DC02\_transcript\_deidentified

**SUMMARY KEYWORDS**

data, people, interviews, sharing, researchers, qualitative data, identified, participants, archive, tweets, project, context, qualitative research, consent form, research, big, published, metadata, depositor, case

**SPEAKERS**

DC02, Sara Mannheimer

**Sara Mannheimer** 00:00

I'll record now so I don't forget, thank you for reminding me. I have a point in the guide where it says, and now we will start recording. But um, yeah, so my thought is that big social data is this complicated new type of data. And whereas qualitative data is well established, and curation practices are more well established for that. But both have similar challenges. I did a big lit review where I identified challenges in each of the different types of data and publishing and sharing both of the different types and identified the six issues that they have in common. So I know you said you didn't agree with my premise. Tell me more, or do you want to go through the interview first?

**DC02** 00:51

Let's let's go through the interview first, and then let's chat about...

**Sara Mannheimer** 00:57

Okay.

**DC02** 00:57

...where I think this may be true and how I think it may also not be true.

**Sara Mannheimer** 01:03

Okay. Let's see. So, the interview is structured around these six key issues that I identified. And so we'll go through them one at a time. And I have seven questions. Some of them have um, have like a sub question. And I think the interview should take about 60 minutes, it might take up to 75. And I also asked you to identify a specific example of a recent time when you either supported curation for big social data or qualitative data, or observed someone firsthand, curating one of those types of data. So did you have an example in mind?

**DC02** 01:50

Yeah, I mean, I have obviously lots. Yeah. So anything in particular, any type of research, you're you're interested in, like more interviews more other stuff?

**Sara Mannheimer** 02:05

I guess, you know, and it doesn't even have to be one example. I'm just trying to get at a recent, recent specific experiences. So I wanted to like, focus on...

**DC02** 02:18

Well, I mean, let's, let's make this a kind of baseline example [adds a link to the videoconference chat to a dataset that is published online]. And then I can also draw on other examples.

**Sara Mannheimer** 02:28

Perfect. And then if there, if other things come to mind, as we're going through the questions, you can also sort of pull those in. It's not that... Okay, great. I've actually heard of this study [that is, the example provided in the chat], so I'm a little bit familiar with it. Cool. All right, let's get going. So the first question I want to ask is about context. So I have two quotes here, sort of setting up what I mean by context. So for qualitative data, "Qualitative research is a process that may include deep and prolonged contact and connection with your research subjects attempting to understand the subjects within their own context. And qualitative data are therefore highly context dependent. So context is a source of data meaning and understanding, ignoring it under using it or not recognizing one's own context driven perspective will result in complete missed meaning, and a misunderstanding of human phenomena." And then for big social data context, has some more nuance. So Halavais says in 2015, "When we collect data from social media platforms, just as we, when we collect data in traditional state spaces, context matters. But the context of a social media post may be absent or difficult to understand social media posts are by nature, short pieces of text taken from the larger context of personal and public life. And the out of context effect is only compounded when data are amassed at a large scale." So that's sort of where I'm coming from. And so during your example, what challenges did you encounter, if any, when you were trying to capture the context in which the data was collected?

**DC02** 04:18

Yeah, so the degree to which context matters varies a lot in qualitative research. So if you have a multi method piece, like the one we're talking about, right, these were one off interviews. So the transcript, kind of like our interview now captures, I would say probably 95% of the interaction, which is distinct from if you think of something like ethnography as the as the opposite end of the spectrum, where any given interaction only captures a fraction of the experience and even if you have textual record of all the interactions, you still wouldn't have the emotional component, kind of having lived that component that's part of ethnography. So the degree to which context matters varies a lot between different types of qualitative data. And for this particular study [the example DC02 is using to structure this interview], this was a very well set up and documented study. So we had extensive documentation right off the bat from the researchers who are information scientists who think about documentation and metadata. From the start, and we didn't have to do a ton to ask. So we typically have some general things that we think everyone, like, regardless of approach, one would like to know about qualitative data, right? How long were the interviews? How did you pick who you talked to? What was the setting of the interviews? Who conducted the interviews? What happened between interview and whatever qualitative data we're getting? So was it transcribed? How was it transcribed? How was it de-identified if it was de-identified? When was it, when was it collected? Can you share the interview guide? Those sorts of things, and...

**Sara Mannheimer** 06:27

Oh, sorry, to interrupt. Do you ask for like demographic information of the people if they're de-identified. For more context?

**DC02** 06:35

We saw that's often included. We don't, we don't routinely ask for it. But in some cases, I don't think people collect... I don't think for example, I'm pretty sure. I'm not sure if they even collected it. Right? Because that's also, depending on what you do, demographics matter more or less. They're also... if you do articles... uh, if we do data accompanying articles, because demographics a part of all the reporting guidelines for qualitative research -- are part of COREQ, and SRQR and whatever they are, they're almost always included in the, in the article already. And, and yeah, so we have kind of a checklist of things that we always like to see. And then we look at the project, as we understand it, and ask some additional questions about stuff that might be interesting. And in some cases, people come to us with very little. So we ask a lot of questions. And then in some cases, people come to us, where we essentially have almost nothing to do and then just make sure that it's represented in a somewhat structured form.

