DC07\_transcript\_deidentified

**SUMMARY KEYWORDS**

data, collected, tweets, people, context, researcher, documenting, twitter, qualitative data, hashtag, irb, api, research, qualitative research, question, published, journalist, project, set

**SPEAKERS**

Sara Mannheimer, DC07

**DC07** 00:01

That's always good to do. I've done interviews where I forgotten to record.

**Sara Mannheimer** 00:11

Terrible! Alright, so I reviewed the research with you, there's a little review at that top about my research and my goals. The interview is based on a literature review that I did, where I identified six key issues that are sort of... that are in common between qualitative data reuse and big social data research. And so the interview is structured around these six questions, the topics, the issues are data quality, context, data comparability, informed consent, privacy and confidentiality and intellectual property. So we'll go through some introductory questions, then a question for each of those issues. And then one wrap up question. It should take about an hour.

**DC07** 00:38

Okay.

**Sara Mannheimer** 00:38

All right. So first, tell me about the types of data you usually curate and then and what your interests are regarding data curation.

**DC07** 00:45

So yeah, I guess. It varies. I mean, so I work [at a university]. And [my unit] works on like helping students and faculty with using, you know, computation in their research. And so that can either be, you know, using computers, to help them ask research questions, right. So, you know, like, for example, collecting data from social media, and, you know, looking at it in some way. But it can also be researchers that are interested in studying computer systems. And from a humanities perspective, that that's more like, you know, people that are looking at the history of particular, social media platforms or something like that. So it kind of is like, it's like a broad area. And, you know, one of the things that the center has done, you know, before, I've only I've been there for [a number of years], and the center's been there for [a number of years]. I can't remember [the exact number of years].

**Sara Mannheimer** 01:24

That's a long time.

**DC07** 02:40

Yeah. I mean, it's kind of a long time for [this focus area]. I guess. So. Yeah. Because it's not that it's a pretty new field, I think. But but over that time, you know, they've worked with different students and researchers on different projects. And so the [unit] has accumulated data, you know, as part of that, that work. And so, yeah, so I guess, you know, some of the data could include, you know, just like, like websites that they built, there's a bunch of those. But also, you know, like, we had one researcher that was looking at [data from] early American books or something. And so we've got a collection of that data as well. And then, you know, [through work on] a social media archiving project, we've accumulated data related to that as well. So researchers that wanted to study particular things. We've we've sort of accumulated some of that.

**Sara Mannheimer** 03:50

Okay, great. And then to structure the interview and help us like have more concrete examples I'm asking people to come up with at least one specific example of a recent time when you either cura-, or it would be curated big social data for sharing or advised or collaborated with social media research on researchers on data collection or analysis.

**DC07** 04:15

Yeah.

**Sara Mannheimer** 04:19

And we can kind of talk about a couple examples, but to have one in mind, it's helpful for the question.

**DC07** 04:33

Okay. Yeah, I mean, I can think of one. Okay. Yeah. Do you want me to tell you what it is or?

**Sara Mannheimer** 04:40

Yes, please. Yeah.

**DC07** 04:44

So I guess [a project team I work on was] approached by a researcher who was interested in studying [a social movement that had social media presence] and and specifically looking at how several of the activists that were involved in the early days [of the movement], how they had died. And sort of like looking at their the the, the traces that they'd left in Twitter. And yeah, and this was a researcher that that was a journalist actually, that wanted to write about, about this topic, right, like, how these these activists had died.

**Sara Mannheimer** 05:47

Okay. This is interesting, because journalists have like different professional standards and values and ethical code. So this will be a cool example.

**DC07** 06:00

Yeah.

**Sara Mannheimer** 06:02

All right. Ah, and have you curated qualitative data in the past as well? Or is it just just social media data?

**DC07** 06:12

Um, no, I definitely have. Mostly as a researcher, so yeah, you know, as part of my own research, I did a lot of qualitative work. So.

**Sara Mannheimer** 06:23

It might be helpful to think of an example there.

**DC07** 06:28

Okay.

**Sara Mannheimer** 06:28

It takes us a little longer. I think we can get through it. But it's nice to have people who have done both, it really gives some good perspective.

**DC07** 06:36

Yeah. Okay, yeah. I mean, so for that, I mean, one of the, so I [some time ago] finished my, my PhD. And I studied, I did a, an ethnographic study of this place [name redacted]. And, and so I was working with them for like a year and a half. Looking at how they, how they curate how they saw how they, how they [conducted digital archiving and preservation]. And, and sort of like what their I was specifically interested in their appraisal practices. So how they decided what to collect and what not to collect. But a big part of that was just creating field field notes. So I was in lots of meetings and in the workplace, with them, creating field notes, and then also studying, like, kind of doing documentary analysis work, like looking at their ticketing systems and in different places where they documented their work. So.

**Sara Mannheimer** 07:59

Nice.

**DC07** 08:00

Yeah.

**Sara Mannheimer** 08:01

And did you publish that data when you are finished? The field notes or, okay.

**DC07** 08:06

No, no, I did offer... so I did a lot of oral history, or not really oral histories, there were interviews like this right with, with them. And I did it wanting to, because they were spending their time talking to me about the history of the place. So I kind of like what you did in your questionnaire, you gave the option of publishing the transcript. Right. And I, I wanted to do the same thing, but they, nobody took me up on it. So.

