DC01\_transcript\_deidentified

**SUMMARY KEYWORDS**

data, curate, researchers, irb, text mining, set, metadata, newspapers, data sets, people, big, strategy, research, context, institutional repository, standards, data management plan, questions, biometric, issues

**SPEAKERS**

Sara Mannheimer, DC01

**Sara Mannheimer** 00:00

Alright, I'm gonna start the recording. Basically, I am seeing connections between the challenges presented by qualitative data reuse and the challenges presented by research using big social data. And so I'm doing interviews with qualitative researchers, big social data researchers, and then data curators to try to build some connections there and see if there are some data curation strategies that can from one meal that you know, from qualitative research that can be used in big social research in order to promote more ethical, epistemologically sound research, and potentially to help scale up some of the qualitative data reuse strategies. So just thinking, I'm hoping that by connecting these two communities, we can make each one better, through data curation. So I'm, so that's basically my project. So I did a lit review, where I found six key issues that overlap between the two communities. And so during the, I've structured the interview around these six issues, and the interview...oh, it says 1.5 but most of the interviews have been taking just one hour. And it could be up to an hour and 15, which is why I scheduled that. So we'll just work through the questions one at a time. And if you have questions for me as we go, just let me know, it's a it's a semi-structured interview. So we can go wherever we want. This is kind of just a guide, but I want to make sure I touch on each of these, each of these issues. Okay, great. Okay, so first question. Tell me about the types of data that you usually help to curate, and then... and what your interests are regarding data curation.

**DC01** 02:05

So most typically, I'm working with my colleagues to curate a lot of environmental, biomedical, biological, bio-sciences, basically, a lot of the hard sciences, I would say more than the social sciences. Although we do do some of those. One big set we typically curate is actually a text mining of newspapers, videos, basically, communications data, that is text mining data, and we curate that set heavily and regularly.

**Sara Mannheimer** 02:39

Does that go into... Do the data sets you curate go into your institutional repository? Or is it a variety of repositories?

**DC01** 02:47

Largely we curate what is in our institutional repository. If it goes outside of our institutional repository, we can help facilitate that. But I would not say we curate that.

**Sara Mannheimer** 03:02

Yeah, okay. All right. And then I also asked you to identify a specific example, have you come up with that, you can also think about it on the spot. But the idea is, this is a type of research strategy called critical incident technique. So the idea is, we come up with an example of a time that you curated either qualitative data or big social data pretty recently, so that you have a good handle on the events from that time.

**DC01** 03:32

Absolutely. So recently, big social data, we have mentioned that text mining of newspapers and such, that is what I would consider big social data. For us, it is an automated process. The workflow into our institutional repository is largely automated with minimal input. by us as the data curators or data librarians, we did give access to members of that team of researchers so that they could actually input their own data sets without us having to, um, and it's the the goal of it was just start to go instantly public, like to be instantly publicly accessible. And that would be my example, as a data curator. I also have a recent example as a data curator and researcher in my own research, where I was trying to text mine biometrics data from a very famous we'll call it tracker, a very commonly used tracker. So my goal is to figure out common overlapping running patterns using those trackers. And this is one of the reasons why I didn't go any further with this research in the future I might but this actually scared me in such a way that I didn't feel like it could, I figured out the most common place for the most women runners to be in the evening. And my brain went "No, no, no, no, no. Yeah, we're not playing the serial killer game." Yeah. So I actually brought that research to a halt. But it included, probably between 40 and 50,000 records of individual runs, and probably between 200, 300 individuals. And that was data I mined from a specific, like biometric social media site, you know, they put its public data out there in the public. And so I observed this, and I finally stopped curating it, and just completely destroyed the data, because it was so upsetting. And I could not figure out there's no way for me to, you know, put this out there, even if it is de-identified and have it be safe.

**Sara Mannheimer** 06:00

That's an interesting ethical situation. I think I read about people using Strava data to identify where military bases were internationally. Did you read that one? So it's like, you could see the military people running loops around the base, and it just like circled a military base? Ah, yeah.

**DC01** 06:25

Yeah. It's... there is, in particular, biometric social data is a particularly big issue. Because we're sharing it without thinking about I mean, even you think about, you know, Google tracking.

**Sara Mannheimer** 06:38

Yeah. Yeah.

