BSR05_transcript_deidentified

SUMMARY KEYWORDS
tweets, data, twitter, people, collected, api, research, conversations, question, big, paper, irb, set, users, context, weighing, disclosures, sexual violence

SPEAKERS
Sara Mannheimer, BSR05

Sara Mannheimer 00:04
Okay, so my name is Sara. I'm a data librarian at Montana State University. And I'm also getting my PhD from Humboldt University in Berlin. And I'm working on this research that the idea is that big social data (like social media posts, and blogs, and photos and videos) and archived qualitative data are similar because they're both presented online for reuse or and when you when you do research, you reuse them. And so I am trying to connect these two communities of practice through their data curation practices. So the idea is that ultimately, we data librarians and repository managers will be able to support this research, more research support the responsible creation of this research, and maybe data sharing as well. So I have conducted a lit review that were identified these six key issues that are present in both big social data and qualitative data reuse. And so the interview is structured around those six issues. So I have like an introductory question six questions about your practices. And then one wrap up question. So it should take about an hour, maybe 75 minutes, depending on how much we talk. Any questions before we begin?

BSR05 01:34
Uh, no I think it makes sense.

Sara Mannheimer 01:36
Okay, cool. So tell me about you the type of research you do, and then the type of data that you either collect or produce.

BSR05 01:44
Yep. Right. I'm a PhD candidate at [university, program]. I focus on online communication networks, and a lot of online political communication. So things related to hashtag activism and polarization and misinformation and all that fun stuff. And so a lot of my, or, you know, almost all the data I work with is some kind of social media data, basically. And most of what I have done, and currently do is with Twitter data specifically, I've done a lot with Twitter data. I've touched other social media data like Reddit, I've been playing a little bit with TikTok. I've looked at the documentation for all sorts of other ones, you know, Facebook, and YouTube and things like that. But most of what I do is Twitter. And so what that looks like is a lot of the times I've worked with data that was collected by someone else at our institution. So someone else who is running something like a decahose feed, where it's just 10%, of all tweets, or something like that. And so I've worked with that. I've also worked with, you know, data that I pulled myself by keywords or something like that, or pulling user timelines directly from the API. I think I've rehydrated tweets. So I've used other people's tweet IDs to get those myself. And then there's another group I work with maintains a panel of Twitter users where they continuously collect timelines for an effort. So different forms of Twitter data.

Sara Mannheimer 03:49
Great, perfect. So I asked you to identify one specific example, when you were collecting data or reusing data that was collected by someone else or preparing your own data for publication or sharing. Can you tell me about a recent example? And we can go beyond the example but it's helpful to have something specific as we talk through the the issues.

BSR05 04:16
Yeah. So we did a project a few years ago, one or two years ago on Me Too. And so this was all Twitter data. And so what that looked like was—I actually met someone who is a friend and colleague of mine now who was at a different institution, and their institution had access to like, all tweets somehow, like they paid like a really big subscription to be able to have really full access. And at one point he had collected the Me Too tweets. So, for the period that was kind of immediately following when it first started taking off, so he had something like the first three months or something like that. And he told me about this, and he was able to share the tweet IDs with me. And so I rehydrated that data. And we ended up using that data for a project on looking at how people disclosed early in the hashtag campaign, and how that may have produced stigma around kind of disclosing for other women to disclose experiencing sexual violence. So it was kind of a sensitive data set. And so we worked with... a kind of standard, just keyword data set, so it was like tweets with Me Too or something like that. And we want to go publish it. I like the idea of, you know, trying to share data, it seems like an important data set, because it is Me Too. But I was concerned because, you know, because it is like a data set about sexual violence, basically. And a big part of our paper was we came up with a way of identifying which tweets were disclosures which were not, and we have to describe the method in the paper. And more or less, that's a method to identify survivors of sexual violence in our in our data set. And so we we talked about this among ourselves. The team I was on, which was my advisor, and then another PhD Student and [their] advisor. And what we ended up doing was, we were aware, you know, we can just put the tweet IDs out there. That's fine, with Twitter terms of service. But what we did was we ended up hosting the data through [data repository], and we were able to get kind of a—they will do like a restricted use agreement that people have to sign to do that. And so I think that involves proof of IRB, or exemption, and they have to sign like that they they won't kind of misuse it and stuff like that. So that kind of barrier to using the tweets felt like a fairly strong deterrent, or strong enough, given the work, you still have to do after that to do anything bad with it.