**Sara Mannheimer** 08:34

Oh, interesting. I mean, it's very hard to de-identify the transcripts, I'm not exactly sure that I will be able to do it. I just, I wanted to be sure. You know, and then I was talking to someone who was like, you know, you don't need to say all the time that you'll deidentify the transcripts. Sometimes it's okay to just give people credit as interviewees. And so, yeah, that's maybe the next study I do. Think about it that way. But.

**DC07** 09:04

Yeah, no, I had some people I for other projects that did want to be identified. And you know, and then we're actually kind of miffed when they found out that they were anonymized right, or?

**Sara Mannheimer** 09:16

Yeah, that's an interesting question with qualitative research. Yeah. Okay, um, was for either of these projects. Did you was there like certain requirements from the grant or from your program that required specific treatment of the data?

**DC07** 09:35

Well, I mean, so for the qualitative, for my own research, I went through IRB, you know, and, yeah, and so, and that's why I actually I haven't published any of the data to because I was just part of the study how it got put together. But for the work with the, with the journalist who's studying the activists, there was no um, explicit, like, like grant...

**Sara Mannheimer** 10:05

Data management plan or something.

**DC07** 10:07

Yeah, no.

**Sara Mannheimer** 10:11

All right. Okay. Let's move on to context. So I think you know, since you didn't publish the qualitative data, well, no, I'd like to, I will just chat about it here and there. But we'll mostly talk about big social data. So look, I have these two quotes that sort of explain the how I'm thinking about context in each of these types of data for each of these types of data. So for qualitative data, "Qualitative research is a process that may include deep and prolonged contact and connection with research subjects, trying to understand subjects within their own context. And so qualitative data are highly context dependent. Context is a source of data meaning and understanding and ignoring context, or under using it or not recognizing one's own context driven perspective could result in incomplete or missed meaning and understanding of human phenomena." And then for big social data, Halavais suggests that "When we collect data from social media platforms, just like when we collect data in traditional spaces, context matters. But the context in the social media posts may be absent or difficult to understand, since social media posts are short pieces of text taken from a larger context of personal and public life. And then this out of context effect is only compounded when data are masked at a large scale." So thinking about these two sort of definitions of context, during your example, when you are working with the journalist, what challenges did you encounter, if any, when you were trying to capture the context in which the data was collected? And then what strategies did you use to address those challenges?

**DC07** 12:04

Yeah, that's, that's a really nice description of the importance of context. And, and I think, so for the journalist, the situation with the journalist. So that the data set that we were working with primarily, was a Twitter data set that had been collected, I believe, for almost a year, it might have been nine months maybe where it was just a dataset where it was collected from this Twitter Streaming API, and looking for any tweet that mentions the keyword [keyword redacted]. And so this was a over a long period of time, right. But well, relative to, I guess, Twitter, maybe nine months is a long period of time, but like, yeah, but it accumulated, you know, millions of tweets, right? Because it was a, it was a big topic over the time period where we collected it, I think it was, like [in the 2010s], if I'm remembering right. And, anyway, so. So that was sort of like how the data was collected, it was collected, before we had any interaction with the journalist, right. So the journalists didn't come to us and say, "Hey, could you start collecting this" and we started doing it. We just identified because of the project, that, that [this hashtag] was something that really kind of like, the hashtag itself started to become prominent, during and immediately after [an event], and we had been in contact with, you know, researchers that were interested in studying it. And so we sort of like proactively, like went and started collecting it. And then, I think in [year]... it might have even been later maybe, yeah, I'd have to look and see exactly when the journalist first reached out to us, but, but, um, but [they] had heard that, that, you know, we that this [project] could potentially help [them] and with this research project that [they] had, and so [they] had identified these individuals that, that, that [they were] interested in studying. So [they] already knew their Twitter handles and, and who they were, and and the fact that they were no longer alive, and wanted to look in our datasets to see what tweets were there, because some of some of the some of the data had been deleted. And so he's looking for traces of what might be left. And so, I mean, the context thing. So important, right? Because, you know, we sort of started this data collection, not totally knowing what the context of the research would be. But we did understand sort of like the context of the [hashtag and the associated events], because we've been talking to activists there, we knew, like what had happened [during the event], we, we knew a lot about how they use Twitter and Social media as a way to sort of mobilize the movement, at least in those early days. And so. So we knew that that context fairly well, but we didn't know the researchers context. And so when that researcher came to us and look for this help, we could look in our data and see what we had from these users, or were, these users were mentioned in the data, but but there were so much of what those users had done online, that was not part of our data set, you know, because of the way that we chose to create it.

**Sara Mannheimer** 16:25

And I guess the journalists probably did research on these people, you know, in addition to the data set, right, so maybe that?

**DC07** 16:37

Yeah, definitely. Yeah. Yeah.

**Sara Mannheimer** 16:40

Is there anything like at [the project], do you... When you're thinking about context, and if you provide the data set? Are there certain like Readmes or certain information that you would provide, in addition to it besides just like when it was collected? And what parameters you use to pull to pull it down from the API?

