that is to create those condensed, more quantitative summaries of qualitative data. So like, in that example of those interviews, they were going through coding. Did the person bring up certain things in their field notes? Did they use certain words in their field notes? Did they express changes in their, you know, emotional state, or their approach to the subject that they were working with in their field notes? They had certain things they were looking for. And then they were able to code those. So then you don't just have these transcripts, you have these more like, consumable, quick, summary tables, right. And that's the sort of thing that I think, is pretty essential to, to the interoperability of qualitative data sets. And just like qualitative data in general, and not everyone does that. And I respect that choice. And I'm not going to force that on anyone as a curator. But I do think that people who are more advanced in the field or people who have been doing qualitative research longer learn that that's kind of the way the world..

**Sara Mannheimer** 37:33

Okay, cool. How about for big social data? Do you see comparability challenges there?

**DC10** 37:44

Sure. I think, particularly because those tend to be huge datasets. So it's easier to have like grand takeaways. When you have a million subjects to be like, well, we feel pretty confident that people behave this way. And when you have those more easily distilled outcomes from a study, I think it increases interoperability, and also like, the ease at which they could approach that research longitudinally, like pull that same data in a year, because it's more straightforward to do so.

**Sara Mannheimer** 38:39

Right. Okay, awesome. All right, we have two more informed consent and intellectual property. So we'll go for informed consent. In this example, what challenges did you encounter relating to informed consent for participants, particularly consenting for like, archiving and future use of the data? And then what strategies did you use to address those challenges?

**DC10** 39:06

Right, so this is something that I find generally quite a bit as a curator, but I also found this to be true in this example. So often, people assume, like consent is consent. Like did you... did your research participants consent to being involved in the study, of course, but I was actually on an IRB. And I know that like not all consent forms are made equal. And that's kind of the most important thing that IRBs look at is like what is actually in the fine print in a consent form. So that's something that as a curator I have to consider, and often, if I'm concerned about a population, if they're vulnerable, or if I am worried that there are a lot of things that could potentially identify them that are included in a dataset, I will reach out and request a blank copy of the consent form because that's not a requirement per se. And in rare cases, I will ask for the IRB approval, as well. And I do that because the wording is really important. And sometimes they'll say, like, do you consent to being recorded? But it says nothing about like, do you consent to those recordings then being turned into transcripts that are then published in their entirety? Yeah. And that is a totally different step. Right? So I have to, as a curator, consider what the consent forms really say, I have to reach out and ask, and I have to at times actually ask for all of the copies of the individual consent forms to make sure that people actually provided consent, if it's a vulnerable population. I think in that example I asked for the consent form. And I had concerns that they didn't really clarify that they would be storing the data in a public repository, it was more that it was being used for educational purposes. And like, what does that really mean? So.

**Sara Mannheimer** 41:35

Would you ever publish the consent form with the data? Or is that just for your review?

**DC10** 41:42

Rarely, it's usually just for internal review. And that's because normally, the consent form actually will like, provide information about the people conducting the study, or the environment in which the study was conducted that isn't in the dataset or anywhere in the manuscript, and therefore, like, we don't want to make that information public unnecessarily. So, say, like, the specific hospital where the study was conducted, but you don't really want people to know that.

**Sara Mannheimer** 42:14

Right? Okay. And so what do you do, if you feel that the consent form didn't ask for enough? And you don't think that the participants consented to publishing the transcripts? Would you decline to publish?

**DC10** 42:37

There are times, particularly with like full interview transcripts, where I will say this isn't in scope for [the repository]. However, a great alternative is most universities in particular will have the option to like store that data in their own repositories with password protection. So then if somebody really has a valid reason to reuse or to access the data set, they can do so with consent from the authors, but [the repository] will not move ahead with publishing that sort of data.

**Sara Mannheimer** 43:19

Yeah. Okay, cool. Um, when you're thinking about content, what similarities or differences have you seen regarding the challenges you encounter between qualitative data and social media data or big social media?

**DC10** 43:38

Yeah, again, like consent is blurry with big social data. Because is clicking the box under the terms and conditions update, the same as consenting to your data being used or reused. I don't want to like say too much there. But, you know, that's, that's tricky. I again, like I think you're less likely to see identifying information in big social data sets, and you're more likely to just get like this dump of a ton of information. And with qualitative data sets, you do have to be a little bit more careful because they're more nuanced, and are more likely to have gone through a traditional academic form of consent or form of IRB review. For with big social data, it's like—I could ask for a consent form, but I don't think I would get one.

**Sara Mannheimer** 44:50

Right. Yeah, there's a couple of projects that like, have plugins that you could have a pop up where People say, you know, they do content to your research project.

**Sara Mannheimer** 46:40

Okay, let's move to intellectual property. Did you encounter any challenges about intellectual property when you archiving the student data? So I guess this is like the idea that the people who wrote the field notebooks basically have copyright over their own words and thoughts, but the consent form, you know, they consented to be part of the research and to have those thoughts published.

**DC10** 47:30

Yeah, I don't think people who are concerned with intellectual property, like protecting their IP, are going through [an open repository]. That's just like, not their zone. And if they are, it might be like, there's a separate, there's software that exists behind a paywall, or like in a separate repository somewhere that is relevant to the data that [the repository] stores. But we're not the ones that are going to store that content.

**Sara Mannheimer** 48:04

Right, so you just link out to it. Yeah. Okay, sweet. Um, are there any other issues or challenges that arose during the qualitative example or the social media data that you've curated that I haven't asked you that?

**DC10** 49:10

I mean, at the end of the day, curation is subjective. We are a team of curators with different backgrounds and different ideas about what is useful, what is properly contextualized, what is quality data. And I think like if you could, if you could just like build robots that could do this job. That would be maybe good, maybe horrible. How good the robots were at their jobs, but there is a human component to curation that should be recognized, and that's why some datasets end up being higher quality data sets than others at times is because someone would be willing to take the time to reach out to someone doing qualitative work and say, Hey, you would really benefit from having a summary table.

**Sara Mannheimer** 50:19

Yeah. Right. Well, and that's true of the data creators too, you know, they're all human. So it's true. There's just so many elements. Humans are complicated.

**DC10** 50:33

That's why we study them. Right?

**Sara Mannheimer** 50:37

Awesome. This has been some great, thank you so much.