DC08\_transcript\_deidentified

**SUMMARY KEYWORDS**

data, tweet, twitter, people, archive, users, social media, study, researchers, api, thinking, survey, qualitative data, individual, metadata, context, question

**SPEAKERS**

Sara Mannheimer, DC08

**Sara Mannheimer** 01:28

Okay, so the idea the research that I'm doing is looking at the similarities between big social data like social media and blogs, and archived qualitative data like interview transcripts, field notebooks, and diaries. I see similarities between the two. But I also see that their respective communities of practice are under connected. And so I had the idea behind this research is, I'm interviewing qualitative researchers, I'm interviewing big social data researchers and I'm interviewing curators who have worked with both types of data to train identify some strategies. Maybe from qualitative data, that's data curation for qualitative data has been pretty well established over time. And the issues are well established. Trying to connect those two, the more sort of emerging strategies that we have for social media data. And also thinking potentially there might be benefits the other way thinking about scaling up qualitative data reuse. So I, the ultimate goal is to help data curators, though, so like my community better handle qualitative and big social data responsibly, to support ethical, epistemological and legal data sharing. So I've done a review of the literature where I identified six key issues that are in common between the two types of research. And so the interview is structured with a question for each of those six, it's context, data quality, data comparability, informed consent, privacy and confidentiality and intellectual property. So I'll ask a few introductory questions. I'll ask a question about each of those issues. And then we'll wrap up so it should take about an hour from now.

**DC08** 03:36

Yeah.

**Sara Mannheimer** 03:39

All right. So tell me a little about like the data you generally work with and the types of research that you do.

**DC08** 03:58

Sure. So I'll tell you both as an archive director and as a researcher, so I'll start with the archive director. So I direct [a program at an archive, and I also work with archiving social media data and social media archiving infrastructure]. And sort of how we answer questions about what is ethically and scientifically responsible for caring for that data. We're not done with those answers yet. But my job is to get answers. So that we can build infrastructure. As a researcher, [some details redacted]. I study political communication and social media, and I'm normally studying elected officials. So I pointed this out, because the people who I study have different expectations about what will happen to their speech. And users have different expectations about what will happen to their speech. So when I am sharing data that I have collected, it's generally from public figures, not from private citizens. For the most part, the I have a new project though, where I studied white supremacists, and they're, some of them are public officials, and some of them are private citizens. And so how I develop data sharing policies for those data, I'm not sure yet. So I have things like the history of [name redacted] a white nationalist discussion board that we grab data from, or from [other far right social media sites], in addition to Twitter, Facebook, Reddit, Instagram, Twitch, and Tumblr. Does that give you at least some space about what we're talking about? Okay.

**Sara Mannheimer** 06:49

For sure. Yeah, this is great. Okay, so the next for the rest of the interview, I want to structure it around, I want to choose at least one sort of specific example, somewhat request specific example of a time when you either well in for you, I might ask you to talk a little more about a bunch of things together. Like often people just have one of these areas that they're hot they do. So qualitative data sharing big social data for sharing, or working or doing your own research with social media or qualitative data? Yeah, so it helps to have a couple of specific examples that you go to so that we're not talking really generally, you know.

**DC08** 07:34

Yeah. Um, so let's see. So I have an example of data that's in our archive, which would be [a dataset of tweets from a hashtag that was deposited by researchers at a university]. And I have my own data, there's one called [redacted], and that's a collection of [tweets related to U.S. elected officials]. And then on the qualitative side, we have, um, what if we go with the qualitative responses to [a national U.S. survey]? Okay, so that is not about big social data, but it definitely has it hits on all of the issues that you're going to raise about qualitative data. So the open ended responses to the [survey], which is the one of the most popular studies at [the archive].

**Sara Mannheimer** 08:24

Okay, nice.

**DC08** 08:25

So I'll try to stick to those three examples. And I'll try to point out which example am I talking about when I answer.

**Sara Mannheimer** 08:30

Perfect, perfect, and you can go to other examples as well. But it's just nice to have something in mind as we work through.

**DC08** 08:37

Sure. And all of these are [available online]. So you can see their metadata records, if you care. And I can send them to you, if you remind me, I can get you the real links.

**Sara Mannheimer** 08:50

That's okay. I can find them. Thank you, if not I'll ask. Okay. So were these examples, part of grant funded projects or like projects that required specific treatment of the data, like did you have, did they have data management plans that you had to follow?

**DC08** 09:09

Yeah, um, so the only one of the ways that I collected the data from is the [U.S. elected officials' tweets] one, and I'm thinking about how I want to answer this. It did not have a specific funder. I used my internal funds from the university to collect and archive that data. The [U.S. survey] was funded by the [a national funder]. And the [twitter hashtag] data set. I don't think that was part of a grant but we'd have to look, it's definitely a project that [name redacted], who's a PhD student at [university], and [name redacted], who's a faculty member did, but I don't, I don't remember if they had specific funding or not. So for the [U.S. elected official tweets] dataset, there wasn't anybody else with whom I had to establish expectations for data sharing or preservation. The, again everybody in there was [an active elected official] and even Twitter recognizes that our responsibilities to those users are different. So for instance, I don't delete tweets that were then deleted. And we kept tweets from accounts that are no longer in use. Because they were public statements by public figures. So they, they don't go away.

**Sara Mannheimer** 10:34

Yeah, and the...

**DC08** 10:38

Sorry.

**Sara Mannheimer** 10:38

On the Twitter's Terms of Service that it says, public figure?

**DC08** 10:42

So the Twitter Terms of Service say that if you know that a tweet has been deleted, then you need to delete it from your records. But I'll say this as both an archivist and a researcher, my responsibility is not to Twitter. So I don't actually follow Twitter's Terms of Service. I try to follow ethical guidelines for my communities of practice instead. And sometimes they're in line and sometimes they're not. But what's best for Twitter is not what's best for science or society. So that's not my problem.

