BSR02_transcript_deidentified

SUMMARY KEYWORDS
wikipedia, people, data, documented, context, paper, database, community, big, conflict, metadata, interviews, bots, librarians, project, method, cases, dumps, feel, public

SPEAKERS
Sara Mannheimer, BSR02

Sara Mannheimer 00:01
Alright, so basically the project I'm working on is about connecting two communities of practice, using strategies from qualitative data curation to support big social research. So my hope is that by connecting a community that's more established and thinking about complex issues around data sharing with a community that is newer and hasn't done as much thinking in that regard, we'll be able to maybe use some of the data curation strategies that we've developed for qualitative data sharing, to support big social data sharing, or big social research as well. So I'm talking to big social researchers, I'm talking to qualitative researchers who have shared their data and I'm talking to data curators, with the goal of developing some strategies that librarians and data curators can use to support these two types of research. And so I did a lit review where I identified six key issues that are challenges for both types of research, context, data quality, data comparability, privacy and confidentiality, informed consent and in-, intellectual property. So this interview will be structured around those six topics, and I'll ask a question for each of them. And I have a intro and a wrap up question as well. So we'll have about seven, seven questions, eight questions. And it should take about an hour, we'll see. Okay, you ready to start? Okay. So tell me about the type of research that you do. And then what type of data that you produce generally.

BSR02 01:44
Okay, um, I do too many different kinds of things for my own good. I very much a kind of disciplinary nomad, I kind of have, you know, I feel like one foot in, in in everything, which sometimes leaves me like, I'm on a on a on a chasm, but I am in my, my current, I don't know. So I'm currently in a joint appointment between the [communications department], which is in the social sciences, and the [data science unit], which is [engineering focused]. And so I am literally, in between these two fields. When I'm when I kind of have my, my kind of communication hat on and my kind of qualitative social science head on I am often more of a, someone who does interviews and who does ethnography, I was trained as an ethnographer, to doing a lot of like, in person participant observation of studying how scientists work. And so I've done a lot of kind of actually similar kinds of topics about like, how do scientists like and I've worked with data, especially, it's been a lot of time studying ecologist and in my grad program, and how they were, you know, trying to do kind of global ecology and data integration and things like that. I've also spent a lot of time with a more kind of decentralized communities of knowledge production and dissemination. So I've done a lot of work on on Wikipedia, and how Wikipedians collaborate, and how kind of conflict emerges and how they resolve conflict. I've done some work on on Twitter on how like social people who are users of Twitter, I've done, did a project on on people who have been harassed on Twitter and the subject of kind of coordinated harassment campaigns, and how kind of their experiences and so I feel like I could have sometimes my, when I'm on the more like social science of things, my methods are often more, you know, interviews, observations, qualitative kind of

small data kind of stuff. But I have done more work that has led me into the, I guess, more of the big social data space, especially when it comes to, and those have often been in collaboration with people who have more of the computational skills. And so I so I would say, the most were the work that I've done with more big data methods has been on Wikipedia, where the data is open and accessible. And so we've done a lot of things on like, I've done some stuff on, like, what are newcomers experiences, like when they join when they first start contributing to Wikipedia? And so in that case, you can, you know, there's like interview methods, and then like, opposite, like more like just reading of the archives in the forums, and, you know, looking at how using kind of more traditional humanities inflected kind of methods and theories. And then I've also done more kind of trace data analysis of, you know, so like looking at, like, who are the people who actually stay and leave, we can sort of segment those people computationally. We can see if, you know, did did risk doing kind of more more things where there are p values, you know, where it's it's regressions of, you know, did this thing happen to a newcomer and did that, you know, does that predict whether they continue to contribute or whether they stopped contributing.

Sara Mannheimer 04:57
Hm, okay. I'm wondering if I should interview you as a qualitative researcher. So I think, it's actually really perfect for me. Because I, like I have sort of seeing that these two communities aren't connecting as much as they could. So I feel like you're the perfect person to talk to you about.... Okay, so we'll try and like, yeah, I will see, let's talk about a big like....so the way I'm doing this is using critical incident technique, it's called. So using like, one specific time that you, one specific project, it can kind of help get it some more detailed specifics. But you can also feel free to go beyond your critical incident if you need to, or if you want to. So let's first identify an example when you either collected big social data for research or used big social data or prepared big social data for sharing, or considered sharing your data. Maybe this Wikipedia example is a good one.

