BSR03_transcript_deidentified

SUMMARY KEYWORDS

data, users, book, author, data sets, records, link, recommendation, research, gender, published, context, ratings, public, people, set, scraping, project, integration

SPEAKERS

Sara Mannheimer, BSR03

Sara Mannheimer 00:00

Lest I forget, okay, there we go. So, first, can you tell me a bit more about yourself the type of research you do and the type of data you produce.

BSR03 00:15

I am working on, so, recommender systems researcher is my primary background, with a human computer interaction focus to my training. But I do a lot of things particularly looking at issues of bias and discrimination, and search systems, recommender systems etc. I also do some other work on just making more effective recommendation research experiences. But with that we need a lot of data for.... a recommender system is a personalized machine learning algorithm, we need data to train on to learn people's preferences. The recommender systems community has a long history of public datasets. Um, going back to well, originally, one of the earliest ones was the each movie data set that came from the movie rating and recommendation system operated by DEC in the 90s. And they made their data available to the research community. When they shut down, they transferred the data from their system to the GroupLens research lab at the University of Minnesota who used it to seed the MovieLens platform, which then produces the data sets that are probably used more than any other single data set for recommender systems. So it's that's a community that has—the current data set dumps have data from somewhere between 150 and 200,000 users who have used this thing over the years that they know they're participating in academic research. Minnesota provides dumps of the database for the recommender research community. Those kinds of data sets that need a lot of the core recommender systems work is needing data that we have, basically, we need to know what items users like whether it's movies, whether it's books, whatever. So I've done a lot of that with with doing that with movie data, particularly the movie lens data set. But then also working on research paper recommendation, which I mine Bibliography linking or link to bibliographies from the ACM Digital Library, they were giving me in-beta data dumps for a while. Um, and I've been doing some work on books where that data is, again, we need the data of what users have liked. And so there's an old data set that's been around for 15 years or so from a website called BookCrossing. That predated Goodreads. I don't know how popular it was, I think it still exists. Um, but also, other groups have scraped and made available for academic research data from, from Amazon reviews and from Goodreads. And so those are the data sets I largely use. For the books and the kinds of data I use, there's the one of the tricky things is we don't actually care about the user identity. So it can be they can be deidentified, and I'm not going to go try to re identify them. There is research on redefining data users from these data sets. It's remarkably effective. But um, I just need to know that there is a user number 17. They liked these books. And so I can use that to train my recommendation models and and look at user behavior and look at the recommendation results and effectiveness.

Sara Mannheimer 04:03
Cool, very interesting. Nice. Okay, so some of that is like social media data, others is established datasets that and it seems like you don't do any scraping or downloading yourself.

BSR03 04:20
I haven't done any scraping myself in a long time. In grad school, I was doing scraping a little bit, but I had Yeah, I haven't done that in 15 years.

Sara Mannheimer 04:33
All right. Um, so I think I sent you a follow up email asking you to identify a specific example of a recent time when you collected big social data or used social data to do research. The idea here is we want I want to be specific as we work through these issues. So having one project in mind is helpful. We can also sort of like, move around between projects, if there are some that apply more to the questions that we have in mind. But do you have a recent project that we might use as this specific example?

BSR03 05:16
So probably the book data. So I've been with the, with some of my book research, particularly, I've been wanting to study. But generally, I've been wanting to study bias and discrimination and recommendation, particularly with regards to content provider attributes. So is the search engine biased towards or against content creators of particular ethnicities or genders? Um, getting that data is difficult. And I'm, and so I settled on books for two reasons. One is that, well, maybe one of the two reasons but books have, unlike, say, movies, books have an identifiable primary creator that we can think about, like there is and there might be two or three authors, but there's an author for a book. And while we probably want to be fair to everyone else involved in the book production process, too, you don't have the like movies, you want to be fair to directors, producers, actors, there's 17 of them. authors, authors are better defined than what we would want to do with say, movies. And then coupled with that the availability of rich data about books and authors. So that we can actually be able to identify authors, discrimination, relevant characteristics in many cases.

Sara Mannheimer 06:56
Okay, so it's whether what is recommended is biased.