Sara Mannheimer 07:58
So you've published just the tweet IDs under restricted access in [data repository]. Nice.

BSR05 08:03
Yeah. So that was, that was kind of the full story of the data that we worked with. There's a portion in there where we also bought just a little bit of missing data from Twitter.

Sara Mannheimer 08:18
So you used.. or, your collaborator used the Twitter API to pull down tweets with the Me Too hashtag and then you've got some additional data. Okay. And was the project grant funded? Like, so, did you have like a data management plan or anything that you needed to follow?

BSR05 08:39
Um, I think it's funded by an internal grant, but no one kind of told me I had to follow any particular guidelines with it. So it mostly just had it on our lab computer, which, at the time only I had access to, and that I had set that up to be as secure as I could, but yeah.

Sara Mannheimer 09:06
And let's see, Do you have plans for storing, retaining and deleting the data in the future?

BSR05 09:15
I still have the Me Too data somewhere. I think it's on backed up hard drives at this point. It's actually not on the lab computer anymore, I think because I had to reset the whole computer. But I do have a backup of the tweets themselves somewhere. I probably wasn't planning on deleting them anytime soon, because it's a useful data set that we come back to so I've actually used that data set again in a later project. A case study. Yeah, the, the tweet IDs are kind of posted for perpetuity.

Sara Mannheimer 09:56
Perfect. Okay. So we'll start with the six issues which are context, data quality, data comparability, informed consent, privacy, and intellectual property. And so with context, I have a little quote to help you to help communicate what I mean by this. So Halavais suggests that 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 can be absent or difficult to understand, because posts are by nature, short pieces of text, images, videos, etc, that are taken from the larger context of a personal and public life. And then the out of context effect can be compounded when data are amassed at a large scale. So coming from this idea of context, can you tell me about a time, if any, during the process of your research with the Me Too Twitter data, when you considered the issue of maintaining and understanding the data's context? So like context about the community where the data was collected, or context about the Twitter users themselves?

BSR05 11:13
Yeah. Yeah. So we did a mixed method study. And so what that looked like was, we did a big inductive qualitative analysis where we identified themes in the Me Too tweets. But a big part of that was your like, the broadest way you can sort the tweets for those that are disclosures and those that are not. So there's a whole set of those tweets that are disclosures, and so we read 1000—1000-plus tweets that were—

Sara Mannheimer 11:47
Oh, just manually, okay.

BSR05 11:49
Yeah, yeah. So maybe we sampled 2500, or something like, is about half of them that were disclosures in the end, and is about 60% of those that, like were stories of some sort, rather than just the hashtag Me Too. And so I, I was glad we read those, because you really got a sense for what happened at that moment, and who is talking. But there was there were some tweets, sometimes where you, like, to understand you had to go... Like you couldn't tell just by the tweet text, like you had to go see, like, who was saying this? Some of them are a little ambiguous, like, are they mocking Me Too? Are they serious? Like, what is it? Is this person disclosing or is this person just making a general comment? And sometimes it's hard because sometimes it was profiles to more anonymous, or they weren't established figures or anything like that. And we come back to this question sometimes, like who those people are in terms of what communities they're coming from, because you didn't have you don't have demographic data for those folks. So it's kind of a, almost like a sleight of hand that we do in the paper, the whole paper is framed around women, but we never actually like, check who was a woman or not, it's just kind of going off the assumption that most people are disclosing more women. And so that's something we kind of think about sometimes is the fact that we don't, you know, we don't know what

communities were underrepresented either in those disclosures. And, yeah, and then, I guess, just maybe this is more technical, like the context of like, when you're like doing qualitative analysis, and you're reading the tweet. And it's just like a text box in front of you versus like seeing the tweet with the picture in it. And, like the username and the person, like, there's a lot of times when it didn't become clear until you went to the tweet itself, versus just like in a spreadsheet or something like that.

Sara Mannheimer 13:59
Right. And when you were publishing the data in [data repository], did you do anything—like did you address this in the description that you put there? Or is most of the way you address this through the paper?