**Sara Mannheimer** 07:57

Yeah. Okay. And I have one follow up question about the strategies that you use. So most of your strategies are linking the data set to the article and adding metadata to describe all of these.

**DC02** 08:13

Yeah, and our approach to metadata is heavy on narrative, unstructured metadata, we have structured metadata, of course, about I think, 20 or so standard fields that we try to have completed. But I would say the most important and part of the metadata is narrative and unstructured. So that our accompanying documents like the informed consent, recruitment, interview guide, especially, but it's also often a lengthy data overview, a data narrative that describes -- in terms adequate for the research, right, which varies quite a bit -- the relevant parts of data collection, and in some cases, even analysis, if we have analysis outputs as part of the data, so say, codebooks, etc. That could be part of what what would be what would be included in contextual, like, how did we come up with this codebook which, of course is very important in qualitative research, because the codebook often emerges from the data rather than preceding the data as in quant.

**Sara Mannheimer** 09:27

Yeah. Okay. Have you curated big social data, like, do you have...

**DC02** 09:39

Nothing big. No, we have some social media data, but I think we don't have anything that is even above 100, post tweets, etc. So certainly nothing that would be even remotely qualified.

**Sara Mannheimer** 09:52

Okay, so we'll skip the 1a's, 2a's etc [numbered subquestions in the interview guide]. So let's move on to data quality. So during this example, what challenges did you encounter, if any, when you were trying to document data quality? So for qualitative data that might be bias, or quality of methods? And then what strategies did you use?

**DC02** 10:17

We would be hesitant to document data quality in the sense of quality of method or bias. I don't think that's really something. I think it's rarely the curator's job. But it's also just given the variety of perspectives and qualitative methods. People wouldn't agree, and I think our role is to be a somewhat neutral party and trying to judge this would impose one particular perspective on data. And I don't think that's, that it would be feasible for us or added, or, or okay for us to do. So I actually think of documentation as the most important part where we shape data quality, because documentation, as your preceding quotes suggest, effect how data can be understood, and, and interpreted. Whereas for the data themselves, we do very little to enhance data quality, because standard things like you know, missing data doesn't really make sense. In the context of qualitative data, bots don't exist, bias and quality of methods I already talked about. Right. So yeah, so so our main impact on quality is actually the quality of the documentation and description, rather than the quality of the data.

**Sara Mannheimer** 11:48

I guess I'm trying to think of like examples where data quality would prevent the data from being reusable. Um, yeah, I guess, when you look back at like your examples of publishing the codebook and publishing all of these different types of metadata, you can see how that if there were any issues, you would just document them there as well?

**DC02** 12:17

Well, we would try to fix that, right. That's what I mean, right? That that's the places where we would try to, try to get more information and, and get things into... Well, most often, it's actually getting things at all, right, by getting an interview guide. And getting a codebook is often requires asking and oft-, at times even just telling people what you're looking for. One thing is when we curate for transparency, rather than reuse as for some qualitative data, articles where full data can't be shared, right, so we'd get a matrix of excerpts, something along those lines. And that's often something that researchers have thought about, right? They get told by a journal, you know, "You need to share your data, we look at the informed consent," and they're like promising only to share aggregate data or something like that. We're like, "you can't just give us the transcripts. It won't fly with your with your consent. But you could write like, give us for all the different codes, themes, notes, whatever you'd call them in your research, a description or coding strategy, and then one or two extended excerpts. And that's kind of, I think, the place where we typically interact most directly with depositors in how the actual deposited data looks match for transcripts. Their transcript, we look at the identification, which is something we'll talk about later, but we don't judge impact or would even know how to judge the quality of that.

**Sara Mannheimer** 14:03

Yeah. Okay, thanks. This is helpful. All right, on to number three data comparability, so during your example, and what challenges did you encounter if any relating to comparability or interoperability of the data set? So my examples are missing data, different research questions, different methods, metadata interoperability, and then