**DC01** 06:40

So I hope those examples suit your needs.

**Sara Mannheimer** 06:42

These are perfect. And I think, you know, you don't need to use the same example for each question. But it's helpful to have something specific so that you can think back to that project when we're talking through the questions. Were either of these projects grant funded? And did they like, did they require specific treatment through the grant?

**DC01** 07:03

The text mining of newspapers project in our institutional repository is grant based, and I think it is still grant based, it did not require any special treatment as far as things like anonymization and such because it was coming from public sources, newspapers, radios, that sort of thing.

**Sara Mannheimer** 07:26

Okay. And did So did they have a data management plan as part of their funding?

**DC01** 07:30

They did. Okay. And was, that was... that preceded my time, but it um, I believe it was looked at and checked out by one of my colleagues.

**Sara Mannheimer** 07:47

I know that data management plans can change too, as you continue through the practice.

**DC01** 07:55

I have recently, that's one of the things we do in my position is examine data management plan drafts. And so I have recently done two data management plan drafts, examining their, I guess veracity, as for their sharing for their open training, and both of those examples are, involved human participants. So it's definitely changed the way I interact with DMPs. When I'm considering social data. Yeah.

**Sara Mannheimer** 08:29

Um, let's see. Okay. I have, we've talked through this. So the data is published in the IR, and then for the second example, you decided not to publish it, in fact, to destroy? And you said that the data is openly accessible in the IR right away? Okay. And do you have plans for storing, retaining, deleting data in the future? Or will it kind of just go on indefinitely, this newspaper data?

**DC01** 09:00

At that point, because we consider it active it will go on indefinitely. We are currently developing sort of our standards now that we've realized forever is not actually forever. We're thinking about a life of 10 years and then considering cold storage after that. It's no longer public after 10 years, but we still retain it in the event it is asked for. And we're currently working on what that looks like, what does cold storage look like post 10 years. And that is if it's inactive, once it goes inactive, and the researchers, the creators are no longer touching it. And it's not actively being touched or accessed on a regular basis, then we're thinking about how we want to create an assessment plan. And then, just for storing and retaining, we have a whole work process a workflow for that. That's established.

**Sara Mannheimer** 10:04

Okay, great. Alright, so now we'll move on to the six issues which are context, data quality, data comparability, informed consent, privacy and confidentiality and intellectual property. So we'll have six questions, with some sub parts. So first, for context, I have these two quotes here for qualitative data. "Qualitative research is a process that may include deep prolonged contact and connection with research subjects, you're trying to understand the subjects within their own context. And so qualitative data can be highly context dependent. Context is actually its own source of data meaning and understanding and ignoring context when you're 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 Halavais suggests that "when we collect data from social media platforms, just as when we collected data in traditional spaces, context matters. But the context in social media may be absent or difficult to understand." And I think the same goes for any type of like big qualitative data that's coming in, like your newspaper posts are shorter pieces of text, and they're taking from a larger persona and public life. And the out of context effect is compounded when data are masked at a large scale. So this is to try and help you understand where I'm coming from in terms of context. So during your example, what challenges... and you can choose either one or both. What challenges, if any, did you encounter when trying to capture the context in which your data was collected? And did you use strategies to try and document context to support future users like metadata or linking related datasets?

**DC01** 11:59

So in the context of the newspapers, the text mining example, in many ways, newspapers sort of capture their own context. So that data is already contextualized to a certain extent, just by things with dates and the newspapers, we did set up particular metadata, so we know what newspaper, what country it's coming from, what region, etc. Those data sets include a lot of metadata that we preemptively created and laid out sort of a curation strategy for the researchers, since they're adding it on their own. We said, "Okay, here are the standards you need to meet for your metadata, we need this information per data set." Okay. Um, in the other example, I did try and contextualize the, I'll call it the biometric data. I did try and contextualize that data to a certain extent. But that contextualization involved metadata information, including geo tracking metadata that was just not shareable and therefore cannot be curated in an ethical way. Right, I could not find a good strategy for that.

**Sara Mannheimer** 13:15

So providing context sort of compromised the anonymity of the participants.

**DC01** 13:21

Exactly. Or even as anonymized, as I could make it, it put participants at random in danger, or could have, should I publish something like that.