**DC07** 17:06

Yeah, so they do usually have Readmes, yeah, these data sets, right, I could send you an example of one if you want. Normally, they're fairly minimal. I mean, there, you just, you just mentioned, to the really, I mean, the really important thing is to mention when it started when it stopped. So when you're when we're collecting from Twitter, and then the parameter, so like, which endpoint was used, is really important as well. Because if you're collecting it from the streaming endpoint, that means that you're going to see the data in a particular way that so the data will, you'll collect it as it's being created. So all the engagement data around each post won't, won't be there. So like, how many times have a tweet was retweeted or replied to or all that stuff. Because you're getting it live. It hasn't had a chance to be engaged with yet so.

**Sara Mannheimer** 18:07

Oh okay.

**DC07** 18:08

So that, so knowing that, so. Yeah, so. So that's really important, whether it's the Search API or the Streaming API. But then, like, when it was created, and and with what software. And I think we tended to use the same software, but but yeah, like the version of the software is important sometimes like, because there are bugs and things and yeah, knowing like, that a data set might be impacted by like a bug in a piece of software that was used to create the datasets, pretty important. We also put, I mean, it's just kind of like boilerplate, but like, you know, this data is not to be... so for us because we created at [my university] at least the stuff that I was doing, you just put like a label on there, this is for [internal university] use only. So [this university] use only. Not that it's going to stop anybody, like if they actually got the data set. You know, it's that isn't going to stop anybody from using it. But um, but the idea is that it would be anybody that had access to the data, and then maybe was being asked to share it with someone else like that they know that it was not intended to be just like, put on the web or distributed within, you know, just anybody, right? Yeah.

**Sara Mannheimer** 19:38

Okay, interesting. I think we'll get we'll get to get more into that sort of Terms of Use as we continue to talk.

**DC07** 19:46

Yeah.

**Sara Mannheimer** 19:47

Um, like, okay, as you are working with the social media data, thinking about context as you archive the data set and provide it to researchers. What similarities and differences do did you see between like the curation strategies you use to address these context issues for qualitative data and for big social data? And I guess since you didn't quite get to the publication point of your qualitative data? You might, yeah. But thinking about, like, how you address context in both types? Do you have thoughts on that?

**DC07** 20:40

That's kind of a hard question.

**Sara Mannheimer** 20:42

Yeah. Yeah. I'm like, will you just do my discussion section for me? No, just if you have that. You don't have to stress too much about it.

**DC07** 20:56

It's it's really interesting, I think, because so when I was doing my, my own qualitative research, like I had chosen that site, right, for particular reasons, partly because it proved the opportunity presented itself, but also because once I found out about it, I was like, oh, yeah, that would be a great place for me to, to study what I'm, what I'd like to research. And, and a big part of the project was like, kind of understanding the context like, like you said, so well, and that, you know, in your description of, of why context matters for qualitative research, like, it is, like, understanding that context is almost like the goal, right? Like, it's like, the, the point of doing qualitative research is to understand the context. And in order to get insight into some something, you know, and for me, that was, that was like, how they decided to [collect digital archives]. But yeah, like, I didn't, I'm not sure I totally understood the context when I was when I was starting it out. But you know, at the end of it, I felt like I did and, and, you know, building a thick description of like, like a workplace in this case was like, the whole point. But like, one... Yeah, it's almost like the exact opposite for when we did the well, I mean, it's similar in the sense that, like, the reason why we collected that [redacted] hashtag was because of the engagement that we had had with the project [related to the hashtag]. Right. So we did, we did have an understanding of that context, right, like with I mean, and this was largely, not just me, but like other other participants in the project that that were part of the, the activism and protests and, and that knew people that we could talk to an interview and and yeah, so I mean, I said, it's the opposite. But it really is kind of like the same thing. Like, like, that was the context that that drove us to collect the [redacted] hashtag. But, but, but in some ways, like that context is a little bit distinct from the context of like, okay, I'm going to get every tweet that mentions [the hashtag], right, and save that. Yeah, like, the context there is, like, almost unknown, right? Like, it's like, yeah, we had certain expectations about what we were going to be collecting. But that hashtag gets used in lots of different ways. Right, that, that we didn't, and still don't fully understand. Right? And, yeah, and I'm not sure that yeah, so so it's almost like there's there's two different types of context going on sort of like the context that motivated the creation of the data set. But but the context of the actual data set that we created it yeah, it's it's bigger it or not bigger, but it's, it's almost like it's just it overlaps a little bit, I think, with what we were sort of hoping to collect but but just in the nature of the way that that the data was collected it it includes context that we just don't know, you know, yeah, yeah.

**Sara Mannheimer** 24:56

This is good. Thank you. And yeah, yeah. Okay, cool. Well, let's move to data quality. During the Twitter example, did you encounter any challenges when you were trying to document data quality, like missing data or bots, or even like bias or the quality of your method of collecting? And then what strategies did you use to sort of help clarify those issues for future users?