BSR05 14:17
Yeah, I think in terms of... so we did with our, with the Tweet IDs, like we didn't really say anything about like, which we'd identified as disclosures, anything like that. So in that sense, that context, isn't there. The only context so it's really there in the description is just, you know, what kinds of tweets they were and those were the ones containing Me Too the what period they cover. And, you know, anything we knew about any other kind of those like details of how we sample, but nothing really more about those communities that they might be coming from.

Sara Mannheimer 14:59
Um, Do you think that your research was affected by the sort of fuzzy context that happened here? And then what strategies did you use to sort of make up for that in your research?

BSR05 15:23
Yeah, I think I think just generally, doing the mixed methods, where the qualitative proceeded the quant[itative] really helped us understand the data better. And I could see the project having gone down like the wrong route without the qualitative analysis, because it really gave us a sense of what was going on in the data, because we had to look at so much of it.

Sara Mannheimer 15:52
What was the quantitative analysis part? What did you do next?

BSR05 15:56
So the quant part was when we, so we had the tweets, labeled as disclosures and not disclosures. So we used that as a training set to build a predictive classifier for the rest of the tweets. And then most of the rest was a network analysis.

Sara Mannheimer 16:15
Okay. Cool. All right. I think that's good. Let's move to data quality. So can you tell me a time, if any, during the process of this research, when you considered the issue of data quality? So you mentioned, like not knowing whether the tweet was making fun of Me Too or actually disclosing? That's one example. Or bots or another example, or missing data? I guess you did talk about buying some missing data from Twitter, too. But are there other examples when you took that into account?

BSR05 16:55
Yeah, there's actually so when my friend originally collected the data... so, Me Too was October 2017. So I think [my friend] collected it in February or March 2018. So already, that was a few months. And then I met [my friend] the summer of 2018. And then I don't think I rehydrated the tweets until December of 2018. So by the time I rehydrated them might have been over a year since Me Too first

happened. And so even comparing against [my friend's] tweet IDs, so I had all the tweet IDs and then I had hydrated some of them. I think we were missing about 25% of them. And it's because either they had been taken down or probably more likely they'd been deleted. And so I always kind of wonder what those tweets were. Because some of our results... Those tweets if they were taken down for certain reasons, it's almost like our results could understate like, the parts of me too, that weren't positive. So it could have been folks who didn't have a good experience from disclosing online or it could have been, like, for example, you didn't find a lot of... surprisingly, we didn't find that many tweets that were, like harassing people or disparaging the hashtag, even though that's common now. But it could have been because those have been, were more likely to be taken down or something like that. So that data missingness is something that was kind of always in the background, and there was just nothing we could really do about it. So that's for the major thing when I think of data missingness. So you had mentioned... remind me of some of the other things you mentioned?

Sara Mannheimer 16:57
Um, bots... and

BSR05 18:56
Oh bots, yeah, yeah. When we were going through like, you notice there's some there, luckily, like it didn't seem like there's too much. But there's like some wonky stuff, like, you find spam kind of data. And it, you know, has the hashtag in it. But it's not related to Me Too. We did some things with the follower graphs of the people in our data. And that's always a little tough to work with, because the Twitter API rate limits on those are so strict that you can't always get as much of those networks as you want. So yeah. Oh, and then the missing period. Yeah, my friend had happened to miss something like a 12 hour period when [they were] collecting data. So the tweets, the tweet IDs weren't even in the data set from [my friend]. Yeah, we didn't think it would affect the final analysis and it really didn't, but it's kind of... I only noticed it because I went down to doing like a time series that was at a particular resolution, like if you kept it at a certain resolution you couldn't even tell. And so I could have easily missed it.

Sara Mannheimer 20:19
And so let's see, what strategies did you use them to, like communicate or clarify these issues to, like in your paper or to future users?

BSR05 20:37
In the paper, just trying to like be straightforward about it was seemed like the best bet, like 25% of tweet IDs—like that is a fair bit. On the other hand, there's so many papers out there that are made with like, 10% of all tweets. So like, it's a lot better than those at least. So, so yeah, we, like being straightforward about communicating that there. And then with the, like, kind of the description of the data online, kind of the same thing, just trying to be as clear as possible, how the data was collected. And then like, I think most people who work with Twitter data regularly know this, but kind of emphasizing that, you know, this data is collected at this time. That was like a year after Me Too. And so what this includes is Me Too tweets, but not those have been deleted by the time of data collection. So

Sara Mannheimer 21:36
Yeah. Did you work with a curator at [the data repository] to like, help work through how to just like, include these descriptions with the data? Or did you deposit just on your own?