BSR02 06:07
And I've got I've got the exact paper in mind that that was a big one. Yeah, I can share the link with you too, because we actually tried to go into a lot of methodological detail about how we did it. I've actually got, I can share with you this massive like, just to give you a scale of this, I'll do a screen share real quick, but oh, no, nevermind. But yeah, I can, I can share it with you many ways... the project was around, this is actually a project we did that was a rebuttal of another study, that, in our view, was making a claim about Wikipedia, and they were trying to detect conflict and Wikipedia. And based on sort of my ethnographic understanding of what goes on in that community, I think that they were just absolutely wrong. And so, and we've been very public about this, it's a was a, you know, we, we made a lot of criticisms of how they were operationalizing conflict. And we thought that that sort of didn't align. And so we, we did this sort of big data, small data project, where we wanted to find all these cases where their method had identified these cases as conflict. And we did not think based on, you know, my ethnographic understanding of the situation....

Sara Mannheimer 07:23
Oh, okay.

BSR02 07:23
...that in the Wikipedia sense of things, like it was not conflict, because it was actually either cooperation or just like a non-event. And so what we tried to do is I worked with my, I was I was a lead author on the project, and I was working with my longtime computer science collaborator. And actually, we kind of flipped our normal roles that we did. So I was actually taking the lead on the Big Data part of things. And he did more like he did interviews with some of the people who were involved and kind of

talked to them about their experiences of these events. And so we were kind of trying to, like invert that traditional like, we're like, typically, like, they do like the method first, and they did the findings. But we wanted to kind of start with these cases and kind of see where their method broke down. And then we wanted to then replicate, we wanted to kind of produce our own quantitative metric, or not metric, but like quantitative operationalization, where we can feed in all this like edit activity data about who does what, and like, who makes an edit to a page, and then who like, undoes that edit and replacing it with something else, and then put, who puts it back? And, and so we wanted to have have our own method for detecting conflict based on those kinds of those kinds of metadata and records. And so we did that. And yeah, we did a whole lot of work documenting it. And you know, it's something that goes through and can scrape through the entire full history of every change that's made every Wikipedia article in the multiple language versions that Wikipedia has. Actually, I think we only did five or seven language versions, but like, still, like, it was a big effort. Yeah, you know, so that's, that's the context of this. We ended up making our entire data and pipelines, so not just the data, but also all the all the scripts, all the Python, all the Jupyter notebooks, all the builds, like that's all on GitHub and Zenodo. It's all well documented. I have this like, massive, ridiculous flowchart diagram of like, "Okay, we've got the, the Wikipedia SQL database that exports into this format, and then, you know, now it's in a XML format, and then we have the XML parser." And so I don't know, it's, it's, we actually have that pretty well documented, if you want to ever follow through on that later. So.

Sara Mannheimer 09:37
Okay, yeah, I'll take a look. Perfect. That's a great example. Okay, so your data collection method was scraping the web when you did this?

BSR02 09:50
So thankfully, Wikipedia has their database in so they have a SQL database that they use for production, you know, that like that. Like the servers themselves that are hosting Wikipedia, you know, are structured in that way. And they regularly process actually I can I, at the time, it was someone named Ariel whose job it was to make I actually have this all I've got this this thing that I've got my documentation up next. And so they regularly export that database to XML. And they publish those to the web. So we just could download the XML documents.

Sara Mannheimer 10:25
Nice. Okay. Perfect. And did you have grant funding for the project that required specific treatment of the data?

BSR02 10:34
It was, so we didn't receive any money from Wikipedia, but because the data comes from Wikipedia, that data is is Creative Commons licensed. And so you know, we have to, you know, continue to, yeah, so it's like in in our processing of it. We released that under the same license. And I also was funded under the so... I was at [Institution] at the time, and that was funded through a grant by [Funder]. And so they have a, an open access for the for the for the publications, but not for the data. So the publication is, is open access, because it was funded by this grant.

Sara Mannheimer 11:24
Okay, cool. And did you have a data management plan for the project?

BSR02 11:31
No, I'll just be honest. We didn't have we didn't have a DMP, like we had, I would say we had plans about data management, we did not have a DMP. If that is a distinction that makes sense.

Sara Mannheimer 11:45
For sure. For sure. Alright, so you didn't have written plans, like a written policy document or anything, but you knew what you're planning to do?

BSR02 11:55
Yeah, we've worked with this data before. So yeah, we were we were pretty ready to do that.

Sara Mannheimer 12:01
Great. Okay, so you, and you published in Zenodo and GitHub, right?

BSR02 12:07
And we put it everywhere. I think it's on there, it's on, The California Digital Library has its own? Yeah, lots of copies, keep stuff safe.

Sara Mannheimer 12:18
So okay. Are you, do you have information science background?

BSR02 12:23
Mm hmm, I'm a mutt so yeah.

Sara Mannheimer 12:28
Like you talk like a librarian.

BSR02 12:31
Some of my best friends are librarians actually, the [data science unit where I worked was situated in the] library. Oh, no, we actually, that was like the neutral spot on campus. And so I got to know a lot of librarians. My PhD is from a school of information, although we were one of the schools of information that like got rid of our library program, back in the '90s so.