**DC07** 25:33

So yeah, I mean, we had a lot of difficulty... and largely manifested in communication between me and the reporter. And so we had, we never met in person, it was all either through, you know, video conferencing or email, and and we're always kind of like going back and forth about what, what the data included and what it didn't. And, yeah, I think that that was, that was challenging. So I'm remembering one, one particular case where, so we collected the [redacted] hashtag. And that did give us some insight into like, you know, when these people that that [the reporter] was studying when they use that hashtag, right, because then they were part of the data set. But then there were all these things that they said that were not part of the data set, because we didn't collect them. So we did retroactively try... So users that still had an account on Twitter, you know, they had pass-, they were no longer alive. But but we, they sometimes some of them still had public accounts. And so we went and pulled their timeline. And using a tool that basically scraped the Twitter website, so it didn't use the API. So those endpoints I was talking about earlier, it didn't, didn't use that. And, but we had one user, the scraping tool didn't work. So you can see it, you can go to the website, you know, the Twitter website and see it, but the tool couldn't get it. And part of the reason for that was that the user had been what's called shadow banned. So I don't know if you've heard this term before, but so they're, they're still there on the Twitter website, but they're kind of like, sort of removed from the search index. That's, like, one way I would explain it, but so you can't actually find their, their content when you're searching or looking at a conversation. So it's been kind of like, not banned, but sort of just like hidden a little bit. And then Twitter for a longest time, said that they weren't doing this and denied that they were doing it. But, but there's lots of evidence that they that they do do it. And anyway, so we one of the users have been shadow banned. And, and so like, try to explain like, why we couldn't get their data. And in that case, was very complicated. And it was complicated for us, too, because we didn't totally understand it. Like, like I mentioned, like Twitter themselves said that they didn't do it. That might be different now, and they might have admitted it now. But anyway, that was that was complicated. And, and I'm not sure that we really came up with a solution in that case, other than, like pointing at other people that were having similar problems and sort of like, just just kept talking about it until we both felt like we understood what was going on at it better.

**Sara Mannheimer** 29:10

And then when the researcher published their article, what did they like, try to explain some of that to, you know, like readers, or like, do you have documentation thinking about the data curation side of things? Like, is there a way that you or the researcher sort of documented these issues that you that you encountered? So like, if future people want to use this [redacted] hashtag data set that you have, like, would... Are these things documented for their knowledge?

**DC07** 29:48

Well, it's kind of complicated because the journalist hasn't published the article yet. I'm not sure if they will, either. It's possible. It wasn't that long ago, so they might still be working on it. But but and the data set itself, so the [Twitter hashtag] data set has had its tweet IDs published. So. So the so that people can, you know, rehydrate the data, if they want to work with it. But we don't we didn't make that data set available other than as tweet IDs. And we didn't. This scraping of the accounts, that was just something that we did with the researcher and have not published. The only documentation that exists for, for this is the emails back and forth.

**Sara Mannheimer** 30:53

And thinking just generally about social media data when you're collecting it. Do you see similarities and differences between the way that you would address data quality? Like, do you have a way to like, do you have like a quality control mechanism where you have certain accounts, you know, are bots or there's like certain in your Readme, you say, you know, like in your Readme, you said the likes and retweets wouldn't be available? If it's on the Streaming API or things like that, like? Um, yeah.

**DC07** 31:31

Yeah. Yeah. I mean, I think that probably could do a lot better job of documenting, like, why it matters, like what endpoint that data came from. I think it's there implicitly, like, I'll mention, like, which API was collected from, but but for each data set that we created, I didn't repeat the same information, like, you know, because it's from the Streaming API, the retweet, and, you know, favorite, and reply counts will all be zero, right, like, and so it's there implicitly, it's not it's not made explicit. And I think that that's, that's probably a problem. And in terms of, so what your question was about, um...?

**Sara Mannheimer** 32:19

Well, I guess I was, I am also asking about similarities and differences. When you're thinking about data quality for your qualitative data. Like, you know, maybe you would review the data to make sure that you've talked to the right people or, you know, and then with social media data, it's kind of like, were you able to collect the people that you were interested in? You know, I guess thinking, yeah, from a data curator's perspective. Yeah.

**DC07** 32:55

For some similarities, I think maybe in like, how you're describing, like, sort of returning to the data and sort of like, looking at it and, and kind of like, yeah, reviewing it. So with, I guess, with my qualitative work with a field study, I mean, like, yeah, like, the whole, it's almost like, the whole process was like, like writing stuff down, and then returning to it later, and then trying to, to come at it from different angles and stuff. So that was often like my field notes, right? Like, just basically going back to them. And, and doing like, demoing practices around that. So basically, like looking at the field notes, and then trying to find patterns within there, and then looking for gaps. Like, oh, like, I, this person, I mentioned them, and I should I should really follow up with them. Because they, you know, that one of my informants like, mentioned them, and then, you know, like, so basically doing that work. And I think that the, and for the quant, the, I guess, quantitative social media data side of things, it's almost like it is kind of the same, like practice in a way to, um, so, like, I mentioned, that the, you know, the [Twitter hashtag] data collection was happening over a long period of time, you know, like, eight months or whatever. And, and part of that was making sure that it was still alive, right? So because it's a it's a program, right, that needs to run for that long of a time and and so to do that, you need a server or a computer doesn't have to be a server, but just any computer that can stay with the power on for that long, which sounds simple, but, but actually it can be hard sometimes. And, and then that it has enough disk space right? To save the data um, that it still is connected to the API. It may look like it's running, but it's, but it's not like something's gone wrong. So a big part of doing that work was like the, the program that was running wrote a log, basically. So a file, almost like the field notes, right, in a way, like, it's we're like documenting what the process is, what, what's being collected, how much has been collected, what time it's been collected. And so we're like remembering to periodically return to that just to check that it's still on. And that's something that I could have done a little bit of a better job around. You know, like, putting a note into my calendar, like, hey, maybe after a week has passed, go and take a look at it again, or just, yeah, simple. I think I did end up doing something like that.