**DC02** 14:43

Right so in terms of metadata we tried to have... and that's typically not a problem for project level metadata. To have standardized metadata we use. A simplified DDI codebook But also have clean mappings to DataCite. And especially with the most recent DataCite kernel updates, I think we can map almost any metadata field to DataCite's. So in terms of findability and like the interoperability as it's defined in FAIR. That's actually hardly ever a problem. In terms of interoperability, how well could other researchers use this? That really depends on on on the project, right? I think the nature of qualitative research is that it's somewhat personal, right. And the good project was kind of if you think along an epistemological spectrum wasn't particularly personal, right, like [the researchers'] personality and positionality, they're right there, like both active open science advocates, etc. So that certainly plays a role in who they talk to, and how those people responded to them. But it's not like, you know, Alice Goffman, being a white woman in a black urban community type of positionality. So so it's kind of more... the researcher doesn't loom particularly large as part of the data, which makes, I think, it much easier to be used and reused by other people. I think you can actually just treat this as... if you've read the documentation and the data, you could just run with the data and reanalyze them and do something useful with them. And qualitative data aren't standardized the way like survey questions are. And yeah, I don't think they should be necessarily, right? So synthesis is harder, but can be done right? You just have to be aware of [the fact that] questions were asked at a different time by different people in a different context. So you don't expect, you know, one on one, psychometric mapping of answers, but you can still understand similar trends and similar projects. So in that sense, coming back again, to context, right, if you have enough context, if you understood how the data were generated. You can use them in a comparative reuse, even synthesis context.

**Sara Mannheimer** 17:31

Have a little side note, have you seen that happen with some of any of the data in [the repository where you work]? Have you seen people like...

**DC02** 17:43

So it's not a ton, we do see a fair amount of pedagogical videos. So we do see classes coming in and looking at different data sets, doing encoding and analysis on them, comparing data to published articles, those sorts of things. We have - there is a special issue coming out in [a journal focusing on qualitative research] where they grabbed one of our data sets and had it analyzed by four or five different researcher groups. And then, more has been done on the larger collections at UK Data. So as you know probably, Libby Bishop has written about secondary use a fair amount. Florian Kern from Essex, a political scientist, has, and I'm sure he'd be happy to share that with you, has a manuscript where they actually did a secondary analysis of a development project, a qualitative development project. So it's... I forgot exactly what the topic was. It's Sub Saharan Africa, and particular types of development projects. And I think they looked at 25 different UK Data data sets. So, and we're quite pleased with what they were able to do. So it's still comparatively rare, but we're starting to see more of it. Right? [In a smaller repository], the probability that you find a ton on your specific project currently, it's just not that high.

**Sara Mannheimer** 19:13

Right. Right. Wonderful. Okay, question four, informed consent. So in this example, what challenges did you encounter, if any, relating to informed consent?

**DC02** 19:27

So this is not a particularly good example because [the researchers], of course, were on top of this. The informed consent clearly spelled out exactly what they would do with the data. And that's highly atypical. So so let's maybe move to some other projects, because this is something that looms large. And what we do so we review the consent for every data project that we get in that involves interaction with human participants. Which isn't all of them, right. We sometimes have historical data, etc, with other challenges. And, you know, on the one hand, you have a couple of examples. Those are either people who are really tuned into data sharing debates, or people who have talked to us at the data management planning stage, where the content is crystal clear and spells out, you know, we are going to share data with other researchers or even has a separate consent for, you know, are you okay with having your data shared with other researchers or in a repository or even in [a specific repository]? Those are kind of the languages that we look at as completely unproblematic. There's clearly data sharing built into the consent, that happens, we're seeing more of it, but it's clearly not the norm. So, and then we see kind of a broad area that, on the other end of the spectrum, have very clear statement that data will not be shared, that data will be destroyed, etc. And those people we basically turn away, or as I described before, we suggest kind of aggregate data publication issues that we don't think qualify as data sharing even in the minds of the of the participants. And then there is the gray area in the middle, where there's like a large silence about data sharing, where people don't say, "no one will see the data, I'll destroy it." But they also don't say they'll share it. And there, we go kind of through a complicated -- internal currently, but hopefully, eventually public -- decision tree. And how we make those decisions that takes into account both the researcher assessment of what participants were actually expecting and thinking, an assessment of the degree to which the data can be deidentified. So if there's not- no clear consent for data sharing, we have to be very confident that the data have truly been deidentified in order to consider even, sharing it. And even then we'd only do it if they're relatively harmless data, right? So if you have data that are super sensitive, and you don't have clear permissions, we tell people that that's probably not okay to do. What's interesting is that IRBs are only... have limited help here, because a lot of IRBs think once data are de-identified, they're no longer human participant data. And so they kind of wave their hand. Not not in a unified way, right. Like, in the U.S., it's kind of 50/50. Some IRBs say you can't publish and some IRBs say no, that's fine, and it's also not really our job. And then so we get into this weird situation where people come have like, the IRB says, "No, sure, you can share that." And that's like, we [that is, DC02's organization] don't really think you should, though. And I think there is a particular perhaps unexplored issue with qualitative data. There... that may be similarly applied to to social media data in that, as opposed to survey data, participants would always be able to recognize themselves in deidentified data, right? If I see a survey, and that's been deidentified, I can, I can not find my role. If I see 100 deidentified transcripts, it takes me 20 seconds to to recognize mine, which means participants know that their data is in there if they ever were to access it. It's also interesting in the European context, right, the GDPR makes this distinction between anonymization and... I think deidentification, I think that's the label they use. And if I understand GDPR correctly, qualitative data is just not... it's not possible to anon-, anonymize it.