Sara Mannheimer 12:51
Okay, well, let's jump into these issues. The first one is context. So I have a quote here from that says, "Halavais suggests that when we collect data from social media platforms, just as when we collect data in traditional spaces, context matters. However, 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 a larger context of personal and public life, this out of context effect is only compounded when data are masked at a large scale." So I guess in your case, these posts that we were working with are like people's edits, right? And like their interactions between each other as they're working on Wikipedia. So this is sort of where I'm coming from, in terms of context, just thinking of like, context in the social media realm, similar to context and like a qualitative space where it's like, there's so much that you don't know, and that you may not be able to document. So can you tell me about a time if there were any, during the process of your example, when you considered the issue of maintaining and understanding the data's context? So like contextual information about Wikipedians themselves, or contextual information about how you collected the data or who you are and who your collaborators are? And then what were those considerations and how did you address those issues?

BSR02 14:16
Yeah, yeah, yeah, so context was the heart of this, because, you know, that was the point of this paper is that we felt that the original study had not kind of gotten the proper context, you know, and so we felt

like they took a very kind of, I guess, data driven is the word for like to use maybe, but you know, you know, kind of came from that, that that approach. And, and so we wanted to kind of put the context back in. So that was a central part of this. And part of that was that you know, both me and my collaborator, you know, I I'm more heavy on the qualitative humanities side. He's been trained as a computer scientist, but we both have had more than a decade of experience in Wikipedia. Me more from like an ethnographic perspective, we've been a long time kind of studying this project, he just as a volunteer contributor, and you know, was just very always involved, and then, and then made it a part of like his dissertation work as well as a computer scientist, and then actually was then working at the Wikimedia Foundation kind of doing, you know, when we, after his after his PhD, and so both of us had a kind of understanding of the culture of Wikipedia, and the culture of interactions between between people. And also like, between, there's a whole, like, part of the paper of a lot of people really centers on automated accounts on like bot accounts. And kind of, because those are often and so one of things that if you don't, if you like, like one of the things that if you don't know a lot about how those accounts work, you might like it when people hear bots, and they think, "Oh, it's like spam bots," or things like that. But one of the things that I know a lot about, I spent my dissertation writing about it, but it's about all these, like pro-social bots that are very, like, authorized by the community. And they're, there's a lot of debates that happen, like some of them, you know. And so they do kind of like a lot of like, routine work maintenance work, kind of finding replaces cleaning stuff up. And, you know, integrating data, and there's all kinds of community processes around approving those and raising issues when they look like they're messing up or going outside their scope of approval. And so we really knew a lot about that. And so we wanted to make sure that we were treating that with the proper context, and that we understood that, you know, if there was a particular pattern of interactions, you you need to actually look at the content of what those interactions were, and you needed to also look at, you know, who the who these accounts were like, if they were bots accounts, we need to treat them a little bit differently. And if there's certain kinds of like, and so there's certain things about like, the Wikipedia bureaucracy that I feel like we we knew about, and so we were able to bring that to bear. But to be honest, like, it's not data management for that is, like, hard, and I don't think I think we kept a lot of that in our heads. And I feel like we were much better about documenting everything we were doing on the kind of quantitative side of things. But you know, we did, we did sort of do a lot of work with Jupyter Notebooks, and, like, starting off with the quantitative sort of approach, where we would just use like, their, their metric and their their sort of method for identifying conflict, cases of conflict. And so we'd throw those in. And we had a Jupyter Notebook where we could just like, look through all these, and then we could sort of filter in, and we could sort of look through and and I could sort of print the full context of an interaction and say, "Oh, I don't think that's right." Or "I think that is right, I think I think their method captured this one, this seems like real conflict to me, this doesn't seem like real conflict to me." And then we could then like, go back up to the top of the notebook and like fiddle with some some things in the in the in the kind of the programming language, and then run it again and see, oh, does that exclude the thing that we want excluded? Does it keep in the thing that we want kept? And so that was kind of our process of like, putting the context back in. And then we saved all those notebooks. And those are available.

Sara Mannheimer 18:08
Nice. That's so interesting. Okay, and did you think about like, this is coming up about comparing data comparability, but like, I'm thinking about previous studies that you and your collaborator had done with Wikipedia? Did those come into it at all? Or was, did you just use the knowledge that you had about context? Like, when you're publishing this data, would you connect it to other articles you've written and say, here's how we know that, that Wikipedians conflict looks like this for Wikipedia?

BSR02 18:46

Yeah, it was, it was largely, I think we wanted to especially for this project, I think we want to make it also the kind of thing that you could give, like, it was like, a bit pedagogical and kind of walking these through from the beginning. So I think we did a lot of that just kind of from scratch, even though we could have. But actually, I think though, I think this paper was like more of a critique of a lot of the other prior work that even maybe we had done, that that didn't, you know, that was more based on just, let's just do the metadata. Let's not look at content, let's not look at context. Like it's just that that's just harder. And we've, we've published a lot more papers, and we've published other papers that are just, you know, just kind of counting things with metadata and kind of looking at trends up and down and not not looking at too much about the contact, content of what's actually changed. So.