**Sara Mannheimer** 35:56

To support completeness. Yeah.

**DC07** 35:59

Yeah, yeah.

**Sara Mannheimer** 36:03

Great.

**DC07** 36:04

And yeah, that's another angle on that, too, is like, we chose to collect this [redacted] hashtag and kind of kept it kind of simple, right? In a way, like, it's just gonna get any tweet that mentions that, but, but there's all kinds of variations on that hashtag that could be collected, right. And this is not for that, for that particular data set because of the significance of that hashtag. We were pretty comfortable, just keeping it simple like that. But other cases where we've done collection like this, you kind of learn more about the collection that needs to be created as you're collecting it. So you've learned about new hashtags, you know, it's, it's, it's a thing in motion, right? and stuff happens. And so people want to revise it, right? Like add, add new hashtags, in, as it as it's going. So that makes it really difficult. I think, too. It makes it harder anyway, to take us as a data set, because that data set wasn't, like created, like, with one kind of set of parameters. It's it was, it sort of changed over time. So if you're gonna make any kind of statistical kind of, like, statements about it, like it becomes difficult, because, yeah, like, it almost be like, if you, you did a, you did a survey or something, and, and then just were, like, change how you were the questions or something in the middle of it or.

**Sara Mannheimer** 37:36

Right.

**DC07** 37:37

Yeah. And then tried to make statements about the survey data afterwards.

**Sara Mannheimer** 37:43

Yeah, that reminds me of qualitative data in that, like, when you do a qualitative study it you don't expect that it will be able to be statistically, you know, like, applied to everyone. It's applicable just to your population, you know, and then you can, like, make inferences from there. But.

**DC07** 38:01

Right, right. Yeah. They're very different, right? Like you can you can learn more about your research question as you're, as you're kind of doing the qualitative work, I think.

**Sara Mannheimer** 38:10

Yeah.

**DC07** 38:11

And, and you can adapt things and.

**Sara Mannheimer** 38:14

Yeah.

**DC07** 38:15

Yeah.

**Sara Mannheimer** 38:18

All right. Let's move to data comparability. So during your example, your social media example, what challenges did you encounter relating to comparability or interoperability of your data set? Like metadata interoperability, or if people I don't know if you've worked with people who like combined different data sets that [a social media archiving project] has created?

**DC07** 38:46

Oh, well, actually, yeah. I mean, so with the [Twitter hashtag] data set, right, like that was I mentioned that we got that from the API directly. So. So that was the Twitter's JSON representation of what a tweet is that that was the thing that we were collecting. But when we when we were working with scraping the data from the timelines of the different users, so I mentioned that, you know, we wanted to fill in the data set by collecting the tweets that different people had on their timeline. And so we use a different tool for doing that. And it scraped the data. And so it did not get that kind of canonical JSON representation of its tweeted, it created like a spreadsheet, CSV file that had certain columns in it, but but not other ones, right. And so one way we kind of tried to sort of I don't know what the right word for it, like correct for that, I guess is that we, we used the scraping tool to get the IDs of the tweets from the timeline. And then we use the other tool for interacting with the API to get the JSON representation of those IDs.

**Sara Mannheimer** 40:15

Oh, okay.

**DC07** 40:16

So it kind of like, like a hybrid approach. But doing that did allow us to use tools that were expecting to have the JSON, you know, in a specific, yes, specific representation available. It allowed us to use those tools with the scrape data.

**Sara Mannheimer** 40:39

Okay. I'm gonna skip the comparison question on this just for time, because I want to get to the other ones, um, the unless you have something in mind, like, clear similarities or differences regarding data comparability for qualitative data and big social data.

**DC07** 41:03

Um, yeah, I mean, I definitely have problems with compare-- comparability of qualitative data, like I had different. So the, like, data that I written as field notes compared to interview transcripts, like, you know, that interview transcripts were literally like what people said, and then my field notes were what I remembered people saying, you know, like when I went home and like scribbled in my notebook, yeah. And they were often at odds, right? Like, I, I'd forgotten some things or yeah, and people use different language, right to talk about things in a language that I wrote down in my field notes. So but those are generative sources of data. I think, like noticing those those differences, I think and like, right, yeah, I think that, for my qualitative work, those were almost like guiding post to like, like, finding things that were interesting, right. Like, where I was sort of missing something. Yeah, there's maybe with the data, it's like, proceed more as a as a problem or a bug. Right. You know, and less interesting, maybe, but.

**Sara Mannheimer** 42:26

Yeah.

**DC07** 42:28

I don't know if that's helpful. Yeah.

**Sara Mannheimer** 42:32

Okay. Okay. Great. Let's move to informed consent. In your example, what challenges did you encounter, if any, relating to informed consent for participants? I'd say particularly consent for future use. But since your data set is kind of like staying private, except for those tweet IDs, you can just think about like consent, in terms of for research, you know, like these folks who were in the data set that use the [redacted] hashtag, what were your thoughts as you collected the data about their consent for future use of the data set?