**Sara Mannheimer** 11:11

Okay, interesting. Cool. All right. Well...

**DC08** 11:22

But I've said it enough now that it's time for me to just own it.

**Sara Mannheimer** 11:25

Oh, it's very interesting. This is a huge question, of course, as I've been talking through this with other people. So I will talk more about that when we talk about like, intellectual property and some of these other issues. So sure. Um okay. And so they're all published [in repositories], you said?

**DC08** 11:46

Yep. Both the [twitter hashtag] set, and the [survey] are restricted use. So you have to apply for access. The [elected officials tweets] data is [openly available], which means I deposited it, and it's available for use in the form that I deposited it. So it was not curated by professionals. It was prepared by some graduate students.

**Sara Mannheimer** 12:11

Okay. All right, great. So we'll jump into the six issues. The first one is context. So I have two quotes to sort of like blurbs, talking about how I'm thinking about context. So for qualitative data. "The idea is that qualitative research is a process that may include deep and prolonged contact and connection with research subjects, trying to understand the subjects within their own broader context as people in society. And qualitative data are therefore highly context dependent. Context is a source of data meaning and understanding and ignoring context, under using it or not recognizing one's own context driven perspective, could result in incomplete or missed meaning and a misunderstanding of human phenomena." And then for big social data, "When we collect data from social media platforms, just as when we collect data in traditional spaces, context matters. But the context of a social media post may be absent or difficult to understand, because social media posts are by nature, short pieces of text taken from a larger context of personal and public life. And then this out of context effect can be compounded when data are masked at a large scale." So that's sort of like what I mean by context. And so during your three examples, or one, if there's one that's more relevant than others, what challenges did you encounter, if any, when you were trying to capture the context in which from which the data was collected? And then what strategies did you use to sort of document and help future users understand that context?

**DC08** 13:54

Sure. So the... let's talk about the data that we collected. So this would be the [elected officials tweets]. So in that case, we had we collected the data by using the Twitter API, and we did follow the rules and collecting the data. Their rules I should say. And then we had to say we had a larger dataset, and then the [elected officials tweets are] actually a subset of that whole collection. So we have the [approximately one million] tweets that [elected officials] sent during [a specific timeframe]. [A few more details redacted]. But we used metadata that's provided by Twitter to indicate whether or not they're retweets. So there are two ways that retweets can be indicated. Well, three, really, they can start with an RT, they can start with a QT or an MT or there can be an indicator key value pair in the metadata from Twitter that says is retweet and it and that changed actually have to look up when exactly but so those text based approaches, the user supplies, the RT, QT, MT at the beginning of the tweet. Those are still available and have always been possible. The API returning a retweet has only been possible since Twitter introduced the retweet button. And they have actually now introduced this quote tweet button. And they've changed the functionality of retweets. So what that field in their API represents has changed as different versions of the API and different versions of Twitter, the software have been released. I don't think that we indicated that at all, actually, in the dataset. I think that we explained how we define... what a retweet was. But we didn't say and Twitter has changed this. We do make note that during that collection period, Twitter also changed the length of a tweet. So it was 140 characters until 2017. And then it became 280. So if you the maximum number of characters that a tweet could contain changed during that time. So if you're looking at something like length, you have to sort of complicate your model a little bit to account for exposure or potential length where the potential length changed. I don't think we indicate in the paper that we wrote that that's true. I don't know that we did in the data set. Yeah, we also there are other things about Twitter that have changed since that time that we don't indicate, we do talk about how Twitter has, each Twitter account has an ID, which is a numeric string, generated by Twitter, and then it has a handle which is user defined, and handles users can change handles, but they cannot change IDs. So for instance, my handle is [Handle 1], and my ID is some string some integer, and I can change my handle to be [Handle 2], but my ID would stay the same. But if I did that, then I could start a new account called [Handle 1], and it would have a new ID.

**Sara Mannheimer** 17:10

Yeah.

**DC08** 17:11

So connecting the handles and ideas to individuals is a challenge. And when [politicians] do things like go from being candidates to being elected officials, or [change political positions], sometimes their IDs stayed the same. And sometimes their handles and IDs both change. And so we do have to explain why do we have more people in this data set. And it's because people come in and out. We also... the raw tweets themselves are not that useful for studying political communication. So we needed to be able to connect their Twitter selves with their selves elsewhere. Luckily, [this group of elected officials] has been studied for a long time. And so there are entities that generate unique identifiers for [this group]. So all we had to do was to connect these semi unique identifiers from Twitter, to the unique identifiers from [standardized studies and databases]. And so these others are longstanding studies that, that provide identifier, so we connect to their identifier, and then we provide some basic information as metadata, so that could be gender, party, district, etc, that we couldn't do for just public users, because we don't have unique identifiers for them. But we do for [these elected officials]. So we were able to capture some of that, and then provide that as metadata in our own study, because it's hard to go get it. And so we just made it available for people. So every tweet comes with all this metadata about the author of the tweet. Okay, what else was in your parentheses here? So what do we use to document it for future users? The documentation for using this data within our team is extensive. And the documentation for using this data outside our team is like a paragraph. So I mean, it's a little embarrassing for me as an archive director to admit that, but we do make a lot of assumptions about researchers who will be reusing our data that we don't make about members of our team. So for instance, almost every semester, we get a new undergraduate student who joins the team as part of [an undergraduate research] program. I don't expect them to know anything about Twitter when they come in. And so we do talk about these things like the length of the tweets, or when did the interface change or what's an API, but for a data reuser who's not on my team, I kind of expect them to already know if they're looking for [elected officials'] tweets, and so I don't I don't leave it out to be sneaky. I do it because it's hard to document data. And so I just let them carry their assumptions and mine forward, and then we don't put very much information in there.