**Sara Mannheimer**  24:09

Yeah.

**DC02**  24:10

I probably went a little further into privacy and confidentiality, than you wanted. Oh, one other strategy. So so and and I hinted at that. We try to do a lot of advising templates early about support for informed consent, to get into that as early as possible, right. Like, whenever someone comes to us to ask about, for example, for an NSF project, [they ask,] can you give me a budget? Even if they don't ask, we always ask, you know, have you thought about consent for data sharing, because that is a problem. We give workshops. We bring this up all the time, we have templates on our on our website. And that's not particularly unusual. Right? So does [specific repository], so does [specific repository] and so does probably every actively supporting data librarian in the country. But it is it is an important part of what we do to improve the informed consent situation. Interestingly, and that's the last bit, we also do active research on this quite a bit. And I have a recent paper out where I've worked with two researchers from [institution] - the paper's out on [preprint site], you can find it there, where they had opt-in consent for data sharing. And we were interested in how that type of tiered consent, tiered consent for data sharing, right. So they... there was an additional option, you know, you can, "would you also be okay with us sharing the data with other researchers? You can still be in the study if you don't." So it was a tiered consent, where data sharing wasn't a condition to being in the study. So we were interested in how many people would say yes, and that, but then they also asked follow up questions at the end of the interview, like, did you think about this? How did you think about this? What motivated you? And one thing that we hadn't really thought about, but that came through, is that people don't really understand the question, even though it was fairly simple form of the question. And so that's kind of troubling.

**Sara Mannheimer** 26:23

Like they didn't understand what it meant to be sharing the data? They're sharing the transcripts or, huh.

**DC02** 26:32

Yeah, and there are a couple, if you look at the paper, they have a couple of quotes where they are like here are the people where we didn't get the sense that that they actually caught what we meant by by data sharing, which is obviously problem for the informed part of the informed consent.

**Sara Mannheimer**  26:48

For sure. Yeah, I mean, in a couple of these interviews, I sent the consent form, just like right at the beginning, before we had logged on to Zoom. And as soon as we'd logged on, I said, "Oh, let's review the consent form." And they go, "I already signed it." And so I feel like we're all so used to just like clicking the button, whatever, I'm sure it's fine. That like, yeah, I can see it being a problem.

**DC02** 27:15

And then what's the other thing... oh right. We also have, and that's not mainly me, that's mainly [my colleague], who I think you know, who is working with IRBs. And so has research projects on how IRBs' thinking on data sharing is, is evolving. And that also, of course, for example, you now see that an increasing number of IRBs, Cornell was one of the early ones start putting language about data sharing into their templates, which is, of course, will have a much bigger impact than us recommending.

**Sara Mannheimer**  27:53

Right. I mean, the problem is that data repositories... people still tend to think of them as an endpoint, you know, whereas the IRB, everyone knows happens at the beginning. Yeah. All right.

**DC02** 28:07

Ok I think that's actually now all that I have for informed consent. Yeah.

**Sara Mannheimer**  28:11

That was great.

**DC02** 28:12

Thank you.

**Sara Mannheimer** 28:14

I look forward to reading that paper. And I'm thinking, well, maybe you can tell me if this is wrong. But I was thinking maybe that I would add to my consent form the option not to have your full transcript shared, just to give people that option, but maybe it doesn't. Maybe there's no purpose...

**DC02** 28:32

I, no, people really like that. I would, generally people rarely opt out. But one thing that we found is that the people who did opt out, they understood exactly what data sharing meant, and they didn't want it. So I think, which is fair, right? This was abortion research. So it's also like, a couple of levels more sensitive than what you're asking me. And so no, I'm a huge advocate of tiered consent forms, the only place where I've done it and where I wouldn't do it again, is in survey research, because you can't do reproducibility if you have people opting out of data sharing, because then the public data set and the study data set are different. So that I made that mistake once and I'll have to live with like the five people who opted out of data sharing for the survey, making slightly different reproducible public and private data, but no, otherwise I'd always do it.

**Sara Mannheimer**  29:30

Yeah. Okay. Okay, on to privacy and confidentiality. We have discussed some related things here. But what challenges did you encounter in this example, or others, relating to privacy of the people represented in the data? And then what strategies did you use to address the challenges?