**DC07** 43:24

Yeah. I mean, I guess the sharing of the tweet IDs is like. Yeah, I mean, the reason for that, at least for our project has been totally motivated by issues of consent. Right. Like, that's, that's the reason why we chose to do it. But, but it's not terribly satisfying. So. So the, yeah, so so we collected the [redacted] hashtag. And obviously, like, we didn't ask anybody, whether they wanted their data to be collected. Maybe it's not obvious, but we didn't. We didn't ask anyone. And we collected millions of, of tweets, you know, from hundreds of thousands of users. And and so when it, I mean, so Twitter has, they have rules around like how you can share data. And that you have to agree to when you like, use their API. And, and so some of our thinking was like, motivated by that right by just sort of like understanding what the limits of what Twitter wants to do with the data are. But uh, and those are like, yeah, they're kind of weird documents, like they're, they both at the same time, they, they both, like protect their own business interests, right? Because they don't want anybody just having their data, right. Because they want people to pay them for it. But, uh, but also like, it's kind of couched in the language of, like, we respect content creators, like, they own the content that they create. We want, we want to, like, honor their voices, and, and so like, so their terms of service kind of do both those things at the same time, but for us, like, and they and they basically promote this idea of sharing tweet IDs, rather than any data of the tweets themselves. So that...

**Sara Mannheimer** 45:48

And they require that right?

**DC07** 45:50

Yeah. And they, they've sort of changed their rules over the years a little bit around it. At one point, they, they limited it to, you can only share, I think it's a few hundred thousand, maybe it was a million, I can't remember, but some some, some numeric limit on how many IDs could be shared. But then they kind of backpedaled a little bit and like, unless you're an academic institution, in which case, you can share more tweet IDs. But, uh, but yeah, for us it I mean, the nice thing about doing that is that it does, it doesn't, it doesn't give, it doesn't require any consent to be given to be part of the data set, which is, you know, problematic, but it, but it does give some agency to the people who created the content to, to remove their tweets from circulation, right. So if the tweet ID data sets online, and they have some tweets that are in it, but they choose to protect their account, or delete some tweets or delete their account, then those will no longer be able to be turned back into data again. So they're effectively removed from circulation, but that doesn't mean that anybody who has them already will remove them. Right. So that, like people that have already collected them, will still have them. And right. So, yeah, it's it's really a kind of not a very satisfying solution at all. But we were we were also trying to not only think about consent, but also researchers that were using the data sets and wanting to share the data that was, you know, underneath their their studies, right, that they wanted to publish the data for, for reproducibility for for people that are just trying to understand what was going on in research. So, so we're kind of trying to balance those two things. That's how we ended up with the consent, but also the reproducibility I guess there.

**Sara Mannheimer** 48:08

Yeah. Transparency.

**DC07** 48:11

Yeah.

**Sara Mannheimer** 48:11

Yeah. I feel that's another similarity? Well, I should let you talk. Is like, with qualitative data, it's not exactly reproducible, either.

**DC07** 48:25

Yeah that's true.

**Sara Mannheimer** 48:25

It's very specific to one point.

**DC07** 48:28

Yeah, I agree.

**Sara Mannheimer** 48:36

Okay, did what... do you have any thoughts about your qualitative research you did? On informed consent there as it relates to your these informed consent issues? Or do they seem pretty distinct?

**DC07** 48:53

Well, I mean, that's the nice thing about the qualitative work was that because I had to go through IRB to do it, right. And and, you know, I did so I was at [a university]. So [my research site also had an IRB]. And so I actually had to do IRB in two places, so I had to... I had to do it with them. And then also with [my university]. I forget... one of them had yeah, anyway, like, it was complicated, but I had to like think about it twice and from different perspectives and but with with data collection from me, maybe things have changed. I haven't done this recently, but for the the data collection from Twitter, like, normally, like, if you go through IRB, they're like, "Are you collecting public data? Okay, well, then you don't have to do this." And...

**Sara Mannheimer** 49:49

You didn't even have to get like an exempt exempt status or anything like that?

**DC07** 49:54

I mean, it's so so actually, for this [Twitter hashtag] data set, we didn't no, we didn't even do that, we didn't. Mostly because it wasn't part... So a lot of IRB gets framed in terms of a study, right. Like, I got this research question. And in this case, we weren't. We were like, well, we think this hashtag is going to be significant for historical reasons. And, yeah, you want to, you want to collect it. And, and so there was no research question to even engage with an IRB.

**Sara Mannheimer** 50:31

Right? You were like, archivists, just collecting stuff.

**DC07** 50:36

Yeah.

**Sara Mannheimer** 50:37

Interesting.

**DC07** 50:39

So there, it was kind of different. But, but when the journalist interacted with us, we also did not engage with IRB then either, you know, arguably, like, maybe it should have, right. Even though it was wasn't a, like a study, like a, like a, but it was research right it was journalistic research that was going on.

**Sara Mannheimer** 51:09

Yeah. Yeah. Have when you work with, like, academic researchers, have they gone through IRB before using your data?