**Sara Mannheimer** 20:20

Hmm. Interesting. Okay.

**DC08** 20:25

Yeah. I think I answered linking. So we link to [an official political site], and to a crowdsourced information about [elected officials] on GitHub. And we also, because we use the [standardized] ID and provided you can link to lots of other datasets about [elected officials] through that identifier.

**Sara Mannheimer** 20:44

That's really nice. I think this example is like somewhat clean, about context, you know, you've got a lot more information than you normally would.

**DC08** 20:54

Yes. So like you said, so in contrast with the [Twitter hashtag] dataset, so in that data set, they collected tweets over a two week period that had the hashtag [redacted], in it. So this is a very different population than [the elected officials]. In part, because people who are using the hashtag, some number of folks who use that hashtag, we're really putting themselves at risk of backlash or harm by using that hashtag. And yes, they did use a public hashtag on a public forum. So none of these are private tweets with the hashtag, they're all public tweets with the hashtag. But a user who's participating in a large international discussion [about sensitive topics], I would expect different expectations about who will access that data and in what ways than a public figure making a statement on a public forum. And [the creators of the Twitter hashtag dataset], the depositors for that study agreed. They've also followed Terms of Use from Twitter, so they don't share the tweet, they shared the IDs of the tweets. And so if you want to use that data set, you have to apply [to the archive] for access. And then what you get if you access it is a spreadsheet with the tweet IDs, and then you have to go collect them yourself from Twitter by using the ID to query the API. And so we added some documentation to the study that says, here's how you would go about recollecting these tweets from Twitter, using Hydrate which was a tool created by the Documenting The Now project you've probably encountered. Yeah, so we, and we tested it with Hydrate, and then we were able to provide those but the one thing that happens there is that all social media data is ephemeral, especially when compared to other digital objects, like digital books, or things that have been digitized on purpose. But things that come up in these sort of political moments, or really, social moments like me to where [tweets that related to social movements], they're subject to pretty rapid deletion. Either the account or the tweet itself, or Twitter decides that it's not acceptable. And so the loss for like, that data is lossy in a way that data from [elected officials] is not lossy. And so I don't know that we put any put a note on that collection. But if somebody would have asked me, What should they expect about social media data, there's a paper that was in JASIST, a few years ago, that found somewhere between five and 30%, of data gets lost right away, basically. And so I would say expect data loss, up to about 30% for just your average collection from social media. And so you shouldn't be surprised if from these million tweets, you're only going to get 700,000 or less, and that we won't be able to accurately quantified the missingness. So we won't know why they're missing or what was in the missing set. Which missing data itself is not necessarily a problem. But if you don't know what's different about the missing data, then you have to be more careful about the inferences you draw from the available data. Yep. Yeah.

**Sara Mannheimer** 24:26

Yeah. Yeah, I mean, yeah, it makes it so that you like we don't save social media data really, for reproducibility. It's not just, that's impossible. Probably.

**DC08** 24:41

Well, we're going to try to save some for reproducibility. But it's going to require that folks be willing to share their data at collection time rather than an archive time. So you know, normally when people at least this is something that we're working on it somewhere that we would like, when you start a data collection that you plan to deposit, tell us about it, and we'll start a collection in parallel, and we won't use it, we won't release it. But to reduce some of that loss, but it's hard to know, because when you're talking about ephemeral data and digital objects, especially you don't always know what you wanted or what you needed until it's actually too late to get it again. Yeah, and so this isn't something that [the archive] is gonna be able to solve ourselves, we're going to need researchers and community members and other stakeholders to be willing to deposit data also. And then if [the archive] is engaged in data collection, that's a radical change for an archive's normally received collections, you know, actively generate them ourselves. But that's something that we're going to try to do [where I work].

**Sara Mannheimer** 25:45

Hmm, interesting. And then you'll compare the data set that you collect with the one that the researcher submits, or it would be....?

**DC08** 25:52

Potentially, yeah, but it might also just be that they sort of people can signal to us that there is a particular conversation that may have scientific utility or analytic utility in the future. And so we may want to be collecting it. So for instance, [a social media] project from [a university], it was a [grant] funded project that [a faculty member] ran. That data can never be recollected because of the API changes that Twitter put in place. And the lossyness, and so [the faculty member] has the set. But the Twitter Terms of Service prohibit [them] from sharing the actual set. And so there's some amount of data and [they have] posted the links you can, [they've] posted the ideas, you can go get the ideas for that. But this is, it's like, I want to say it's in the billions of tweets that contain [a certain] hashtag this is scientifically and socially, incredibly interesting. But the technological infrastructure to regenerate it is really expensive. And it can't happen because Twitter deleted some of the data forever, it's just gone. Unless [this faculty member] is allowed to share it. So there's, we'd like a heads up early of things that might be of broad utility like that. So we can't do everybody's project, but things that are likely to be useful to more than one set of researchers we'd love to help with earlier in the process.

**Sara Mannheimer** 26:22

And do you get permission? So you don't get permission from Twitter for this? You just say "We're archivists, we collect things the same way as you would like archive the web," or something?