**Sara Mannheimer** 13:38

Okay, interesting. And for the newspapers, how do people use those? Is it all coming in... It's all coming in as plain text? And so people are using it for like, big data research?

**DC01** 14:00

I believe. So yeah, that kind of comes in almost as a corpus, you see, okay.

**Sara Mannheimer** 14:07

All right. And then I have these one a... what similarities and differences... but since we're just talking big data in your interview, we'll skip those. So data quality, during your example, what challenges, if any, did you encounter when you were trying to document data quality? So that could be missing data or in social media, it could be bots, or bias, or the quality of the method.

**DC01** 14:42

In the example of the text mining, the newspaper text mining, because we were not the ones actively doing the data collection and because we set it up so we wouldn't have to check each individual data set, which is the [usual] practice... when a data set is submitted to our institutional repository, at this institution, we check it pretty thoroughly for anything that might be missing, anything that may make it unusable or non reusable. We've basically checked it for FAIR principles or, um, I think one of the things we always consistently run into when any any sort of documentation is researcher willingness, when we're trying to curate their data, their willingness to give us sufficient metadata or sufficient readme... But their willingness to participate, because often we'll give them feedback, like say, okay, hey, you're missing a chunk of data that we can ascertain, or bots or bias, we'll see quality of method and what kind of question the method but their willingness to address that is sometimes like non-existent. And that will just lead to them not going through our institutional repository, but just going through another domain repository, where standards aren't as high. So in the example of the text mining, I still occasionally look at those ones. And I will add, I tried to do a lot of "Yes, and..." strategies for talking to researchers like, "Yes, this is great, and... it will be even more accessible, even more reusable, which means you will get cited more often." I always like to emphasize like, if this is reusable, they have to cite you, so if it's more reusable, you'll get more citations. That always seems to work out for me. And then [reading from interview guide] "what factors helped you and the researcher better communicate/document data quality issues for future users? What factors prevented you from communicating data quality?" I will say in general, any factor, the factor that often helps, and this may sound rather strange, that helps communicate to the researcher is if there's a grad student involved. If they have a grad student, with a grad student is almost always willing to take direction. Established researchers are not like very much kind of like, ain't nobody got time for that sort of thing with the real researchers. And I would say that's the preventative thing is time for those researchers. So the other strategy, I would say, that has helped us in general, and really helped us with the text mining sample. And it's helped us when we've sort of executed recently, with a lot of data, sensible social and non social is providing templates and canned language, and providing templates and canned language. And that's been, so I mean, I can have like canned language and templates ready to go, it's going to happen. But I have to have those on hand. And I've had to start making templates on a regular basis that suit a given project. So in the case of social data, I now include, here's your IRB information, because you may get this DMP through but when you have to fill out the IRB, you're going to need to do that before we're willing to even look at your data. Now, and we recently actually got into contact with our IRB office and are now starting that conversation, which has been a great strategy. So now we can inform as researchers come to one or the other of us we can, each of us can inform them of the other standards.

**Sara Mannheimer** 19:07

Nice, okay. This is really fun for me, because I feel like I'm getting new ideas for my own work.

**DC01** 19:15

Let me know, let me know.

**Sara Mannheimer** 19:15

As a data librarian I do similar things reviewing DMPs. Okay. So I think that's good. Do you have data quality issues, like you know, maybe like OCR problems, or? Yeah, nothing like that. Just trying to think of other things.

**DC01** 19:45

Maybe is there a specific example of OCR issues?

**Sara Mannheimer** 19:49

I guess I was thinking like, if they're pulling it if they're automatically pulling it mining text from newspapers, maybe there are issues where the transcripts don't make sense. So are need to be cleaned up or something. But they do that before they deposit the data?

**DC01** 20:05

Yeah, I think that's our standards, you need to do that before you deposit the data. And again, if we're looking at a long term project, which is what that text mining project is that has been in place for a long time is those standards are set before we agree to house that data. Um, we're actually in the middle of another project that's similar-- not based on social data, but that it's like, you have to make sure you're hitting a certain standard, before you put this in our institutional repository. And here's the standard and basically sends sort of a— it's almost an MOU where it's like you agree, this is the standard you're going to meet. And we're agreed upon these standards. Other than that, I mean, I definitely will say we have... my supervisor calls it as good as it gets, sometimes it's gold star, here's a DOI. Yeah, like this is as good as it's going to get. And for that we have standards, it's gotta have a good title that's findable. Like someone would be able to recognize what's, the data set's gist is, we need at least a few sentences on what's in the data set. And ideally, we need to readme but we're willing to like slide on that if it's sort of built into the data set itself. So if an Excel file, they sometimes have a sheet in the Excel file, that's an explanation.