**DC02**  29:48

Yeah, so the data were - participants were assured confidentiality, although I think [the researcher] told us, most of them wouldn't have cared, and the content of the interviews wasn't at all sensitive, like literally not at all, which is rare. Like typically there is at least some professional risk involved or some reputational risk. So the degree of worry was limited here, but still we always read through all transcripts that we get and flag for the positives... issues with deidentification, and those are often contextual, right. And that's typical for qualitative data, right? It's not that one piece is a problem, but that a couple of pieces taken together might be problematic. And right, the stuff that you can automate, like finding proper n-- proper nouns, etc, is easy. But that's also typically not sufficient. Then there is also the problem between how much did you take out, versus at what point are you starting to, to hurt the data. And for more, in depth interviews, you always reduce the richness of the data when you're deidentifying. In this case, not so much, right? Because, again, people were talking about [a low risk topic], which doesn't really require a ton of personal background or identifying information. But even with patient data, right, which is we're getting a fair amount of public health or like patient experience data, those sorts of things, you're losing some richness. And then with... so for example, if you want to de identify this interview, good luck, right? I'm doing research researcher interviews right now. And it's like, I thought we were going to share deidentified copies. But I'm essentially arrived at the point where I don't think we're going to be able to do it, because every second sentence is like about this very specific thing that everyone in the researcher's environment will know exactly who they are.

**Sara Mannheimer**  32:20

Right, so what will you do in that case?

**DC02**  32:26

We probably will share because we won't fully code the interviews. Right? So in the sense of not every sentence will be coded in the interviews. So we'll have sparser coding. We'll just share a complete a complete code export. So we have every... and then that will still need some deidentification but much less. Right. And you don't have all the different... Yeah, so so that's probably what we'll end up doing. And that's kind of the types of things that we talked to depositors about in in these scenarios. What can you do that is still interesting, often more in the sense of transparency than reuse, that preserves the data. And in other cases, I would look at access restrictions. But there, you also have to think about, you know, is it is it worth it? What do access restrictions protect, especially if you're looking at researcher interviews, right. Other researchers have the prime problematic group to get access. So so the access restrictions wouldn't in this particular case, wouldn't help us much. But that's certainly something that we think about in other scenarios. And the most common scenario for, like simple access restrictions is actually again, contextual, right? If people are concerned about if they do interviews in a well understood group, right, like in small town, etc. Right? It's going to be impossible to deidentify the interviews so that their neighbor doesn't recognize them. But it's very easy to prevent their neighbor from not getting access to the data. So that's, I think, is where we get a ton of bang out of access controls because they're very easy to implement, right? I just need to verify researcher identity, brief research plan, I don't even need secondary IRB or like fancy signatures, right. I can have something that we can handle in 24 hours to get someone access to the data. And that's kind of the access controls. I'm fondest of because they solve a fairly complex problem in a very low bureaucracy way.

**Sara Mannheimer**  34:54

Yeah. When [the organization where DC02 works] does access restriction Then like, do you have standard restriction like restriction levels that you use?

**DC02** 35:08

We have kind of, we have a basic kind of level of how we treat the data in terms of data security. And we have fairly clearly defined levels, like what data can still stay on the servers where they aren't encrypted. So for example, data like this, that aren't particularly sensitive and have low risk, but may may still, you know, have some leftover identifiability by acquaintances, we leave them unencrypted on the server, right? People would also email them. So it's kind of the same idea. And as we get more sensitive, the data get removed from the server, because it's not encrypted, and they get stored internally, on a specially encrypted server, those sorts of things. So so those protocols vary. For the access controls, we have kind of standard language for a couple of common scenarios. So right, like we'll just check a research plan and verify identification is by far the most common thing we do. We have template language. We do have template language for -- and that's actually an interesting one -- for depositor-approved access. And we don't love those, because they pose some challenges to long term accessibility, right? What do we do when the depositor goes away? Or we can't get ahold of them. So we built a fallback clause into those right, if we can't don't hear back in X weeks from you, we'll use these criteria to provide access. [Gives an example of a project that used this type of access control].

**Sara Mannheimer**  37:31

But you did, you did publish them, and you have them just under restricted access, and then [the researcher] gets to decide who sees them?

**DC02**  37:39

They're deidentified obviously. And [the researcher and their team] get sent the access request, including the identity of course of the requester.

**Sara Mannheimer** 37:52

Right. Very interesting.

**DC02** 37:56

Anything else in there....? Yeah. So I think that that, that really, I think we've covered all three of the scenarios you describe there, in some depth.

**Sara Mannheimer**  38:08

So, okay, and these are just examples I thought of, so if there are other things that you have experience with, you can go beyond the the parenthetical.

**DC02**  38:19

I mean, we do have on-the-record interviews, right. I think people need to remember that not everything needs to be deidentified. It's the norm, but some interviews are just perfectly fine to give, to give an on-the-record interview. Journalists do it all the time.

**Sara Mannheimer**  38:35

Right. Right. And in fact, then journalists often don't-- wouldn't publish the transcript of the whole interview. And whenever I've been interviewed by a journalist, the quote is always like, huh, did I say that? Even just like MSU news, people, you're like, did I say "flywheel?" I would never say that. Alright, so this is our last question, intellectual property. So during your example, did you encounter any issues regarding intellectual property concerns? Or other examples.