**DC08** 27:35

Yep, yeah. And actually, you can use web archiving to archive social media too. So you can generate WARC files. So there was the George Washington University Library folks, actually built a tweet collector that was used WARC files and captured the HTTP requests instead of the API requests. And so that's another way to do and talk about context. So the API removes individual tweets, and even groups of tweets from their original context of posting. And we can never capture their contents of viewing. So I don't I can't ever there's no API calls that I can make that says, show me Sara's timeline view, Thursday at noon, when this tweet was posted. So I don't know if you saw it in real time. Did you see it later? Or did you never see it? Even if you're following that person? And so the sort of the reader end of Twitter we have very little information about. But why did I talk about... Oh, the API, so you get different contexts. If you use a web archiving tool that generates and captures the HTTP request, which contains some different timestamp information and browser information and some of the sort of associated context because the HTTP request will tell you what shows up on those sidebars on Twitter, also, but an API call would not do that, you just get the data that's available in the API, so not in any situation. And so there are different times where you might want to make the WARC archive version rather than a JSON archive version. And then if you're scraping that can be different to that. You can use a scraper to get some of the timeline information, but then you're violating Twitter's Terms of Service. And if you don't have the [university] general counsel's office backing you like I do, maybe you don't want to do that. Yeah. And I mean, I've never asked my students to do that, because they're in a precarious position. If they wanted to, and it was necessary for their research, I would absolutely go to bat for them. But I'm not going to tell a contingent worker like here, go run afoul of one of the largest tech companies in the country. Yeah.

**Sara Mannheimer** 29:50

Let's move down to intellectual property actually, now that we're talking about this. So like during these examples, what other... We've talked about some of these challenges about trying to align with the terms of service. And the strategies that you've used. So you either follow the terms of service or decide that it's worth it to go against them, right? And what do you what are you feel that like you are your responsibilities around intellectual property and Twitter's like license to use the or to like, control the content? Tell me more.

**DC08** 30:30

So I think more about what my responsibilities are to the people who posted the content, and to those who can learn from its existence, so to users and to science, I do have a responsibility to Twitter. It just does not trump my other responsibilities. So when I think about what are my responsibilities to the user, I think that when a sort of average user has deleted a tweet that is innocuous and holds little analytic utility than my obligation is to follow the user's expectation that that tweet will be deleted. But if that would make science harder, so for instance, around the time of the Boston Marathon bombings, Twitter was still quite a popular way for people to respond to crises that as the social media landscape has changed, that that message has gotten diverse, but [at the time of a recent major crisis event], Twitter was your real time social media platform. And so there was each individual person who was trying to identify: where did the [event] occur? Where can people get help? [Where are there dangers? Were people searching for each other?] We have, because so many people posted the information that they had at the time, we have an opportunity to study crisis in a way that is not available for other crises that occur. So what is a crisis and community response to an event like the [this crisis event]? So our obligation there is, our obligation to science and society, I think outweighs our obligation to any one individual user. But it does mean that we need to work harder to deidentify and disconnect that data. So if the purpose of the data is to understand crises, I actually need very little information about the user. And so there's not much reason for me to keep a particular deleted record attached to the person who deleted it. In fact, it would probably be better for me to sever that connection. And so I us-, I will lose some analytic utility by making that break. But I preserve the user's privacy expectations, and reduce their risk of the identification, while making it possible to do the science of crisis that having the deletions facilitated. So each time we make a choice, it's actually a complicated decision. So there isn't one like blanket, this is what will work for so far. I do think we give pretty wide latitude to public figures. Where a public figure as I have defined, it is an elected official. So we also study [a country that is not the U.S.]. And if you're an elected official, or you're a publicly declared candidate in [that country] or the U.S., there's, I think that there's some level of celebrity after which a person becomes a public figure, but I'm not sure where that line is. And it probably depends. But I think some of those folks would get a different, we would treat those accounts differently than yours or my individual user account.

**Sara Mannheimer** 33:29

Have you ever gotten in trouble for this?

**DC08** 33:36

Yeah, not so far. So I was [at a conference] a couple of years ago, where I was talking about this, and there was a guy from Twitter who was there, he was like, "I don't think you should do that." I was like, "Well, I don't think you should work for Twitter." So I sometimes get pushback, but we've never gotten in trouble. And I'm not trying to, like I'm not trying to attack Twitter necessarily, or to reduce their ability to generate revenue. I just like... delete [a tweet] because somebody said to... is not necessarily the best use of that data. But you know, we have to be careful too, because I don't want to say, oh, we're gonna keep everything for always. And I never delete anything because users posted it publicly. But like, that's not right, either.

**Sara Mannheimer** 34:25

Yeah, I mean, I feel like just segues into some questions about informed consent. You know, it's like, if a user... like if for qualitative data, you got a user's consent, and you say, I don't exactly know how this data might be used in the future. But I know it will be used by an academic researcher, and you can provide all these terms to like how the data will be used. But with Twitter, since the users aren't maybe even aware that they're being archived, like what other types of challenges with consent did you encounter when you're like going ahead and publishing this data, even if it's for sort of, like, very regulated use within [a data repository]?