**Sara Mannheimer** 21:26

Got it.

**DC01** 21:29

But that's, I mean, in particular, the, and I would say, when it comes to social data, the standards get higher really fast. And then we always inform researchers that we are not equipped to keep that data safe, that they have to be sure it is safe to make open, because that's what we're going to do. Yeah.

**Sara Mannheimer** 21:49

Okay, great. Great. All right. On to our number three, during your example, did the issue of comparability or interoperability come up for either of the data sets? And if so, how?

**DC01** 22:05

Yes, for comparability and interoperability. We always try and ask for non proprietary file types, so plain text, CSV, that sort of thing. So that it's as interoperable as possible with as many different types of other data. So if you want to start merging or joining data sets, people can, be easily manipulable.

**Sara Mannheimer** 22:35

How about like, metadata standards for social media, this can be an issue like, you know, if you're trying to combine a Twitter data set with a Facebook data set, they come down in different formats with different metadata fields and trying to go join them to get that can be a challenge.

**DC01** 22:52

Absolutely. And I think that's we try to mean, we're not super active in this, I would say here, we're still mean, like, we're still we're still evolving. As the science evolves, we're evolving with the science. But we asked them, that's part of the readme usually is we ask them to describe what their fields are, where they are, how they got them. And that all some of these issues get solved to by requiring non proprietary data sets of CSV is very straightforward. So even if the data set is perfect, it can be made perfect by the next person who wants to reuse it for their purposes. And then, again, missing data is something we search for, and we try and weed out as much as possible. And that's something that is a standard, we will set if there's an obvious massive gap in a data set is we asked that they either explain it and why it still has to be there, or they get rid of it.

**Sara Mannheimer** 23:56

This is good. These questions, you know, all of these issues are kind of unresolved, which is why we're talking through them. So like these answers, where you're saying, "you know, we're not really sure," that's exactly what I'm looking for, you know, like, if you're not sure, I want to know. So thank you for that.

**DC01** 24:13

I will definitely say when I'm not sure. We all know that the data is a, data is a moving target and a living beast. And it's like I get asked questions about ethics that all the time and sometimes I'm just... I don't know!

**Sara Mannheimer** 24:26

Yeah, I know. Well, I'm hoping that this research will at least give a little bump to that... that will have it. I don't know that it will result in some strategies that we can use to support ethical use of big data.

**DC01** 24:41

I think the big thing right now is just make it open. And we kind of have to leave it to other people to decide whether or not it's useful data, whether or not it's good enough for them to use but right now at least it's open. At least it's there.

**Sara Mannheimer** 24:58

Yep. All right. Well, let's move on to an informed consent. So in your example, what strategies did you use, if any, to support informed consent for participants, particularly consenting for future use of the data? And I guess this would be more applicable to... Well, no, I think this could apply to either one of your projects.

**DC01** 25:20

Absolutely. Um, so the text mining one colloquially does not require informed consent. Because they're a public thing. They're public items. However, I mean, there's a lot of discussion right around whether or not that's ethical if people who were quoted in newspapers or on social media actually intended for their work to be turned in, their, their quote, to be turned into a data set or a piece of data. In terms of informed consent, number one, okay, backing up just a little bit more often than not, when people submit DMPs to our, for our feedback, they're going for a grant, and that grant preempts their application to the IRB. So I support one of the ways I support informed consent for any data set is just to say, "Hey, you've got a DMP here, you will need to fill out an IRB, make sure you do that and give examples" and my email back to them or my feedback, like, here, just get them set up with the template, here's the website for the IRB office, you're going to need to do this. And then when a data set comes to us at the other end of the....

**Sara Mannheimer** 26:43

Can I just ask you a quick question? You have them get IRB approved, prior, like at the same time as the DMP. Is that what you're saying? Or like....?