Sara Mannheimer 19:41
Okay, let me ask you this last question that's on the bullets here. Since you published your own data here, what strategies did you then use when you were publishing to communicate the context of your data to future users, you know, so you'd gone through the you know, process of the study that you were critiquing and redoing. What about for yours? Like, if someone's trying to reuse your data? What strategies did you use to make sure that they understood where you were coming from?

BSR02 20:12
Yeah, so we have, I actually want to, like, refresh my memory here, and like, I'm going to the repository itself. Let's see, we have, yeah, no. So we do a pretty like this is I'm actually pretty proud, like, I led the effort on the data and, so I'm first author on that paper, right. And so I'm going through just this documentation file. And, you know, we talked about each of the data sets and how we generated them, and really made it so that like, you know, if you're going to rerun our project from scratch, like you'll get the same, like, it's reproducible in that sense. And talking about how it's sort of getting all these different stuff from different places. And then and then I don't know, so actually, we have, it's more of it's more of like just giving documentation about like, what each like, like, I have a big table of our data, our data data dictionary table, to kind of say, like, okay, here's the field name. Here's the like, description of like, how we got there, and like a little bit of context, like, where in the pipeline was it created. And then some examples of each of those. And if like, that kind of captures, because like, some of these are ones that are like, straight from the database. And you know, they're they're, if you download the raw XML dumps, you'll get them. But then some of the if it's like, ah, no, this is, this is a metadata. This is like another metadata calculated field that's like generated in this script. This is our, like, we do a lot of stuff of like, we built the parsing engine for like parsing the actual text of certain things. And so it's, you know, here's, here's our, like, calculated type of interaction of like conflict or not in conflict, here's where that comes from. But really, I think we were imagining that the data documentation itself, and like, basically everything that we upload to Zenodo, and Figshare, and all the good places. I don't think that stands alone. Like, I feel like you need to read the paper. And we open the paper with like, a vignette of a case that and we like, present it, you know, like, I feel like you really need like, and I don't know if you put like a warning in that, but I feel like I was not, I don't think that if you read everything that's in Zenodo, you'll get any of that context, you'll just get a reproducible pipeline to like, I want to get the context, you need to read the paper, and I don't have any, like, I think I actually tried this a little bit. And I was just like, I can't figure out a way to do this. Like, if you want to get the context, read the paper, because the paper is everything about how we got the context into the data. You know, if that makes sense.

Sara Mannheimer 22:52
Right. For sure. Yeah.

BSR02 22:53

So if there's any, so maybe, maybe your research might figure out some ways to like, help us with that. But I think I actually just remember struggling with being able to do that. And I was just yeah, forget it. I'll just, you know, the paper will stand out.

Sara Mannheimer 23:05
Yeah, for sure. Huh. Okay, side note, what motivated you to go to all this time and effort to make sure that your data was so well documented and available?

BSR02 23:17
[Detailed explanation redacted at request of participant. Summary explanation below at 27:02].

Sara Mannheimer 26:39
Okay, cool. Yeah, well, I guess part of what we wonder as data librarians is like, What does motivate people to like, do a good job, because it's not really like to do a good job of data sharing. Because there isn't that much external motivation right now. It you know, it's like, as long as you share, and you show that you've shared and shared, that's like, kind of all that seems to matter. So,

BSR02 27:02
Yeah. In a way that I'm more comfortable sharing, which is that I was contesting a paper that was widely circulated, got widespread coverage in in mass media, was led by a professor at [a well-regarded university], who led a prominent group, led by a person who has given keynote addresses at major conferences and [these conferences] were continuing to celebrate his work. We were contesting that. The stakes were very high. So we wanted to make sure we got it right. And we wanted to make sure that people had trust in our work. And that we knew that this might be something where we had a bunch of back and forth, because they also didn't document their methods sort of as well as we would have liked. And so we wanted to make sure that everything was as documented as we could, because it was going to be a matter of kind of contestation and debate.

Sara Mannheimer 27:52
Okay, okay.

BSR02 27:53
That's for the record.

Sara Mannheimer 27:56
Great, let's move to data quality, I want to make sure, try and get us finished.

BSR02 27:59
Okay, we can a few minutes more. Yeah.

Sara Mannheimer 28:02
Okay, so tell me about a time during the process of your Wikipedia work when you consider the issue of data quality. So you kind of talked about like their misunderstanding of bots. But I guess you could also talk about if there were problems with data quality in the Wikipedia database, or if there was any other quality of your own method and how you addressed that or how you communicated the quality of your method when you ended up publishing the data later?