**DC08** 35:10

So I think that, um, we've done a couple of things. We're one....so when I'm publishing about politicians, I don't check with them. But when I'm publishing a paper and I quote a user who's not a public figure, I check with them about whether or not it's okay for us that, okay, if it's not, I can use a different one, like, nothing that I'm studying has only one example, which is one of the things that makes it different from qualitative research. Like, there isn't just one person who can tell me about whatever. You know, like we did a [study about how people interact with elected officials] on Twitter, the patterns of [interaction] behaviors and sort of different rhetorical strategies that people employed, were pretty robust, there were only [a few] of them. And so if my favorite example of one of those categories was not comfortable being quoted, then I could just pick a different one. And so it didn't impact my ability to communicate the results of my science, I think that is one thing that's been useful. The other is that we can do so the idea around informed consent is to balance risks and benefits to science into individuals. And so informed consent is one way to do that. But especially if we're thinking of informed consent as an ongoing process and relationship between researchers and research participants. It's not like you just sign a document, and then whatever use for always is fine. That's not what anyone means when they sign an informed consent document, right, you and I had a conversation about my expectations about how you're going to use the data and with whom. So if you were to go use my data to try to make money, I might get annoyed, because that isn't something that we actually discussed in informed consent, and I use the money motivation, because it connects to social media data. But also, if you were going to use it to study, I don't know COVID vaccine uptake. I might not object. But I would be surprised and really caught... That's an interesting use that we hadn't discussed, right. So it's not like informed consent is perfect. And so anything other than informed consent is bad. It's that we ought to be considering risks and benefits to individuals and to science when we're making these judgments. And one way that we mitigate that at [DC08's workplace], is that right now, all [big social data], we plan for it to be restricted use, meaning that you'll have to apply for it with a specific research question or topic in mind. And that has been through some level of review at your institution before you apply for the data. And then once you apply for the data, our hope is actually that, and I say hope because it, it takes funding to do it. We'd like the data to always stay at [the archive]. So we'd like folks to bring the analysis to the data. And then we'll review the analytical output for disclosure risk, just like we do with qualitative research studies that are held [in the archive]. And so that instead of reviewing all of the data on ingest, we review all of the results on download. So if you, you know want to do, one of my favorite papers using social media data is when April Williams wrote about black women tweeting during How to Get Away with Murder. It's a really fantastic paper. And I love the second screen, and the focus on black women and their identity, their sort of co-construction of identity. And so in that case, for instance, if we had that data set, and it was restricted to use, if you wanted to use it to study their identity construction, like April did, you could do that, but we wouldn't let you take out their individual identifiers. So your output would have to be the patterns and aggregations. But you don't actually need individual identities to do that science. If your science is about individuals, then we'd have to have a different conversation.

**Sara Mannheimer** 39:00

I see. Okay, so you're not just sharing all of that you wouldn't be sharing all the data wholesale, it would just be people come in, conduct their analysis, which then gets shared with them. Hmm. Yeah. Very interesting.

**DC08** 39:11

Yeah, that's two reasons where... one is this sort of privacy and confidentiality protection. The other is that it's more sort of environmentally and technologically efficient. So it doesn't make sense to make multiple copies of petabytes of data, like that's, hello, Bitcoin, this is not what we're supposed to be doing. And so if we hold the data in one place, and you bring your compute to the data, then the it's the storage costs of the data are lower, and the privacy and the risks of data loss are lower because we have it in our.. sort of enclaves. And then if the compute needs that you have are different from the compute needs that I have, you bring yours and I'll bring mine and we can right size them. So it's more efficient, technologically and environmentally for the data to stay in one place. And then for people who bring their minimum computation set to the data instead. So it's cheaper and better for the world. And less likely to have loss or leakage. And it's probably more privacy protecting. So there's so many reasons to do it.

**Sara Mannheimer** 40:26

Yeah. Hmm. Where, where did this? How...I guess... Well, that's a huge question. But like, how... Where has your thinking come from? Like, how long have you been working on these questions? And then did you just like read a bunch of people? Did you do interviews? Like, how did you develop these strategies that you're thinking through?

**DC08** 40:45

Um, so I started working on this shortly after I got to [the archive]. So [a few years ago]. So I mean, building data infrastructures takes like a decade. So I've been working on it for [less than 5 years]. COVID time is not real time. So it's like, yeah, let's call it [less than 5] years. And I started by sort of thinking about what are the technical challenges and what is technically different about social media data from surveys, or GIS data, and variable level indexing, and full text search are really important for social media data. And they're not as much for surveys. Variable level indexing is becoming increasingly important as things like GDPR would require the removal of observations rather than whole datasets. But the idea that you would index digital objects at the variable level, or the observation level is a pretty big change for archives. So that was one, like, what is technologically possible to do? So if you think about projects, like the Library of Congress, where they got the whole tweet archive, for instance, but they couldn't technologically provide for it. It's not that they didn't want to, it's that they didn't have the computational resources. And really, at the time, we didn't have the technology, the indexing or search technology to facilitate access, so why bother when they have so many other products? So that was where we started, like, technically, what could we do? And then a lot of sort of... in parallel was socially and ethically, what should we do. And that's where the move towards restricted data and data in place became, like, that's just efficiency from like, this is really expensive and huge for our archive. And so facil, we would have to upgrade our the wires coming into our building to allow people to download this data. That's probably not the best use of those resources. So we had to think about those. And then I went to [a conference], and I talked to the web archivists about, what are some of the issues that they face and how is this like that are different from....

[Brief interruption]

**DC08** 44:35

We were talking about the web archivists [at the conference]. Yeah, thinking about what those are. And basically, we move toward the JSON or items from the API, because WARC files are huge. So the it wasn't clear that social media researchers actually needed the data that came in the HTTP request. So we didn't think we were losing something that was analytically useful by changing the data collection. Because if you're studying web archiving, you do want all of that information. But social media data, it wasn't clear that it was necessary. There might be particular use cases where it will make more sense for us to get a WARC file than a JSON file. But most of the researchers who I talked to ended up putting it in a CSV anyway, and so they're losing all this metadata along the way that maybe we didn't have to collect all of it. And then [a few years ago], I'd have to look at my own paper to remind me what year we did it. But we ran a survey of researchers who use social media data in their research to try to understand their data management practices. And then this year, we're running a survey of the social media data generators, so people who use social media data and what are their expectations? So I know Casey Fiesler and Nick Profieres wrote a paper a couple of years ago about the expectations of Twitter users. I find that paper pretty frustrating from an archive perspective, because it doesn't consider ways in which social media data is similar or different from other types of public or administrative or non design data. And I'm not at all surprised that folks say no, you can't use it when you ask them about it in isolation, and so we're.... our instrument is about different types of data that are generated about an individual. And under what conditions would it be appropriate for whom to use it? So location data, social media, data property records, sort of putting it in context, the census, that that paper also assume that there's no way to protect identity once you've gotten it. And you and I just talked about [reviewing deposits for risk of personal information disclosure], and like that's.. the archives have been doing [these reviews] since there was an archive, sort of so de-identifying the data on its way in is not the only way to protect respondents.