**DC07** 51:22

Yeah, in some cases, yeah. Yeah. I mean, often, it's been waived, right, because I haven't really worked with, like, private data before. Yeah, it's all been public things that are, you know, public on the web. But, uh, but yeah, like, definitely. But sometimes it's something that we had to, or I had to ask about, right. Like, it wasn't always the researcher, you know, saying I need to do this. Mostly, I think part of the reason for that is because I mentioned that the center that I work in is any in the arts and humanities. So it's, we did have some interaction with social scientists, too. But but many of the researchers are humanists, and hum-, like humanities and IRB are kind of like oil and water a little bit like they I don't think I don't even think a lot of people know about IRB in the humanities, as a, like, have much experience, like, going through the process of doing it. Right. Yeah. But there have been cases where we have not not this specific one. But there are cases where it has been done.

**Sara Mannheimer** 52:46

Interesting. Alright, let's move to privacy and confidentiality. During the when you were working with the [Twitter hashtag] data set, what challenges did you encounter, if any relating to privacy for the people represented in the data? And maybe we have some overlap here. But then what strategies did you use to address the challenges? Like, I guess the tweet IDs is one way to do it. So people can make their tweets private after the fact.

**DC07** 53:21

Yeah, that's true. Yeah. Yeah.

**Sara Mannheimer** 53:26

I'm also thinking about like, the the way that you collected the data using the hashtag, like, did that make this data collection different from just like pulling down any tweet that had certain words in it, you know? Like, are you assuming that if people are using the hashtag, they're wanting to, like, add their voice to a public conversation?

**DC07** 53:51

Yeah, I mean, I definitely like using a hashtag is like a performative thing, right? Like, you're often using it because you want your content to be sort of like, included in some kind of aggregate right? Like where people might notice it, if they're watching the hashtag. I mean, [this hashtag] was a kind of maybe a somewhat unique one, in that it really was it was a hashtag, but it was also it became a movement, right that almost like I mean, there have been a few, I suppose, like [this hashtag is not] totally unique, but yeah, so yeah, you're right. I'm not sure. Yeah, just thinking about the qualitative work. I don't know. The privacy for the qualitative work was yeah, I mean, it was largely, you know, through the IRB, like I said I was not going to, unless, unless they gave permission, I wasn't going to publish the data. And, and I did say that, if I was going to use quotes from them, I would anonymize them and, or deidentify them and and also give them a chance to, um, so I ended up sending them. And this is like, what journalists don't do, right? I suppose. But like I sent them, like, the paragraphs that I was gonna be mentioning their, their quote, and to give them some context, and then give...

**Sara Mannheimer** 55:40

Oh, that's very thoughtful.

**DC07** 55:42

....to rephrase it if they wanted to, like if they, if they felt like, it wasn't capturing what they meant. So yeah, I was kind of I did that mostly because I wanted people to be as honest as they could be, you know, and but, uh, but yeah, that that for the [Twitter hashtag] data set, like I mean, these were people that were no longer alive, right? So there was no, I mean, nothing has been published yet. So I know that, like the journalist, did want to publish a piece that had their tweets in it, that was going to be a part of the, of the article that [they] wanted to write. And I never, you know, I shared data with [them], so like tweet IDs, and the text of tweets and things like that, which you can turn into URLs to look at the content on the web. Yeah, I think I even gave [them] URLs for for those tweets, but [they] didn't, [they haven't] published anything yet. And so I don't know if [they were] going to go back to, to the, like, if there were any family that were let... you know, that could be talked to about whether they wanted the tweets to be published that way. I don't know. Who's gonna do that or not? Or if that's part of what a journalist would should or would do. Yeah, I'm actually not totally up on ethics of journalism.

**Sara Mannheimer** 57:17

Yeah, me neither. Yeah. What about when you have this restricted access to like, the full data set that you let the journalist use? Do you have specific, like Terms of Service or Terms of Use, basically, for your, for the data?

**DC07** 57:38

Yeah, I mean, not like formal document that I have them sign or anything like that, although, you know, if I was doing it more, maybe that would be a good idea. But definitely communicating that, you know, the data should not just be put on the web somewhere. That's the main thing I try to communicate. Yeah, like the data and making it available to them. And oftentimes, I work with people at [the university where DC07 works]. And the way that I say that is like, this is a data set that we kind of like, like a, like a library book or archival collection, right, that's at the university that's there for the university community to use. And we're like, talking about the data set in the same way. But you know, in the case with a journalist who's not part of [the university], it's a little bit complicated. And I'm not all I can say is like, you know, the reasons why I think it's important for them, not just to put it on the web, right? That it is people's privacy, even though it might be public data, seeing it aggregated together and an easy to access way could provide visibility to it, you know, that that otherwise be hard to find, right? Like a needle in the haystack. So I tried to communicate, I have tried to communicate that and did with this researcher. But it was all through email and.

**Sara Mannheimer** 59:16

Alright, let's move to our last question or last formal question about intellectual property. During your example, with the Twitter hashtag data set, did you encounter any issues regarding intellectual property concerns of archiving, or using this data? Thinking about like, Twitter saying that the tweets belong to the content creators? Yeah. And then what did you feel were your responsibilities there, you know.