BSR02 28:41

Yeah, it's a good question. I'm, I'm maybe struggling to think of the specifics. This was a few years ago. Definitely, there's a lot of stuff. I don't know, this is the kind of thing when you're working with text, like encodings and Unicode are just very frustrating. So there would often be times when like, something didn't work, because we're converting from so many multiple formats and encodings that sometimes, you know, you get, and also like people, you were dealing with multiple languages, which don't all have just the standard Latin character sets. So yeah, I remember I remember just Unicode being something that I was fighting a lot over in terms of that, in being able to get that in. I feel like I trust the Wikipedia data. Oh, ha! I got an example. We don't trust the Wikipedia database before 2004. We had, Wikipedia exists from 2001 to the present. In 2004, they did a massive redesign and it went from being something that was like very janky and very, like a very like old platform. And they kind of modernized and so they they there is so you still have the Edit history that goes back for like some of the older articles that goes back to 2001 by just after something that like, you know, in the Wikipedia, computational research community, we've talked about it before, like, I know people at the foundation who make the dumps. And so it's just sort of one of those things that, that we know that we cannot be confident at anything, the database says before 2004, before a certain time in 2004, when the new migration happened. And so we actually just sort of limited, we're like, we don't care about those early days. And so we can just start when the new database, we cut off our data set, I believe, starting when the transition happened. And so that's that's the kind of example that like, I know that from just my participation in the rather tight knit, like Wikipedia research community, you know, there's, there's like a comp, there's, there's a wiki for it has its own wiki, of course, and people talk to people document all that all this kind of stuff around, you know, the database was upgraded, has a new format. And so that changes how the XML dumps get processed. So is that the kind of thing that maybe is an example?

Sara Mannheimer 30:56
Yeah, definitely. Yeah. And then I'm also, I think a lot of what you talked about with context is relevant here, too, like documenting your data really well, to make sure that future users understand its quality, and the quality of your scripts and all of that. So I think we can apply those here as well.

BSR02 31:17
Yeah, we also use a couple of like, some of our fields in our kind of model that are not our model or kind of classifier that we built for detecting conflict, non conflict require imputed metadata forms, like, like, how many bytes were changed between one version of a page and another version of the page, and the characters were changed. And so we, we calculate those, and then we store those in a separate kind of table. And so in our doc, we definitely have, you know, this is where this calculated field comes from. This is actually like, this script comes from there. If you want to poke and you want to change how we, you know, in calculated those fields, you can you can do that, if you don't, you know, trust us to make those, those those or you want to do it a different way. So that was also something that was important for us.

Sara Mannheimer 32:03
Okay, great. Let's move to data comparability. So during your example, did you compare or combine multiple big social data sets? Or did you consider comparability or interoperability of your data set?

BSR02 32:20
Good question. Um, I don't, we did have to combine from a couple different Wikipedia sources. So I'm actually just looking at this in terms of our big table. We have one big, which is like the entire edit history of everything that's happened to every article on Wikipedia. And then because one of the things that was very important were these like bots, and different bots do different kinds of things, we had another kind of database of, of, of bots, and like when they were approved, or if they, because like,

actually, we did have some cases where like, someone created a bot, it got approved by the community, and then the bot developer, used their bot to do a whole bunch of stuff that they weren't supposed to do. And because they had a bot account, they could like edit every page in Wikipedia at once. And, you know, do a find/replace across the entire encyclopedia. People didn't like that, because they you know, so we have a couple of those that was like a more kind of curated list of, okay, we've got this big table of all these bots, and when they were, you know, and so we did have that that came from. And so we had to combine those we do have like, that's a little bit where we're like filtering out some of that, that definitely took some like hand curating. And we actually improved, we improved Wikipedia's own kind of internal table of how it manages that, as part of this project. We were like, we're doing this work anyway, why don't we do it on wiki data, which is this kind of relational database. Actually, that's not relational. It's a graph database that they're using for it was very new at the time. And so we just said, "Let's improve that version, and then build the pipeline that will query from Wikidata and the Wikipedia database." And so then we have our, our pipeline, just kind of, we made the edits to Wikidata data, the Wikidata database on the bot table, and then our pipeline just grabs from Wikidata instead of grabbing from our own like hand curated version.

Sara Mannheimer 34:14
Got it. Okay, nice. That's a great example. Cool. Um, do you consult with anyone about like, how to support the comp-, the comparing between those two? Or was that just like you knew how it should work? And so changed it accordingly?

BSR02 34:34
Yeah, I think we, I had to look up because Wikidata was new at the time, and it's kind of what they're using to kind of replace, you know, some older database forms. And so they were using it to keep track of, you know, which bots were approved when, and so I remember having to like read up on the documentation for that, but it was pretty well documented. And so I felt like it didn't I didn't need to ask anyone anything, because there was good documentation.