**DC01** 26:51

We encourage it. We can't police it. I wish we could. Because a lot of people, people don't fill out the IRB until the last minute. Yeah, it's not a good idea. Um, okay. But yeah, absolutely. But when a data set comes in, let's say it's just not for grant, but it's just for a publication and a data set comes in. I'm trying to think if I've seen one recently. I haven't seen one recently, but let's say my that my example of the biometrics data came in as a data set, I would, I would insist on the IRB number being part of the metadata. Okay. And that's something we've started talking about. That's still a rough idea in our at our institution is like, if we're going to house human subjects data, regardless of whether or not it's been de-[identified] or whatever, we need an IRB number to go with it. So that we're discussing whether or not that's going to become a permanent part of our metadata. That's going to be a field like DOI for the data set, IRB.

**Sara Mannheimer** 28:02

Yeah. Okay. And have you talked to your IRB about big social data? Or like, are you, or do you kind of just, yeah. Do you trust them to do their work? And then you do yours?

**DC01** 28:20

It's been so actually, I, since I have come on in this position, I have been the one to initiate that conversation, because I am...Before getting into my current role, I have been a researcher and done extensive social research around [topic relating to human subjects] in particular. I know the importance of the IRB. And so when I came on, I reached out to the IRB office and said, "Are you aware of how data management works?" And they said, "No, no," they had no earthly clue. And actually ended up with a document which I'm willing to share that we put together, I, and basically, my unit put together a document with three key points about like, here's things you need to be aware of with social, what does, what does anonymization look like? What does de-identification look like? What is the difference? You know, what, what can be the harms of not appropriately curating and managing data politically? You know, example, that text mining, if that text mining group were to say, let's talk about.... is it Ghana or and I think it's Nigeria where there's a lot of LGBTQ issues happening. If they were to text mine, those newspapers and then print, you know, quotes out in a data set that can be identified to people like you can put a lot of LGBTQ individuals in really big danger. If that's data set is out in public. Yeah, you know, how often you know the politic, the political people of Nigeria are looking for data sets from universities in the U.S., I don't know, but it's still out there, and it still needs to be thought about. So yeah, we started that conversation with the IRB. And I'm happy to share that document that we gave them basically three core recommendations for how to update their protocol and consent documents that included thinking about what happens to data after a study is completed, the reuse, that sort of thing is that they need to be a lot more emphatic about: it needs to not be reused unless people give specific consent. And that includes, let's say, for example you're using....what's a good example? Let's say you're using a, what is it to get? Have you heard of Taguette? So Taguette is a qualitative analysis tool where you can put in a corpus of interviews, and it does text mining and analysis, content analysis for you with certain tools. It's very cool, but they very specifically say on their FAQs, we are not a secure database. So when you upload your interviews, you got to be sure that it's housing it on your machine, not on their web, their cloud. So and that way, I mean, IRB had no idea. They just said, "Oh, my God, it's an analysis tool. It's fine." No, it also houses the data. So you need to be aware of that, that third party does not unintentionally provide access to it can be very identifying information in an interview.

**Sara Mannheimer** 31:33

How do you spell that tool?

**DC01** 31:35

T A G U E T T E.

**Sara Mannheimer** 31:41

I get it, like a small tag. Sweet. Okay. Great. Yeah, I would like to see that document, if you wouldn't mind sending it to me.

**DC01** 31:51

Sure.

**Sara Mannheimer** 31:52

That'd be really helpful. All right. Let's move to number five. So we're almost done here, privacy and confidentiality. During your example, what strategies did you use to protect privacy for the people represented in the data so that you already talked about your biometrics data example. Like actually just saying, I can't even use this data. Did you have considerations during the process, like when you were collecting the data and trying to doing your research? Did you consider restricted access or other types of like protecting the data? Or was it like if it can't be open, I don't want to use it at all?

**DC01** 32:37

I did consider that. I'm still considering it. But considering the danger, even if the data is anonymized. I mean, just think about putting a map in a paper somewhere with Hey, look, "Here's a point where 25 to 30 women in the dark of night run at the same time."

**Sara Mannheimer** 32:57

Yeah.

**DC01** 32:59

No. Yeah, no.

**Sara Mannheimer** 33:02

Yeah. So even like that, not the full data set, you felt was too dangerous to put out there.