**DC02**  39:15

Yeah, so this one doesn't, right. And interviews typically don't. But we have a fair number of other examples. And let me just walk you through a couple of those. So very recently, we published a collection of hard rightwing, kind of terrorist manuals and related publications. And these are often kind of published as pamphlets, or like by a super independent, unofficial presses, etc. And so we did two things. A) We restricted access to them at the request of the depositor, but we think he's right about that, because we don't want to be a repository for rightwing terrorists searching for pamphlets. So right, again, we want to see some sort of research affiliation. Most of these are still under copyright, and we made a fair use determination right. They are for educational use, we have limited distribution, because they're under restricted access, they were not published with, with a profit motive, so a lot of the Fair Use factors. And we kind of make a perhaps slightly weaker argument, but that grouping them contextually into a data project is some sort of transformation. I don't know if strongly that would hold up in court. But overall, we're fairly confident that this would hold up. And I think other repositories do similar things with this sort of material that's not for profit, published and reproduced for scholarly uses. So that's not I think, an unusual or particularly envelope-pushing Fair Use claim. So that's number one. Number two, is we ask. That works really well, for newspapers outside of the U.S. in the U.S., most newspapers monetize their archives. In Latin America, most newspapers don't. And so we've had several people who went to newspaper archives in Latin America, and then just sent a letter asking, you know, here's what I collected, I'd like to make it available. And the newspaper just says, sure. No license fee, no, nothing. So that works. And, again, of course, in the U.S. this would... I don't think even for regional newspapers, it would fly. But certainly for the big national newspapers, which this usually is in respective countries, the New York Times would send you to the page about, you know, relicensing, there are the costs, and you end up with a $20,000 bill for a data project. So I think, archives are really interesting. So to various parts of, right... A), the stuff often still is under copyright. It's very interesting, we published [material that was part of an archive in the U.S.] and had a very nice interaction with the archivists there. And they told us, they're happy to see. So there's two, two parts, actually, to archives. One is the archives' own rules. And U.S. archives tend to be fairly permissive about digital reproduction of their holdings, although I'm sure there are exceptions. European archives much less so. Like we've we've had researchers get very stern letters from from the German national archives, for even like using a research assistant, because the archive was just meant for them. Along those lines. Let alone... and then kind of as you would expect for the German archive, they have a two page single spaced list of requirements for sharing scans from the archive, which includes no higher than 75 dpi. So they really don't want the digital materials out there. Which just and you know, we don't piss off archives. So if the archive says no, we don't publish. So yeah, so we asked the archive, which is also how we were in touch with the [archive in the U.S.]. And that documents and it's like [some parts of the archival collection] were donated by the [creators of the papers] to the public domain, so you can do whatever you want with those. But the letters in the file, that are part of the collection, and that were written by [other people besides the originators of the manuscript collection] are under copyright and belong to those [people]. And so those, mostly because they weren't essential for the project, which was about [topic related to the creators of the papers]. So the the letters were kind of nice vignettes, but not actually essential. We left out of the collection, they probably could have made a Fair Use claim there if we had really wanted them and we... right, that also the question we... Fair Use is such a gray area, but I only like to make those claims if I really feel I need to.

**Sara Mannheimer**  44:49

Yeah. Could you point to the letters like in the archive and did you you know, say we also reviewed some information from the letters and they're available. I don't....

**DC02** 45:02

I don't think we ended up doing that, mostly because... [phone rings... pause]. Yeah, we've done this in the past, for other projects. I think in this case, because the researcher hadn't really talked about [the letters] in any of [their] research or referenced them in [their] documentation, we just, would just have kind of collected them probably also, because [they] thought they were fun, because the letters are really entertaining. Yeah. I think we just went without them. But yeah, like pointing to data that's omitted because of copyright reasons, is definitely something we've done elsewhere.

**Sara Mannheimer**  45:51

Yeah. What are your thoughts when you do an interview, like would the consent form give permission? It's basically like a license to publish the person's own...?

**DC02**  46:12

I know, there are some discussions about this with copyright, I feel like especially when the consent form has, has the data sharing clause that it's an implicit license to publish. And I know that the legal situation is maybe a little gray, I think it's clearer in the U.S. than and I think UK Data is more worried about this, I think they have actually built-in copyright transfer, or some license into their, some of their consent forms. I would worry that that's a deterrent and also potentially unethical. And unclear what that even means for an interview. So I, I'd worry about writing too much legalese in there. And I don't know. Yeah. Do you know of a single case where this has become a problem?

**Sara Mannheimer**  47:08

Not for qualitative research. You know, it's more of a problem with social media, because the companies are, you know... The text itself belongs to the person who posted it. But then the companies have all these licensing rights over it. So that's where intellectual property can get more sticky.

**DC02** 47:32

Yeah, we do occasionally come across licensing when people use... and we are fortunate enough to know that they used databases for historical documents, right? So you get 19th century newspapers from Gale, right? They're technically in the public domain. But you essentially signed away your firstborn. When when you started using Gale Online. And and so.

**Sara Mannheimer**  47:57

That's a really good analogy. Yeah. Yeah. And so have when you have situations like that, what have you done? Like, have you been able to contact Gale, or like find the newspapers in a different location?