Sara Mannheimer 35:03
Wikipedia is good as a data source.

BSR02 35:06
Oh, yeah. Yeah.

Sara Mannheimer 35:09
All right. Um, the last three issues we'll talk about are more ethical and legal issues. So first, informed consent. Can you tell me about a time any during the process of your work when you considered informed consent?

BSR02 35:25
Yeah, yeah. So we do consider the the database of everything that happens on Wikipedia to be public. There are issues around you know, I know this, like not all Twitter users know that Twitter is public, by default. There's kind of those issues that we think about. We do, because on Wikipedia, I mean, it's, we we tend to treat that more as public data. I've had, I've had IRB kind of determinations of non, you know, like, I forget the technical phrase, but I've, I've gotten one of the documents, it's like from our IRB, if it says, if you're working with the Wikipedia public data set, we don't consider that to be human subjects research. Even though that might be I don't know, personally, I might say that they may be a little bit looser on that than they should be, I don't know. But I do, I do think that because you do have to, like, every time you save an edit on Wikipedia, you actually, you have to license it under a Creative

Commons license. And so there is a like, you know, even if you edit from an IP address that will say, like, your IP, like, if you don't create an account, your IP address will be public. So they do provide a lot of like in the interface, I feel like it's more so than Twitter, in that sense. So I'm more comfortable treating that as a public data set. You know, we did do some member checking. And I do have an existing IRB that kind of covers, you know, so so I wouldn't necessarily, I wasn't sort of treating them in the same way that I was treating interviews. But I did have, like a very broad, existing IRB protocol that did cover the kinds of, like, interviews that me and my collaborator did around.... So basically, basically, what we did is we had these cases that were identified as conflict, and then we would give them to people. And then we would, and we did most of this actually over email instead of interviews. But it was sort of more like, we wanted to have this in a way that we said, you know, "We're going to send this to you." And one of the things that we know we wanted to take care about was whether we could identify those people in the paper. And this is where Wikipedians are actually a bit different when it comes to like informed consent, where Wikipedia is it's a very public space. And so often, the informed consent is not around anonymity. But actually, like, they want credit for what they say, you know, it's like they want to get, and so like, de-identifying them is an act of, I don't know if I'd call it like an act of violence, but an act of erasure, it's an act, it's like, erasing their labor. And so you know, they mo-, and so we found that, actually, so we did double check, we kind of, we would have this interaction with them, we would say, "You know, we're writing this paper, it's covered under this, and you know, we're interested in this we want to use, we're interested in using your case, as an example. Can you talk about whether you thought this was conflict or not conflict? And then, you know, would you like us to, you know, actually detail this case in the paper," and we only used ones where the people said that they would be comfortable having their usernames and their case be described in our, in our vignettes and case studies.

Sara Mannheimer 38:26
And most Wikipedia users use pseudonyms anyway, in their username, right? I feel like most people don't use their real name?

BSR02 38:35
I would say like, the more you get to be a veteran in the community, it's more common, like so your username, people do use use pseudonyms for their usernames, but they'll often then like, their if you go to their user page, their profile, they often kind of put a lot of information about themselves. Some don't. But like, yeah, some some remains, you know, very strongly pseudonymous, where they have a very strong identity, they say a lot about themselves, but they're you don't ever know, but but their actual name is beyond and they might just say, "Oh, I'm I'm an American," or something like that. But yeah, so so there is a culture where like, people are able to select the level of self disclosure on the platform. And, and so that I feel like especially the people we were talking to are much more longtime, sophisticated users, veterans who have then, you know, in these kind of we did, I don't think we did any, any newcomers, like, I think every case we did with someone who is very well established, and so I feel like has, has gotten the chance to set those out for themselves. And that's something that we also kind of know how to navigate that through our membership in the community. I actually was one of the very first when I was a grad student, like, started one of the first conversations about like researcher ethics and still have a page on like, how to ethically research Wikipedia that's on Wikipedia. And that's like still used and yeah, so we think about these things a lot.

Sara Mannheimer 39:53
Cool. Yeah, and your answer there kind of overlaps with the other two issues that we have, which is privacy and confidentiality and intellectual property. So for privacy, we've kind of talked through this idea that you, Wikipedia users may not want their information to be private, they may like expect to be

public. Was there anything else about privacy that you thought about or that you talked about with users?