BSR05 21:48

I think I wrote up a lot of it. And then I think it was [curator name]? at [data repository], who helps kind of walk us through the rest of the process?

Sara Mannheimer 22:01
Yeah. Great. Alright, let's go to data comparability. So during your example, did you compare or combine multiple big social data sets?

BSR05 22:14
For that project, we didn't really. There's like the tweet data we had, and then we bought some, but it's like the same data. Yeah, this one was pretty contained. I'd say. If I were to, like, give a different example, just for this. There's this... one of the groups I work with, they've maintained this panel data. And what that is, is they have public voter registration files, so, voter registration is public in the United States. So they have those. So they're real people, what they did was they match those to Twitter accounts. On Twitter. So what that does is it brings in the demographic information with the Twitter account, so you can start to ask questions about like, what are real people doing on Twitter versus this weird mix of real people and bots and organizations and stuff like that. So that's probably the biggest instance of the data that I work with that's like, combined with a different data set.

Sara Mannheimer 22:48
What demographic info is in the voter registration files, like gender? And...

BSR05 23:37
Yeah, the ones that come directly with it are gender, age and state of residence, some states report race, but there's other fields that are inferred by... it's a it's a commercial vendor who kind of assembles the voter file because they come from all these different sources. So they assemble it and then they add additional metadata that they infer based on proprietary classifiers and stuff like that. So we we've done some work here to like validate the proprietary stuff, and it's not too bad. But they that's how we have inferences. And sometimes it's different. Sometimes they add information for race and sometimes for political party if it's not reported.

Sara Mannheimer 24:32
Did you encounter any challenges when you were trying to combine the Twitter data and the voter registration files?

BSR05 24:40
I didn't do the matching myself, but I'm pretty aware of the process and yeah so just, you know, matching names is a difficult thing because of informalities and stuff like that, multiple people having the same name and same location, and I think they ended up just doing... the most basic thing they did was if there were multiple people in the same location, they just said they couldn't match the file. And so it's really a panel of people with unique names and unique location.

Sara Mannheimer 25:18
And who used their actual name as their Twitter handle, right. Or as their descriptor. Okay.

BSR05 25:23
Yeah. So they actually have to use their full name, either in their handle or their username, which there's a lot of people who don't do that. And that project was... that's a separate project, so that one was IRB approved, because I know, it can be a little fuzzy doing that kind of linkage.

Sara Mannheimer 25:43
Oh, okay. So the previous the Me Too data, did you go through IRB there or consult with them?

BSR05 25:54
Yeah, my my advisor generally has a kind of IRB approval for working with Twitter data and online activism, because that's [my advisor's] much broader gig. I don't think we had to submit a separate one to work with me to data. Yeah.

Sara Mannheimer 26:13
Do you know anything about that? Because I know, like different IRBs have different levels of the way that they sort of see Twitter data, you know, like, at MSU, they just, it's, like, exempt? Because it's existing data.

BSR05 26:26
Yeah, I have seen the IRB... I don't think it was one that went through as straight exempt. And I think it's partly because in the IRB, they also talked about how they would, how they would or would not, quote, individual tweets and things like that. So there's a lot of language. And I saw, I saw one of like, the rough drafts, and about how they were going to, like, make decisions about that based on... and if they would kind of reach out to people about getting their consent to include their quotes or something like that.

Sara Mannheimer 27:06
Okay. Well, that's a great segue. Our next issue is informed consent. So can you tell me about a time if there were any during the Me Too example, when you considered informed consent?