**Sara Mannheimer** 47:06

Yeah.

**DC08** 47:07

So we're running a survey about... Yeah, so and we have two different waves, because there's some sort of folks where we want to understand as an archive about restricted data generally. And I specifically want to know about social media data. So there's a longer instrument for people who generate social media data then for other people who might be in...

**Sara Mannheimer** 47:29

Oh, can't wait to read that.

**DC08** 47:37

I can send you the instrument, if you want to help us make sure that our instrument is good?

**Sara Mannheimer** 47:40

Sure, I would love that. Cool, well, let's talk about privacy, then we'll just jump all around. During your example, or any of these three, what challenges did you encounter related to privacy for the people represented in the data? And then what strategies did you use? You have the sort of like restricted access and the analysis in place, you know, what else comes to mind?

**DC08** 48:08

So restricted access is definitely one of the ways that we do it, I think, in the [national survey discussed in the beginning of the interview], in the qualitative responses, which I have used this is, it's not an ethnography. It's not an interview study. It's just the first one I could think of, but we tried to be careful about how we masked place. So that some of [survey] data has zip codes, for instance, there are some zip codes in which there are zero [of the target survey response population], there are some zip codes in which there's one. And where those zip codes are in relation to one another matter. So for instance, [A state with a large population's] zip codes got recoded to [the name of that state]. Zip codes from [smaller states], maybe didn't get coded as [the name of the state]. Similar, similarly about cities. So the level at which we masked variables for [a small city], is different than well, how we mask variables for [a larger city], because they're different sized cities.

**Sara Mannheimer** 49:07

Yeah.

**DC08** 49:08

And we also so all data that gets archived at [at the archive where you work] goes through [a review to assess risk of reidentification]. And that is a process where the professional curators on the curation team and the project managers for the archives, and occasionally the director of the archive, have a conversation about what is best for this particular study. And in the [survey discussed above] and some other studies of [vulnerable respondent populations], we have worked with the data depositors' to make sense of what is the right level of abstraction for this data. So for instance, a bunch of the data comes from [a different research organization]. And there's a researcher [who worked with us on the survey discussed here]. But this is not just the study, but other studies [this researcher worked on], where people talk about specific locations in [in a city] that if you're not a [person from the respondent community living in that city], knowing that doesn't put anyone at risk. But if you're in [the city and you're a member of the survey respondent community, or you're a person who is prejudiced against that community], we don't want you to have that information. And so some of it depends on, we sort of even in restricted access data, we make particular variables available at different levels of aggregation for different researchers, and under different rules. Part of why we rely so heavily on restricted access at [the archive where DC08 works] is that it's a really big hammer. And this is also why [the survey depositor] chose us to deposit data because [the depositor] had been sharing the data. But they didn't have a good mechanism for if somebody violated the data sharing agreement. So like, what is [the depositor,] a group who's busy trying to [advocate for the survey respondent community], they do not have time or resources to hunt down bad actors.

**Sara Mannheimer** 50:52

And do you at [your archive]?

**DC08** 50:53

Yes.

**Sara Mannheimer** 50:53

Oh, okay.

**DC08** 50:54

Yeah, I do. And if you violate the terms of use in a way that might harm a respondent, we can shut down your whole institution. Just like if you break IRB or you misuse federal dollars, then they can shut down your whole institution. And the restricted data use agreements between [the archive] and investigators are actually institution level agreements. So if you wanted to use restricted data, it's an agreement between [our two institutions], it's not between you and me. And so it means that the penalties can also occur at the institution level. So I have bigger, more effective sticks at getting people to follow us. And, you know, that's in line with my roles and responsibilities, and it frees [the survey depositor to focus on their work].

**Sara Mannheimer** 51:45

Right. Okay, that's great. Um, do? Let's see. So it would what else besides restricted access, like if you were to publish? Like when you did the [elected officials tweets]? Did you, you publish only to tweet IDs for that one, right? Or no, you published everything?

**DC08** 52:12

We did. And then we added tweets, because I got mad about something. And it was like, just put them up. Um, we didn't, because we didn't mask anybody who got retweeted. Where like, if I were going to publish a the tweets and retweets of white supremacists, I would look through the actors that are being retweeted, so that I wasn't exposing somebody to harm. Because often somebody will say, in their Twitter bio or something, and that retweet doesn't mean endorsement. For white supremacists retweet often means attack, that's a signal to their white supremacists buddies that they ought to come to this place. And so I would do more masking of the original accounts. For [the elected officials], I didn't do that. But I don't expect much. I don't expect people to use that data a ton. So I don't think that it's putting folks at new risks by aggregating it. And then confidentiality. Um, I don't know that we did. I mean, we didn't. I don't think we followed Twitter's terms. And we connected it with data about other sources, like we linked it, because that's another way that you can protect some confidentiality is to restrict the data linkages.

**Sara Mannheimer** 53:29

Yeah. Uh huh.

**DC08** 53:31

So it might say you can have this raw data, but you can't connect it with any other data set. And we might remove individual variables, or we might mask the variables. So like, we might roll up zip code to state for everything? It just sort of depends on what the data is.

**Sara Mannheimer** 53:49

Yeah, I guess I'm hearing a lot that like each individual study just requires very different treatment, you look at each one and think about all of the different factors and who the people are and what they might expect.