BSR02 40:26
Not really. I mean, they were they were, I mean, to be honest, kind of very thin interactions that were mostly it over kind of, you know, text as opposed to interviews, we actually did have, you know, we actually found because this this, this original article got a lot of mainstream media coverage. There already was a conversation that was happening, and one of the candidates threads inside one of the community discussion spaces about it, and other people who were saying like, "This is ridiculous, these things seem wrong. I found some cases that don't apply." Yeah, so there already was a lot of like, public, internal, and all those discussions are public into the web. There's a very strong norm against having private discussions in Wikipedia. Everything is supposed to be done on the wiki, or the public chat rooms and things like that, mailing lists. And so yeah, so because of that, I feel like we were we were just in a different ethos. But we did give, because we did actually have an experience, this was like years ago, where there are some Wikipedians, who don't want to be identified and don't even want to be identified in...like, some people create tables of like the most active, like the people who have made the most edits to Wikipedia. And there's a few people who just don't even want their usernames to be on that list. Some of them like hold ideological views that are against like the counting of contributions. And they're just like, "I don't believe that that's something we should be doing. And so I want to remove myself from this list." Others are just like, "No, I don't want my username attached to that." And so there is actually a page on Wikipedia of people who have opted out of those kinds of it was more like people who have opted out of being in those like lists of the most active contributors. So we can also take a look at that. And that's something also whenever I do like a peer reviewed article, that's doing Wikipedia research, like I'll always, like, check that list. If anyone shows up on that list, I'll tell them about it and say like, "Hey, this is a list of like opt outs that Wikipedians keep for, like not wanting to be publicly mentioned," even though officially, it's all public, and it's all creative commons license, and those people would admit that that you have the legal right to do that. But it's more of like, respect for person's kind of...

Sara Mannheimer 42:37
Okay, sure. Okay, so did you go through that list and remove there anything that mentioned any, like, edits that they had made for those people? Or did you still include them in the dataset, but not? Like, talk about them in the paper?

BSR02 42:53
Yeah, so we didn't talk about them in the paper, they're still in the dataset. We didn't remove them. Because there are objections on that is more of like being mentioned as part of an aggregate rather than necessarily... I don't know, like some people are more I don't think I'm realizing like some people are more ideologically opposed to the idea of like counting users, and like by edits in general, we didn't think I think that if we had been I don't know, it's actually a good question about whether we remove them from the data set. To be honest, I don't think it ever came up to us about the like, about whether we remove those people from the data set. I think that we kind of assumed that the real harm was in in mentioning them and mentioning their usernames. We didn't I mean, I think I think we also kind of felt like because we were doing a bit of a rebuttal study, I don't know. But now now, think about the fact that someone else does a harm doesn't mean that you could do that harm again. So I don't know. So I maybe we should have thought a little bit about like whether we should remove those. But it wasn't ever kind of articulated as like, "Get my data out of Wikipedia." I think there have been some people who have tried to like get their data deleted from Wikipedia. I'm not sure how I think about that. It's a good question. So I guess I guess I'm a little split maybe on, on removing people from the record, but I think

it was more of like, we definitely didn't want to like even we didn't even reach out to anyone who was on that list. Yeah.

Sara Mannheimer 44:21
Okay. Interesting. I want... do you know, what happened with those people who tried to get their data deleted from Wikipedia has? Have people been successful in that?

BSR02 44:31
Only only for...Yeah. So no. And Wikipedia is like ideologically opposed, right to any... you know, it sees that as censorship. And often often it is people who have it's not that they're like contributors to Wikipedia who want their previous contributions removed, but they are, they are people who are depicted in Wikipedia, and they want their you know, whatever. Like, like, like a politician has something bad about them and they have hired some people to like scrub the Wikipedia article, but it keeps getting put back, because what could be like, the easiest way to get Wikipedians mad is to like do that. And so then they'll they'll like file a suit and Wikipedia gets sued, like all the time with these frivolous lawsuits, and they'll just say, "Nope, section 230. Like we have the right," you know, unless unless there's like a copyright infringement, in which case, but like Wikipedians are super copyleft. And like, they're like, they they don't want to copyright material for their own kind of ideological reasons. So anything that's copyrighted can be taken down and deleted from the database. There are things called I can actually go into this for way too long about it. There is a process though, for removing personal information. So this is actually the oh, what's it called? It's not super delete. I can find it later. But because sometimes some people get doxxed, where like, they're like, yeah, so some some it could be a public figure or some, you know, they they had privacy rights too or something, or like someone creates an article for someone who's not notable, but someone they're trying to harass, and they put, they put all their personal info and their phone number and their address. And so I don't know, social security number, I don't know, if you just do one of those things, and you're trying to harm someone. There are procedures for if, you know, especially for if like nonpublic information is put on Wikipedia in that way. Those can be deleted from the database. And there are that is that has happened, it happens far too often. People upload nasty stuff to Wikipedia, people upload child porn to Wikipedia, people upload really nasty stuff. And there is you know, it's they do have all kinds of processes in place in the same way that like Facebook, and you know, all the major social media companies have those like illegal content filters that are available, and do those purging of things that are just like blatantly illegal content. But I would say like Wikipedia is, you know, very good at like removing things that are copyrighted, and things that are like illegal according to US law. And those do get purged from the database. But there is a record of the purging.