BSR05 27:28
So let me think just for a moment. It didn't come up directly at all in terms of like working with the tweets, so we didn't discuss it for that. I know the big thing that we talked about it with was in terms of kind of using tweets as examples or quotes or something like that. So the general guideline in my lab, that my advisor has us following, that I agree with, is we generally try not to quote people who aren't like public, big public figures or who, like, wouldn't expect that their tweet could be quoted, so not like the tweet that got two retweets versus the one that got 100,000 or something like that. But most of the disclosures were not like popular tweets, they were pretty small tweets. So what we ended up doing for those was altering, we didn't report actually direct quotes, we altered the text. And we do like altered texts, and we mismatch—mash together, like similar tweets, so that, hopefully, they shouldn't be identifiable. Like, you shouldn't be able to reverse look them up or something like that. And I know there's kind of ethics there to the thing, reporting the data.

Sara Mannheimer 28:57
Yeah. So in the paper, you said, an altered example of a, you know, tweet.. is this.

BSR05 29:09
Yeah, exactly. Oh, and I don't know if I actually say anything about informed consent. There are instances or sometimes like a tweet is, like, a really good example. And it might be that one we use... this didn't come up in that paper. But that is a case where we would, like, reach out via Twitter, like via Twitter DM to ask someone. And I know that's something... my advisor has a book on hashtag activism, and for like the big examples that they have, they reached out to people, since they knew they were going to be kind of memorialized in the book.

Sara Mannheimer 29:44
Yeah. Um, have you or have you ever consulted with anyone about like, practices around informed consent? How have you come to these strategies?

BSR05 29:57
Yeah, so most, a lot of strategies come through my advisor, who has been working with Twitter data and this kind of stuff for a long time, and I noticed [they], a lot of [their[practices come from, you know, working with [their] collaborators and kind of trying on, like black feminist practices and things like that in the literature. And then, you know, I've taken classes where research ethics are part of the conversation. And I've done my own kind of reading in general, on research ethics and thinking about these kinds of things with Twitter data.

Sara Mannheimer 30:40
Yeah, have you found that it's more common when you're looking at research ethics literature to find a discussion of this? Because I feel like, you know, with IRBs not having caught up, there's not like a lot of official guidance. But yeah.

BSR05 30:56
Yeah. This is like one of my little hills that I've been trying to die on, is like to bring those statements a little more forward in papers. So you tend to either tend to not see them? Or you tend to see them as like an, oh, by the way, like, we got it approved by the IRB, or like, oh, by the way, like, this is maybe an ethical issue, but we're at the end of the paper, so like, there's that so. So I like, I tend to not see it with my papers, I've been trying to almost put it towards the front of the Methods section, because ethics should come before the methods really ideally, but. This is my hill to die on, because my co-authors always are trying to like make it shorter.

Sara Mannheimer 31:46
But do you feel that—I guess it depends from project to project. But do you have like a sense that participants expect to give consent for a project like this, like a hashtag project?

BSR05 32:04
My understanding... there's been like some research on this. Is that most Twitter users don't always know that their data is being used like this. My understanding is most are okay with it being—they would be okay with it being used if it was like being used for good ends... like whatever those are. But so like, for Me, Too, though, like I could very easily understand someone being perturbed about being in that data set. And so that's... So it's we kind of try to do it as a balance. Do we think this research is important enough? And given how, if we think it is important enough, you know, what safeguards can we put in place to make sure that this person isn't going to face harm from being in the data set? So the hosting in [a social-science specific data repository] versus just like [a general data repository] or something like that was one of those safeguards that we put up?

Sara Mannheimer 33:08
Interesting. Awesome. All right. Let's move to—Let's move to privacy and confidentiality. So can you tell me a time during the research when you considered issues of privacy? Like restricting access to your tweet IDs is one example. Or protecting the data during the research, or thinking about the expectations of the users? The Twitter users. Yeah.

BSR05 33:40

So the big one is like, we don't want people to be able to easily identify survivors of sexual violence. And that was, we, I think we had several conversations about that among ourselves, trying to figure out the best way to—like if we could share the data responsibly. Because I think if we hadn't identified [the data repository we used] as an option, we wouldn't have shared the tweet IDs. We thought that'd be a little too far. Especially given that like our paper, it's not easy. like it'd be almost easier, just to like manually look up the hashtag Me Too, and harass people through there, than, like, use our tweet IDs, build a classifier after having read your own, like 1000s of tweets. It's not really practical that anyone would actually misuse our method. But it's still like it describes the way of identifying a certain group of sexual violence survivors. And so that was kind of a conversation for us in terms of privacy. There's like smaller thoughts I had about... like no one like, really was watching me with the data and so I was always trying to wonder like, how many copies of this data should I have? I had like one.. I basically had one of my laptop and I had one on the lab computer. And like, where should these be. And it's like, now I don't have the copy on my laptop. But because the more copies there are like, the more chances there are that someone is gonna be able to touch it that shouldn't be able to touch it. And at the time, it didn't seem like a big deal, because I was the only one in the lab. And I was like, the only one working with the data. But over the past year or so, like I've started to bring more people onto the lab and onto the computer. And I have to make sure you know. I love all the people in my lab, and I trust them. But I'm not always going to be the one managing the computer in the lab. So.