**DC02**  48:14

No, honestly, we haven't even tried to contact Gale because it seemed hopeless. And that particular case? I don't... that, the Gale example was quite some time ago. And I think we just used... we just used a literature list essentially, right? I mean, as much as you hate Gale, the advantage is that they're relatively good at digital preservation. So you need the paid access to Gale, but the documents are unlikely to go away.

**Sara Mannheimer**  48:47

Right.

**DC02** 48:48

So it's an access issue, but not a preservation of content issue that you're then facing. Which is not ideal. I like to make access easy, but.

**Sara Mannheimer**  49:02

Okay, cool. Yeah.

**DC02** 49:05

And then we do have, as I mentioned, some social media data, right. So like, selected tweets, etc. And I think you probably have presented on this a couple of times. So you probably know most of what I, what I think about this, but because it's so small scale, you get around the developer API licensing because [the API] was never used. So you know if you just look up 50 tweets by people using the standard Twitter interface, you never agree to the to the terms of the developer API, which are more restrictive. And I actually think that's a reasonable understanding, right? If newspapers, say, were to reprint five tweets from people even if they didn't embed them, I don't think Twitter would complain. That would probably fall in that case under Fair Use or something along those lines. And so I think this is not just a loophole, but this is actually a proper understanding of the interests that the companies want to preserve by putting those terms on the developers slash API licenses. Right.

**Sara Mannheimer** 50:25

What about the the people who tweet, themselves, though, you know,

**DC02**  50:30

Yeah.

**Sara Mannheimer**  50:30

So that your tweet's in the newspaper.

**DC02**  50:32

Right. And so that is a case by case determination. What we looked at, in those cases were tweets that, and I mean, we actually took that from, from, from your paper, right, the idea of the expectation of privacy, right. And these were very clearly public-oriented tweets, right. These were people who are often writing in their professional capacity for a broad audience, and not people sharing private thoughts in what they assumed to be a circle of friends. And I thought, I always thought that was a very useful, useful way to think about social media posts, what did people expect, which happened to them? And that's what we apply. And we've never had the other scenario where we had kind of more private social media information.

**Sara Mannheimer**  51:31

Yeah. Okay, great. Well, are there any issues or challenges that happened during your example that I haven't asked you about?

**DC02** 51:44

Yes, the one issue that actually came up with [the example] that we're discussing discussing internally is our licensing on the data. And so we tend... we have a non-open license on all of our data, specifically to prevent reproduction anywhere else. So with the idea that, especially human participants data are, we don't want them, you know, just flying around wild on the internet, even if they're deidentified, because who knows what deidentified means, right. We'll make exemptions for like, if people want an open license on their non human participants data like documents, etc, We'll apply it, but we default to a non open non standard license. Which isn't ideal for interoperability, but in the end, we think that the ethics and the human participant aspects outweigh the importance of having an open license on... The documentation is always CC-BY license. So this is just about, about the data themselves. And so there we actually, and that's unusual, again, [the researchers in the example discussed in this interview] would have preferred an open license on on their data. And we pushed back a little bit and said, "We don't typically do this, even if the data are low sensitivity." Right, there was a promise of confidentiality. And once they're out there in the internet, and you don't know who has them and how they're being triangulated, et cetera, et cetera. That's not a reasonable promise you can make anymore. So so that was kind of a perhaps slightly unusual challenge, but kind of I do think, our licensing of the data, that might actually be a relevant parallel, I don't know what other people do that provide access to big social data.

**Sara Mannheimer**  53:50

Yeah, you're... so yeah, that is a good parallel. What's the non-open license? Do you have like a standardized something that you use or?

**DC02**  53:59

We have a standard download agreement that's just a custom agreement [created by the organization where DC02 works], but it's essentially education and... teaching and education only, no commercial use, no reproduction, no attempts to reidentify participants. Those are the key points and uh.

**Sara Mannheimer**  54:17

Okay. Yeah, very interesting. Okay.

**DC02** 54:26

And then formats are actually an interesting like, most researchers don't have strong opinions on data formats, but preservation for formats of textual data are actually, a surprisingly interesting and and weird field we typically use PDF/As because, you know, that's what everyone says to use. But in [this] case, [the researchers] wanted the materials also in a more easily changeable format. So they also there as ODTs. And yeah, I go back and forth about the PDF/As, it's nice that everyone is using it, so it's a standard, but in terms of how easily you can actually use it, if you want to explore data from it, if you want to extract text, etc, having something like ODT, or DOCX, right, like XML based formats might actually end up being the better solution. And neither of those is going away. I mean, especially DOCX isn't going away. So so I kind of go back and forth on whether we should change practice on that. One new thing that I'm also starting to think about more is accessibility for data, [accessibility for people with different abilities], which I think tends to be better with, with word processing documents than with PDFs, which are an accessibility nightmare.