Sara Mannheimer 47:03
Right, okay.

BSR02 47:06
And the whole process, there's like a council that like, it's like one of the only secret parts of Wikipedia, but it's still like, its secrecy is documented I find, I've been meaning to write a paper about this. I find it fascinating.

Sara Mannheimer 47:17
Hmm.

BSR02 47:18
I mean, I probably won't, but I can give you pointers more if you're interested.

Sara Mannheimer 47:23
I think this is good. Okay, let's move to intellectual property, which we did discuss because basically everything on Wikipedia is CC licensed. Right. Okay.

BSR02 47:33
That was done. It was easy.

Sara Mannheimer 47:35
Yeah. Yeah, I kind of feel that is kind of all that we need to write here.

BSR02 47:41
I mean, we also released all of our analysis code under—what did we do? I think we probably dual... I think we put a pretty permissive license on it. I think we dual licensed like CC and MIT for the code and documentation. Yeah, because we we don't see this as proprietary. You know, it's, it's, you know, we actually want this to be something that you know, because we were we were we were we were gonna fight. So we wanted, you know, as permissive as possible, remix this thing. We put a Binder up even so that people could I don't know if you know about MyBinder. Yes, we put up we like it works with Binder it works with, like, we're like "fork it, do whatever you want. Use it in your classes." Like no, we really wanted it to be like the spread, we wanted a permissive as possible copyright license. So.

48:32
But you succeeded, right? I mean, no one fought back? It was just those two little like...? Yeah, yeah.

48:38
Yeah, so some, some a couple of data science classes I know from some people have like, use this a bit as a, because we do it's like a good thing for students, because you can like walk through all of our Jupyter notebooks and our pipelines and stuff like that. So it's like a real world project. So yeah, it has gotten some uptake and some other educational settings. So yeah, we were we were happy about that.

Sara Mannheimer 49:01
Really cool. Okay, are there any other issues or challenges that arose during this example that I haven't asked you about that you want to tell me?

BSR02 49:10
Um, we did run into a bit of like, we want to, like the raw XML are actually not the raw but like, even the compressed like Gz, compressed XML dumps are 93 gigabytes, so that we wanted to actually save the version of it, you know, because, you know, we wanted to, for reproducibility purposes, want to get the exact same figures, and Wikipedia, like, deletes, dumps, like, they only keep dumps for a certain period of time, because they have to, like, you can't go get the now like October 2013 dumps from XML dumps because they just don't, they don't yeah, so we knew that they delete them after a period of time, but we wanted to keep our entire pipeline. So that was something we wanted to have, but then we tried to figure out where do we put those 93 gigabytes? It was like it's like one big I don't know, like, it was too big for Zenodo by default, and it was too big for Figshare by default. And so I think we actually contacted Zenodo. And we said, "Hey, like, we know that y'all are at CERN and do a bunch of stuff, can we get an exception?" We didn't hear back from them in the time period that we needed, I think it took them... And so we actually went to [a large university data repository]. So I emailed our data librarian, and our data librarian was like, "Yeah, no, you're not going to be able to upload this to the web interface. But, you know, let's, let's work with you." And our deal, everyone was really, really great in

terms of, you know, at at [large university data repository], like opening up a back door to like, get us in to upload these, like hundred gigabytes of files of the 100 gigabyte file. And so that was, that was a little thing that I was like, "Oh, yeah, like, maybe I actually should've started with y'all at the beginning." So you know, so but we've got it there. And we have everything, the other legs Zenodo and Figshare. dumps have everything but that first one, and we actually redesigned our data analysis pipeline so that it pulls from [large university data repository] to get those if you if you run everything from from scratch at the beginning it no longer because Wikipedia has downloaded deleted those dump. And so now it pulls from [large university data repository].

Sara Mannheimer 51:23
Nice, okay, do you are you using like a transfer? Like a Globus sort of? Or do you just go little by little with like an FTP? Do you know?

BSR02 51:31
No. No, this... No, it wasn't it was it was I think I just SSH'd a staging server and did it that way.

Sara Mannheimer 51:35
I guess 93 gigs isn't that huge but...

BSR02 51:44
Yeah, it's not like massive, but it was just like too big for the web interface. And so and then we were just using, we actually have, there is some there is like a cloud services kind of offering that Wikimedia has they use internally for all their stuff, and they make available to researchers. But we were we were going beyond the bounds of what that was useful for in terms of the storage. And then I actually ran a lot of these just on my I have like a beefy desktop. And so it was just running them locally for that and made sure. And then [my collaborator] also has like a beefy desktop. And so we were so we were like both running those and just synchronizing with get, and that worked.

Sara Mannheimer 52:27
Okay, cool. Awesome. Oh, this has been so delightful. Thanks for talking to me.