**DC01** 33:10

Yeah, right. So I couldn't do any of the data. I just didn't even feel like I could publish a thesis, like a thesis statement here. Like I can't say I'm, I mean, I'm still running this through my head. And if I ever get the chance to redo it, and redo it safely, I will. But, yeah, my strategy here was just to shut it down. And, you know, part of that strategy is, I needed best practice in my head, I needed... I needed to have that switch in my brain that went, well, you know, and look, go through the levels like, okay, I don't want to show individual running patterns, for sure, like, fine, but even to look at it in the larger picture and go, even just aggregated, this can be problematic. And that's something I don't see in researchers and something are always the researchers and something I've really had to work with researchers on is making sure that switches in their brain to go, this needs special treatment. Because even if it let's say that text mining newspaper, let's say that it's de-identified as public, it's part of the, as your quote said, you know, the larger context, it's not about an individual. If that is put together, merged with another data set, there's nothing saying it couldn't be identifiable. And so just having people think critically about this, that I think is part of my goal and strategy. I mean, even with that text mining, just having them think about that carefully. And you know, having it in their brains of their graduate students like going through the data set and doing any cleaning that their graduate student go "Whoa, wait, wait, this is a very identifying sentence here. Do we need to redact?"

**Sara Mannheimer** 34:56

Um do you, are there any like resources that you direct people to regularly or is it kind of just encouraging deep thought, deep thinking?

**DC01** 35:06

A lot of it is just encouraging deep thinking. Recently, actually, there's a researcher called Casey Fiesler. Dr. Casey Fiesler. Yeah, Her work has been helpful....

**Sara Mannheimer** 35:20

Right, she's TikTok famous!

**DC01** 35:23

Really TikTok famous. But also her work just, you know, putting it out there that just because people are on social media and contributing to social media doesn't necessarily mean that that is their consent to being "data-ed," so to speak, to to to verbify the word data.

**Sara Mannheimer** 35:48

Okay, great. All right. Last question. Well, last main question. During your example, what challenges did you encounter, if any, regarding intellectual property concerns of archiving and publishing data? So, for big social data, maybe social media Terms of Service, but for the newspapers, you might have encountered intellectual property stuff with authors or papers themselves? And then what strategies did you use to address the issues?

**DC01** 36:23

So I, I don't know that we ran into anything with the newspapers. I think most of the papers that are being mined are ones that the library provides access to anyway. Or that the researchers are paying for themselves. So as long as a subscription is being paid for, it seems to be fine. Um, however I did encounter, when I started really looking at Terms of Service for my biometric data set, I did find in all likelihood, I'd have to be very careful progressing as a researcher there, because what I did in all likelihood violates user agreements in Terms of Service to some extent.

**Sara Mannheimer** 37:01

Oh, interesting. Okay. And so, did you look at the terms of service? Or were you kind of just making assumptions? And you hadn't quite looked at it yet?

**DC01** 37:14

Before I did not really look at it. Before I started the research, we started executing. And I realized when I started thinking, like maybe this shouldn't go out there public for many reasons. I realized maybe I shouldn't have even been doing this from the get go and extracting this. Shouldn't have searched this data from the get go. And I found like, I mean, it's, there's always a certain amount of leeway and legal language. I mean, it can be interpreted, but I have a saying, and that's "cover the booty." Always cover the booty when it comes to data, because you never know.

**Sara Mannheimer** 37:51

Yeah.

**DC01** 37:52

And so that was another reason, a personal reason why I'm like, I just don't want to go down this road, if that's how it's gonna go. And I need to have, you know, more legal interpretation of this, I need to know that I'm interpreting this correct way that I can do what I'm doing or not do what I'm doing.

**Sara Mannheimer** 38:10

Yeah. Did you... like, what are your thoughts on fair use for data? Do you think that this counts? Like, if you're doing research, is that an educational purpose?

**DC01** 38:26

I do, if it's for educational purposes, it's fine publication is not necessarily an educational purpose, especially if it is sponsored by private entities. This is something that actually came up recently with a data set, is the sponsor, who is a private entity, proclaimed ownership over all the data related to that, and claimed it under copyright. And so we had to argue. We had to help the researcher argue about that and say, "No, no, no. That's not cool." Because that comes into conflict with our commitment as an institution. And to the IRB.