**DC07** 59:50

Yeah. Yeah, I guess some of the issues never really materialized because the article hasn't been published yet. But if it were to be published, I think they would materialize pretty quick. Because, like, there needs to be a decision about whether to make the tweets available as part of a research....[brief interruption] Yeah, so we'd have to, and I know a little bit about like journalists that, that do have some, like, ethics around. Like, if they're going to publish a tweet from somebody trying to get to at least attribute it, right. Attribute it to the person who did it. But, um, but yeah, it's, it's difficult because, like, some, some tweets, it's less important, because they've been seen by so many people and, and, and the person who may be sent it is like, a celebrity or something. And, you know, like it. But in this case, these were, I mean, some of the activists were, were kind of, like, fairly well known, like, lots of followers, but some of them weren't. And so like, I mean, they were no longer alive. So it's like, who owns the intellectual property to it? It's kind of difficult to say, right.

**Sara Mannheimer** 1:01:45

Did you think about Fair Use, you know, since a lot of the what you do is for educational for, like in an educational institution?

**DC07** 1:01:56

Yeah, I mean, I I think so. I mean, fair use is one of these things, that it's so complicated that I never really quite know what.

**Sara Mannheimer** 1:02:09

I know, yeah.

**DC07** 1:02:12

Like, I'm always a little bit leery of making Fair Use claims just because I know that it's complicated. And, but but definitely, like, yeah, working in academia, like, it's, it's kind of like, it's in the water in a way, right. Like that, the purpose of, of using this content is, you know, for education, right?

**Sara Mannheimer** 1:02:40

Advancing knowledge, at least yeah.

**DC07** 1:02:42

Yeah. But, yeah, I feel like um, yeah.

**Sara Mannheimer** 1:02:54

This one is harder, because it's more of like a legal question.

**DC07** 1:02:57

Yeah. Yeah. That's why I'm struggling with it. But, but definitely, I mean, I think you already mentioned it, like the, like Twitter's own language around, you know, content creators owning their content. They don't actually say they own the copyright. I think because they're, it's implied, though, I think that there's been I think there have been some court cases about whether tweets are copyrightable or not.

**Sara Mannheimer** 1:03:27

Right.

**DC07** 1:03:28

And I remember there being some disagreement about it, but but from my perspective, like, the person who created the content, you know, in my mind, they they have they, own-, they have copyright, you know, they have intellectual property. That's their intellectual property. And they're granting Twitter a license to use it in a very particular way. Or not in a very particular way, in lots of different ways. But, yeah, but, uh, and then, and then Twitter is then granting, third parties access, you know, to that intellectual property. And so it's quite complicated. And, yeah, we didn't do it for the [Twitter hashtag] dataset, because there's so many users, but for other projects, other projects and other tools that we're building, we're kind of building with the idea that the the content creators, people created the tweets, they have the intellectual property, and if they want, like if they if we talk to them, and they say, "Yes, I want your or Yes, I'll allow you to have my content in an archive." And that's something that they can do, independent of what Twitter things or not right, like.

**Sara Mannheimer** 1:04:47

Oh, so are you doing some more archiving that's like specific to certain people.

**DC07** 1:04:56

Yeah, yeah. So I mean, so the we haven't done tons of it yet. But so the, we're building some other tools that allow people to collect the data from the API, but then also look through it and basically identify things that they would like to archive. And then to talk to those people about wanting to archive those pieces of content. And for them to say, "Yes, I, I consent to have those." And so the idea there is that then, if they've consented to having their content archived in this way, then then then we can we can archive it, right, like we can actually collect it and archive it. Well, and but it does, I think it speaks to your, the base of your research question, right of like, sourcing where, where maybe, you know, more quantitative data collection, from APIs and things like that, where that fits into qualitative work, because we're kind of treating it as like, at the end of the day, it's really a qualitative assessment, right? Like, like, oh, this is valuable to archive this is video that's part of this tweet, or this image or this conversation that happened or whatever. And, and that's a, like a qualitative assessment, right? That's done by us. We're like looking at the data, and then interacting with that person. And, but, but the API data allows it that to happen, though, because it like it's almost like you, it'd be difficult to know who to talk to, unless you sort of had had the data to be able to analyze it in particular ways. Like, you know, we collected this hashtag like, what, what users were retweeted the most that were that were mentioning that hashtag, and then being able to, to see the try to talk to some of them.

**Sara Mannheimer** 1:07:09

Yeah.

**DC07** 1:07:10

Yeah.

**Sara Mannheimer** 1:07:12

Okay, great. Well, let's wrap it up. Um, I do, do you have ideas about anyone else, I should interview anyone else who's like doing social media research or social media archiving.

[Discussion of who else Sara should interview]

**Sara Mannheimer** 1:10:09

Very helpful. Thank you so much. I'm gonna let you go. I'm sorry that we've got a little over. This is...

**DC07** 1:10:15

I talk slowly so I apologize.

**Sara Mannheimer** 1:10:18

No, not at all. All right, well, yeah, I just really, really appreciate it and I'll let you know when articles come out.

**DC07** 1:10:27

Yeah, I'm really interested to read it.

**Sara Mannheimer** 1:10:31

Cool.

**DC07** 1:10:32

Yeah.

**Sara Mannheimer** 1:10:32

All right. Have a great day.

**DC07** 1:10:34

Yeah, you too.

**Sara Mannheimer** 1:10:35

Bye.