Sara Mannheimer 34:28
Yeah. Tell me more about the conversations you had with your colleagues, and like how you worked out these ideas about privacy and how you would handle it.

BSR05 35:58
Yeah, there was, I would say... There's a nice kind of range of like how people felt about sharing data. So there was one professor who felt that sharing the Tweet IDs is like the right thing to do, it's good to be kind of open with this. Especially since you can share Tweet IDs. And kind of on the other end, there was one of the other professors who was much more hesitant. And so [they] pump the brakes on a lot of those conversations, or [they try] to raise concerns about it. Because—I'm trying to remember, there's like this specific concern. Oh, it's because, I forgot we did some web scraping for this too. I'm sorry, maybe... I don't know if it's relevant. But I feel like I should mention this too, because it's just a another example. So at the time that we collected the data, you, you couldn't get replies to tweets through the Twitter API. So it's like only the keyword tweets. But we've felt that it was really important to have replies to Me Too tweets, because that's the conversational part. And so we're part of like, what we're looking at is like, how survivors might talk to each other or support each other. We have plans for future projects around that. But you can't get those through the Twitter API. And so it kind of put us into position where the only way to get them was to do the web scraping, which is against Twitter Terms of Service, technically. So I collect... I don't have major qualms with like, bending Twitter's Terms of Service with web scraping. Like I wasn't like mean to their servers, like I put in pauses, so I wasn't like hammering them or something like that. And I got the replies, to Me Too tweets, I get the reply threads. But we had conversations, that came up about whether we should use them and whether... not like whether we should tell people we use them, but like, yeah, whether we should use them in the analysis, which would affect whether like we report it in the paper or not. Because to report it in the paper is basically to say like officially, like we broke the terms of service, and you have to justify breaking the terms of service, because... and that's, you know, because the terms of service aren't like ethical rules. They're just a set of guidelines set by a corporate company to like, protect themselves. So the users... so we had conversations about that, too, which kind of... There's one advisor who was like, I do not care what Twitter thinks like, let's just use the tweets. And another who's like, like really does very understandably did did not want to, like, get in trouble over this or something like that. So we kind of

talked back and forth over that, trying to decide. So like, these kind of back and forth conversations where a lot of... I guess like just walking around the ethical markers, like trying to weigh, like, this is the benefit of doing this, versus this is the cost. So with the scraping, it's like the benefit is—you get this much fuller picture of me too. Versus there's potentially legal consequences here. And then with the releasing of tweet IDs, you know, the benefits of open science, being able to facilitate further research on this. Versus the cost, which is like potential harm to the participants. And so I think what pushed that conversation into releasing them was that I was able to propose [the data repository we used] as kind of this—I think someone referred to it like as a walled garden approach, like the data is there, you can see it, but like, there's a wall around it, that only certain people can get through. So I don't know if that answered that last question that you were looking for, but...

Sara Mannheimer 40:57
Definitely, yeah, this is really interesting, because you talked at the beginning about sort of weighing whether the research was important enough to have some, you know, some danger to the participants—some risk. And then you talk about, like, weighing openness and privacy. So I'm seeing like a lot of ideas at odds, where you're weighing is the research important? Is openness, important enough? Is following those terms of service important? So it's interesting to hear how these conversations went with your group. And so with the... this idea of the web scraping, I feel like this is going into my last question about intellectual property. Like, how did you resolve that conflict with the web scraping benefit and costs?