**Sara Mannheimer** 39:05

Yeah. What happened there?

**DC01** 39:09

It's still unresolved, still going back and forth. And I imagine they will be for a while. But yeah, it's it's not the first time I've seen that. And we bring into the issue of in terms of fair use, that is a very Western definition. And this entity at this private entity had, I don't think it was particularly a Western entity. You know, offices everywhere, of course, but when you come into international copyright and fair use law, what are we talking about what I mean, how many different laws are we coming into conflict with here? Yeah. So there's a lot of there's a lot of navigating to be done. And when we entered into those situations, I mean, we just to reach out to legal, we needed legal assistance. And therefore I think a strong strategy there is definitely start building relationships with, you know, legal law school IRB. Yeah. All that jazz. And that that particular example also did involve human participants data.

**Sara Mannheimer** 40:21

Okay. Very interesting. Okay. All right. So question seven is just: are there any issues or challenges that arose during either of your examples that I haven't asked you about that you want to tell me about?

**DC01** 40:40

I think there's a particular issue with the text mining one. So we're, we've only recently started having data sets coming in, like every week, every other week, every few weeks. And until there was the manpower, until myself and my colleague, the other data librarian were added, that wasn't possible, because that has to be managed. It has to be watched, has to be occasionally checked, and we have to answer questions and troubleshoot for the researchers. So until we were added, that just wasn't a possibility to do sort of automated curation. The other issues with data sets, and this would be more of an issue with my biometrics data set is size. Our institutional repository can only house so big of a data set. So often data sets have to be broken up, which is the case of the text mining data set, I believe, we do have ways of housing, we have a back end supercomputer, [Supercomputer name], that we can use to house bigger data sets, but that adds an extra step of we need an endpoint, we need an access point for that data set, which we include as instructions on the landing page in our institutional repository. So the landing page, the institutional repository page still exists, but the data is housed elsewhere.

**Sara Mannheimer** 42:10

Okay. And that... do you use Globus to transfer it?

**DC01** 42:14

Yeah, yep.

**Sara Mannheimer** 42:16

What, um, what system do you use for your IR? Like, what software system?

**DC01** 42:23

[IR software]

**Sara Mannheimer** 42:27

Okay. Okay. yeah.

**DC01** 42:37

Yeah.

**Sara Mannheimer** 42:38

And when so when you break up the data into different parts for the newspapers? How do you decide where to where to cut it?

**DC01** 42:50

They, the researchers are typically the ones decide where to cut it. Um, I think they're mostly doing it they do it like in a... nice thing about newspapers is there's temporal cuts. Okay, here's this month, here's that month. Okay. There's a very big temporal aspect to newspapers, but for, you know, data sets without a temporal aspect, that sort of has to be up to the researcher. And we asked them to consider like, how do people in your domain access things? So recent example is, somebody was putting together a social media data set a web scrape data set, they were going from Twitter and Facebook and I think Whatsapp, or something like that. And they wanted to accumulate it all together, and have it be one big status that was massive. So what we decided to do is go by the platform, so we have a Twitter data set, we have this data set, we have that data set, and we can help them decide how to curate that for accessibility. So we have like, a few different options we have like, we can tap curate each data set separately, we can have a parent and child. So we have overall landing page for the whole project. And then the individual child data sets for Twitter, Whatsapp, Facebook, or we can do, not the, a searchable form where they can actually search the metadata for each individual data set with a overall project landing page. So we try to offer variation so that they can tell us how their constituency accesses data and the way that would be most conducive and intuitive to them.

**Sara Mannheimer** 44:33

Nice. Okay.

**DC01** 44:35

Sorry. That was a long explanation.

**Sara Mannheimer** 44:37

That's good. No, this is great. All right. Anything else?

**DC01** 44:49

I think the only other big challenge I would mention in any given social data set is just the onus to make the right choice for the people who are involved. And that's something that is not fully addressed in big conversations at my institution, like it just hasn't become part of the conversation yet. Like that we need to protect participants and that data curation is part of that. Now, me it's not just about cybersecurity. It's not just about keeping data safe. It's about making the right choices for data that could conceivably harm people.

**Sara Mannheimer** 45:33

Great. Thank you so much.