**DC08** 54:06

So yeah, I do think there's one difference that comes up. So as you've said here about like, what's the same, protecting privacy for qualitative data or big social data? Something that's come up for us is that informed consent is about protecting an individual. But often, when you're looking at something like big social data, our responsibilities are to communities or to groups. But groups can't provide informed consent. So that's another reason to say that informed consent is not enough. It's an indicator of appropriate use of data, but it's not the rule. And yet the [national survey] is actually a good example of this, that the risks are not just to individual people who filled out the survey. The risks are to [the respondent community], generally, that if somebody makes an inference about [the respondent community] from that survey, that can perpetuate bias and mistreatment of [that community] everywhere. Even if they didn't fill out the survey. And so when we're thinking about confidentiality and who our responsibilities are to. It's not just the individuals contained in the survey. And this is coming, I think we have ready partners in Indigenous archives, and Indigenous Studies and sort of feminist data approaches to so this isn't just [the archive where DC08 works], but sort of archives thinking generally about to whom do we have the responsibilities and when, and that does change depending on the data and the depositor and the use that we need to I'm more interested in who ought to have power in this situation where confidentiality is about power, and sort of who has it and who, who can use it. And so if we think about it that way, then it gets easier to say, you know, public figures have tremendous power. And so exposing their public language isn't going to put them at risk. That's a very different calculation than the people who posted [a social justice-related hashtag].

**Sara Mannheimer** 56:04

Yeah.

**DC08** 56:05

And so, right, that it comes down to a confidentiality decision, but it's driven by sort of feminist power question.

**Sara Mannheimer** 56:13

Nice. Oh, this is so great. Thank you. It's so fun to talk to you. I feel like we're like jumping all over the place. And um...

**DC08** 56:24

We're not following your list, but we're gonna get them all.

**Sara Mannheimer** 56:26

Oh, yeah. Okay, I'm gonna go back up to data quality. The second question. So, yep. When you collect social media data, and when you think about making the survey data available.

**DC08** 56:47

Yeah.

**Sara Mannheimer** 56:48

Did you encounter any challenges when you were trying to document data quality, like missing data? Or like for social media, data bots, or thinking not really about, like, quality control, but more like, how do you communicate, if there's a problem with quality, you know, or communicate to the people that your data is trustworthy? You know?

**DC08** 57:10

Yeah. So I think that, um, the, the survey is a little easier, because people tend to trust institutions more than they trust individual PIs. And so we made sure to list [the depositor - a national organization]. So like, [Name] is the PI for that study. But [Name] doesn't work at [the organization] anymore, even. So I think that is one way that we tried to indicate quality was by making sure that they're attached to the record. And then for our data, we tried to describe our collection process, so that people know how we get it. But again, this is... did we actually list them? I don't, I'm looking at the transcript right now. Yeah, the [organization] is listed. So the, the quality of our data depends on the collection mechanism. And so we tried to describe what we did, and which API's we used and which lists of users we used. So, but I don't know that we were explicit about what that means. We don't have I think we mentioned that we don't have campaign accounts, for instance, that we tried to focus on the official accounts that are taxpayer funded. But we didn't get into what does it mean to have a taxpayer funded social media account? Which that's a whole different ball of wax that gets into like franking rules and the ethics of communication that are set by the [U.S. government]? That again, I would expect somebody who's studying [the Twitter presence of elected officials] to have some knowledge about that. But I've worked with computer scientists long enough to know that that's not true. And so I should probably update my assumptions and my metadata appropriately.

**Sara Mannheimer** 58:57

Yeah.

**DC08** 58:58

But updating metadata is a real pain. So I just haven't done it.

**Sara Mannheimer** 59:02

Yeah. Yeah, tell me more about this idea of like, who would be the future users? Like as we begin to share data more, do you think that we'll have to like add additional metadata? Thinking that the users might not always just be other social scientists, but there might be like a broader use base?

**DC08** 59:25

Yeah, I do. I think we're also going to have to do some descriptive statistics ahead of time, because a lot of potential users like a journalist or grant writing nonprofit don't actually need the data. They just need summaries of the data. And that we as an institution, the archive could be more helpful by running some of those descriptive stats and making those available instead of just making raw data available. And so that's thinking about what would be useful for users of these digital records. As do we need to add metadata. I think that I have a couple of colleagues who have, when I asked them to talk to me about what they would want to see in [an archive of big social data], they think I mean archives of the software or of the platforms. So that's been really interesting and thinking about like, what is Facebook version A? How is that different from Facebook version now or like Facebook circa 2021, versus Facebook circa 2012? And what do those mean, and what is what was possible or afforded by the platforms at those times. So we talked about the extension of the tweet link. But if you look at someplace like Facebook, that they've had a bunch of different policy changes and infrastructure changes. And then there are some things that change on the front end, like now you can indicate, for a while it was like an open ended gender indicator. But then there was a category of like 50, the first one had 53, or 54 different genders. But on the back end, it was still coded as male or not male. It didn't matter what you put, like presentation layer and the API layer were not the same. And that's true still, for like geographies that might say that I'm in [City, State], but the API shows the geolocation of my house. So you know, at which layer, are you grabbing data? Yeah. I forget what what I'm answering in that question, with that...

**Sara Mannheimer** 1:01:28

It was like, what's most....

**DC08** 1:01:30

Oh! Data quality.

**Sara Mannheimer** 1:01:31

Yeah.