BSR05 41:55
Um actually, I know, like, something pushed us toward a final decision, but I can't remember exactly what it was. I can't remember. I'm not sure what it was. But I so I think part of that conversation, though, is like, yeah, if you could get the tweets via the API, you could, you could just do that. And so there's part of the question is like, why can't you get them through the API. And at the time, and I still believe this was probably because... it was probably just like, technically annoying for them to like, set up the API so you could get replies. And so the reason you couldn't get replies wasn't because they were like trying to protect users in some kind of way. And that is kind of one of my major beliefs is, Twitter isn't really doing this to protect users. If they're doing it, they're doing it to protect themselves. Computationally or something like that. And so looking at it as the users' tweets, they're just the same as all the other tweets that we can get. It's just a matter of how we got them. And you know, I understand like Twitter... it's an international website, like they cannot have 1000s of web crawlers hitting their website all the time. But so you know, the little thing I told you is just, you know, be nice about going through it, like not making any more requests than like a very fast user could make or something like that.

Sara Mannheimer 42:47
And so then did you end up talking about it in your paper, too, that you had to use the web scraping and said that you felt that it was small enough that it wasn't against the Terms of Service?

BSR05 43:48
Yeah, we ended up... yeah, I think, um, I'll send you the paper after this so that you have the reference for it. I know, we mention it. So we say... and it's kind of, we don't explicitly say web scraping, I think, because what we say is more like we used the package to do it. So we just say, we used this package to get replies. But like, here, anyone who works with Twitter data, you're gonna see that package and know that that's not an API package. I'm not... I'm sure... I think I wanted to include a discussion of it. But it got struck out because of the concerns. The potential legal concerns. I remember I wrote a discussion at one point, and I think it got struck from the draft.

Sara Mannheimer 44:45
Interesting. And so like, Well, I think that's enough, actually. Yeah. This is a common... You know, I've heard data curators talk about breaking the terms of service at times, and I feel like as academics, we're like, we just have different values and priorities than a corporation. And so it's interesting, like, if I'm following my values, which is to get the full data and create a research project that's meaningful, like, Is it okay to bend a little bit?

BSR05 45:24
Yeah, I think I've heard someone say, if you're like a researcher, especially like PhD student, one of the questions you have to ask, and I think is a valid question, is like, what if you did scrape it against the terms of service? And then like, you couldn't publish your PhD because they came after you? And that's like, a scary thought. I don't think it's ever happened to anyone. But I think that's the kind of risk it does pose? Yeah.

Sara Mannheimer 45:57
All right, great. So that's all of my like structured questions. Are there any issues or challenges that arose during the research that I had didn't ask you about? Or that you expected me to ask? And I didn't ask?

BSR05 46:15
No, not too much. I think I always think of the panel data that I work with is the most... I mean, the survivor data is obviously ethically tricky. But the panel that I still actively work with the most ethically tricky. Just because it's matching datasets. So I think we're always trying to balance that one. So like, only using it for questions that meet that level of importance of using the panel? Because I know we get questions about it. But I think one thing... maybe this is just a general comment. When it comes to like matching data sets, or working with, like, these highly sensitive data sets, like sexual violence survivors, is, um, I wish sometimes these conversations were a little more like the weighing stuff. Whereas I feel like it's really easy sometimes to just say, like, that's risky, and so they shouldn't do it. And then there's the other end, it's like, they don't think about it, they just do it. I don't, I think I think, in working with big data, there needs to be a little more room for thinking about, like, this is risky data, but we still can do something with it in the same way that, you know, biologists work with risky chemicals, or not even biology, but like, you know, medical records are very sensitive. But like we have them, we use them. And there's reasons for using them. But there's then like certain precautions go around those. So I think yeah, that's just something to think about sometimes is, how can we work with this risky data, or data that might be risky? And make sure that we're treating it the proper way? So that it's like acceptable to be taking that level of risk?

Sara Mannheimer 48:10
Yeah, that's part of my research. I feel in a way. It's like, how do we... Yeah, how do we make this happen? And I'm hoping like, through the interviews, I'll be able to find people on both sides of the spectrum, like very cautious people, and people who are like, eh, whatever. And then we'll see. We'll see from that, like, if there are ways that data curators can sort of help people weigh those options and help see their side. So yeah, cool. Well, this has been amazing. It's so fun to hear about your project and all of the challenges that you encountered. Thanks for taking the time.

BSR05 50:07
Yeah, I hope it's helpful.