**Sara Mannheimer**  56:04

Yeah.

**DC02** 56:06

And and so I don't know if that's a parallel or not, but that's kind of an unusual challenge that I've been starting to think about more.

**Sara Mannheimer**  56:15

Right, for reuse. Hmm. And, yeah, right. I we've had to take down all the PDFs on our website, for instance, to align with accessibility laws for a public university. So... or you can do certain things to make them accessible. But it's generally... I guess if they're OCRed they can work. But.

**DC02**  56:36

Yeah, I mean, this, they tend to not be quite as bad because they're typically single column conversions from Word. So they're typically fine. But there still tend to be more issues, is my sense. There's a DLF Working Group on accessibility that has started to do some work now, and they seem to be pushing strongly for non PDF formats. And I haven't looked in detail to why they were saying that, but that's more for internal communications and documentation they were writing about. But I noticed that and it seemed interesting, and the other scenario is, right, for automated text analysis, although they're the PDF readers tend to be quite good, right? Like if the RSV text and the Python equivalent read from PDFs quite well. So it's not a huge deal either way.

**Sara Mannheimer**  57:33

Yeah. Yeah, especially if it's a PDF, just of a word processing document, not just not like a scan of a historical doc. Yeah. Oh, this has been really fun to talk to you.

[discussion of who else Sara should interview]

**Sara Mannheimer**  1:01:36

Cool. Perfect. Thank you. Wonderful. All right. Well, what do you think? Are you convinced that there are similarities and overlap?

**DC02** 1:01:47

See, for me, the big concern, the biggest difference, and why I think it's so different, is the relationship between researcher and participant. And I think that's just... I think about this all the time when I curate qualitative data. And it's so immediately in how qualitative researchers talk about their participants and their relationship to participants and what that means for data sharing both on an ethical and protection level, but also on an epistemological level. Like they trust me to get the story, right, those sorts of things.

**Sara Mannheimer**  1:02:21

Yeah.

**DC02**  1:02:24

I think that's so essential for for what makes qualitative research... qualitative research. And how qualitative researchers think about sharing the data and why many of them are reluctant to share the data. Where as with social media researchers, I think it's often us in repositories, and our lawyers who have to put on the brakes, because they're like, oh, let's just take all of OkCupid and just put it out on GitHub.

**Sara Mannheimer**  1:02:51

Well, this is my hypothesis. And I know you're not supposed to tell this to your interviewees, but we're done. Is that because that's the big social data people aren't thinking in these sort of like human subjects ways. And so my hope is that by understanding more about the way that qualitative researchers think about human subjects and about their data, there will be some curation strategies that we can like recommend to big social researchers. So my, I guess, my sort of biggest theory is that by connecting these two communities, they can... like the qualitative researchers can provide some wisdom to the big social researchers and provide more of a like, this is... we should be thinking about social media data as human subjects, qualitative data that has all these same issues of context and, you know, privacy, etc.

**DC02**  1:03:51

Yeah, I mean, if you can get them to do that, more power to you.

**Sara Mannheimer**  1:03:58

I think the connection...

**DC02** 1:03:59

Right. So my, my, I don't want to be too skeptical. I think it's important. And I think even if you introduce a little bit more of the, you know, "let's think about the fact that there are actual humans on the other side of the tweet" into social media research. That the valuable I think, the more I talked to qualitative researchers, right, the... part of this is really kind of an empathy issue. The fact that as a qualitative researcher, you have this direct interaction with people. And you can also notice, kind of the more intense these interactions are, the more concerned and deeply involved into the thought world of their participants researchers are, and it's really fascinating how they then think about that whole relationship and how they're just much more embedded into it. You know, what would my participants want, what are my obligations? And that's also partly training, right? We just think about this more when we train qualitative researchers. But it's also just you talk to people and you realize, you know, they have a deep, rich world of their own, and you can just mess with it. Yeah, just scrape their tweets, that's not helping.

**Sara Mannheimer**  1:05:21

But I feel like it doesn't make it any less true. You know, like, the tweet does come from this deep world of their own. So I guess, I guess it's my hope that like data curators who see both types of data and who think deeply about, like, what it means to share and reuse or even just like make available the data, that we can be the connection, you know, and like, we can sort of bring some of these theories from qual— and this idea of like, being embedded in a community and understanding their humanity. And it's like, even though you're just, scoop!, grabbing a tweet, and there's like, it's still you're still swooping people's thoughts, you know, from their context. But I don't know. It's a big project.

[Discussion of Sara's PhD]

**DC02** 1:08:30

It's so good talking. Good luck on this. Once you have something to read, send it over. I'm happy to comment. Obviously, this this is close to close to my interest.

**Sara Mannheimer**  1:08:41

Awesome. Yeah, I would love that. Okay. I'll be in touch.

**DC02**  1:08:44

Bye.

**Sara Mannheimer**  1:08:45

Thank you so much. Have a good day. Bye.