**DC08** 1:01:32

Yeah. And like I said earlier about the missingness, we there's a lot of social media data missing, that we knew that it existed. And now we know that it doesn't. So that's one kind of missing data. And we can make some guesses about what might have been in that data based on what we have. What we don't have is all of the people who don't use social media. And so social media is not a stand in for the general population. And we don't actually know which some populations that is a good stand in for because we just don't have really good user data. And so I think, thinking about data quality, all social media data is low quality for some questions. Yeah. And so the data quality depends some on what do you want to know. And then there are other questions for which it might seem like it's low quality, but it's actually high quality. So for instance, [a researcher] used social media data to study unemployment. And you might think, well, what is social media going to tell us about unemployment, but it actually is a leading indicator for the jobs report. And so it can be useful for things. Like the Google flu tracker that is based on search, plus social media data, that you might think social media data is not very useful for that. But it can actually help us figure out where do we need to put our limited public health resources to stop pandemics, effectively.

**Sara Mannheimer** 1:03:07

When you think about like, risk and reward, it's like, for public health. And using these resources, it's like, you don't exactly need the data to be super, like accurate, either. It's like, you know, as long as we're somewhat informed, it's better than not informed at all.

**DC08** 1:03:27

Yeah. I think the same thing was true for... but then I'm, like, if you're thinking about the flu tracker, right, it is states that make those allocations. So you need to know: is the flu going to be worse in Minnesota than it is in Wisconsin. And then that is useful enough. You don't need to know necessarily where in Minnesota because your public health officials can figure that out. But knowing if you have one federal dollar, you send it to Minnesota or Wisconsin, you do need to know. But then there was an example where the New York City, whatever department controls food, and food safety was using social media. There was a I forget who the project was with, but they were trying to figure out whether or not they could use social media to predict foodborne illnesses. And there you do need really specific location information because you need to know whether it was the bodega or the deli in this block, that was the source of the foodborne illness. And it doesn't mean like economically, if you can just shut down one, that's a big difference for that neighbor. And definitely for those individuals, then if you can just shut that one and get it cleaned up and go back to business. And so for the same kind of thing, like a public health issue, depending on the question, you might need data of a higher granularity and better quality because if you miss and you shut down every deli on the block, that's bad news for that neighborhood. So it again, it's like, who needs it? But it's okay sometimes for data to be really messy and inaccurate because the accuracy won't actually impact your analysis.

**Sara Mannheimer** 1:05:04

Yep. Yeah. Nice. Okay, let's move today to comparability. That's our last question. What challenges did you encounter relating to comparability or interoperability of your data sets? And then what strategies have you used to address the challenges, you have that good example with using the [standardized IDs] and all of that data on [elected officials], any other strategies you've used.

**DC08** 1:05:32

So there's... Twitter actually makes a big problem for this. But then Twitter ID that it assigns to a tweet is an integer larger than Excel can handle. So as soon as you move from JSON to a CSV, if you open that in Excel, and I think this is true in numbers as well, I'd have to check in OpenOffice, it rounds the number and it is no longer unique. Because the integer is too long for what Excel can say. So if you are migrating your data from one format to another, and at some step, you lose, this might be a data quality issue too. But that has never happened to me in a survey., I have never had an integer so long that Excel rounds it off. But that's true of every tweet generated after like 2015, they're all too long until they all get rounded off, until you can no longer connect them back to themselves in the original data set, or to other datasets that may contain that same tweet ID. And so if you're trying to link data, that can be really challenging. And that makes it harder for it to be interoperable. By changing format to be more accessible and human legible, it actually reduces the quality of the data and the level of aggregation at which you can work. What else may be true. So a couple of years ago, I did a study... I studied [a vulnerable community's tweeting activities]. And they were doing some things around identity construction, and ended up never publishing the paper because it felt too intimate. And so that also happened with a study that I did of postdocs. So when I was a postdoc, my postdoc study was of other postdocs. And it ended up that many of those interviews were therapy sessions more than data collection opportunities. And so I didn't actually publish very much from that data collection, because it was clear that the postdocs who talked to me were talking to me about our shared experiences. And that it didn't feel... it didn't feel okay for me to use many of their stories, even if I had tried to deidentify them, because they had pain and oppression and exploitation. And it just, it didn't feel right. And I talked to them about it. Like what if I don't publish from this. And because often our responsibilities are if we tell somebody that we're going to share their data for it to have increased utility, then we better share it. Or you can't lie to your participants-- be like, you know, we're gonna, if you participate in this controlled trial, I'm going to share my data and we're going to cure cancer. If I don't share my data and try to cure cancer, then I didn't follow our shared agreement. And so with the postdocs, similarly, when I went back to them, I was like, here's my read on this data. I think we started the conversation, assuming this is going to be for science, but that's not really where it went. Are you going to be comfortable if we don't publish this? And they said that they were like, "Oh, I'm glad that you talked to me, because I felt weird after that conversation, but I wasn't sure what to do." So we ended up using a bunch of it to advocate for postdocs at different institutions and within the Postdoc Association, but not as an academic paper. And so, you know, you go back to your participants, and you talk about what is reasonable for this use? And how did our expectations change, as we sort of went through this data collection process? So I think that that was comparable, where this interview study of a vulnerable population, and a social media study of a vulnerable population where in both places, I thought I learned something interesting, but I didn't think that it was so interesting that was worth putting those populations of folks at risk. So the risk to any one of those postdocs or any one of the [people tweeting in the first example] was pretty low. But to those communities, it wasn't like the risks weren't real to those communities. And so it wasn't worth publishing about. And I think that that is how I would think about what's comparable between the two is that you're still making these decisions about what is an appropriate use of this data. And what is the right balance of risks and harms to individuals and communities and benefits to science that we get from this were like, just knowing something you didn't know before isn't good enough. It has to also have some return for the folks who put themselves at risk.

**Sara Mannheimer** 1:10:16

Yeah, that's yeah. Great. Okay, I'm gonna stop. We're at the end of our time, this has been so great.

[Discussion of who else Sara should interview]

**DC08** 1:12:06

Thanks. Okay. Thanks, Sara.

**Sara Mannheimer** 1:12:08

Bye.