BSR03 07:02
Yeah. I've also done some work on bias towards different types of users. But that's a much smaller off the shelf data project. This one... So this one, in order to answer the question, I tied together six different data sources. Um, I have book consumption records, BookCrossing, the Amazon reviews and Goodreads. And those were all scraped by others. Um, the Amazon and Goodreads data both came from Julian McCauley's group at University of California, San Diego. They have, they have a small cottage industry of scraping data sets and making them available. Um, and so that's where I'm getting the book consumption data to be able to train recommendation models and evaluate their effectiveness. But then I need book and author metadata. And so the Goodreads data comes with some book data. But the BookCrossing and Amazon data just comes with ISBN or ASIN. And so I have to go get that data from elsewhere. So I pulled in the, to get book metadata, I pulled in the Library of Congress, open MARC distribution records. And I pulled in Open Library. And then I linked... that has the book information, including who the authors are, but has no author information itself. Like, right, Open Library theoretically does, but it's super, super sparse. Um, so I linked that with the virtual internet authority file to get author information. And that let me build a link of user rated book by author of gender.

Sara Mannheimer 09:00
Nice, okay.

BSR03 09:01
And then I could go do my study.

Sara Mannheimer 09:04
Okay. Cool. That'll be interesting when we talk about comparing data sets. Comparing and combining. Alright, so you've described your data collection method. Is this part of a grant funded project where you were required to treat the data in a certain way or if you had a data management plan?

BSR03 09:23
Yes, the initial data collection predated the grant because it was used to drive preliminary results, but we're continuing to refine it through the grant process. And the data management plan, I specify that I'm going to share the data I can, okay, but note that some data is not going to be shareable either due to upstream restrictions—several of the data sets, I'm linking, I'm not allowed to redistribute. But almost anyone can go get the copy themselves, but I can't provide it. That's like the Amazon and Goodreads data is non redistributable. Um, I make sure I have a backup copy myself in case it ever gets taken down. But it's not redistributable or in other cases that haven't been triggered in this process, but my DMP stipulates that I can also not publish for human subjects protection or other ethical concerns.

Sara Mannheimer 10:19
Okay.

BSR03 10:19
So. But it also says that I have committed to making available the relevant code for reproducing my work, and where it comes from datasets that are public, but I can't redistribute. Basically, I'll give people the tools to go do what I did. And so all of the code for this data integration is open source. Anybody else that can get a copy can and has a half a terabyte of Postgres [PostgreSQL] sitting around can go to it.

Sara Mannheimer 10:46
Okay. Who is your funder?

BSR03 10:49
[U.S. federal funding agency].

Sara Mannheimer 10:50
Okay, cool. All right. And so have you published any data so far? Or will that all happen when you're done with the project?

BSR03 11:00
I haven't published any data, but we do have the codes available.

Sara Mannheimer 11:16
Cool. Ah, all right, sweet. All right, let's move on to our core six questions. So the first one is about context. So I have a little blurb, I'm going to read you to help you understand how I'm thinking about context. "When we collect data from social media platforms, just as when we collected data in traditional spaces, context matters. But the context of a social media post may be absent or difficult to

understand. These posts are by nature, short pieces of text taken from the larger context of personal and public life. And then this out of context effect is compounded when data are masked at a large scale." So that's kind of the issue that we're looking at here around context. Can you tell me about a time, if any, during the process of this research that you're doing when you considered the issue of maintaining and understanding this data's context. So like information about the community where the data was collected, or any contextual information about respondents or platforms.

BSR03 12:28
So we haven't done a lot with context in this data set, the one thing we do do is that we don't mix and match consumption records. So we treat each of the consumption data sources as a separate rating data set, and we don't pull them together. So we'll train a recommendation model on the Amazon review ratings, and use them to do recommendations for Amazon users will train a separate model on the Goodreads data and recommend for Goodreads users. So we're not treating user histories as fungible between sites. Um, that's about the only thing we're explicitly doing with context. Um, as I think about it, and context isn't something I have thought a lot about in the project. But our use of the data is very similar to contexts in which it is already being used, even if it's not the context in which the user intended it. So when you go to Goodreads, and you add books to your shelf, that's the records we get. We have that users added books to shelves, that they rate the book four stars, and we have timestamps associated with those. Um, that's the input Goodreads is using for their recommendation models. And so while the context the user is doing it in order to log their reading, or in order to communicate to their friends, what they're reading, in terms of the technology, like we're doing, with probably a different set of algorithms, exactly what Goodreads is doing to generate your set of recommendations. And so where we're effectively using a very similar context to try to understand how algorithms behave in that context in a way that I think is pretty faithful to the context of what's happening with, with the data in its original situation. It's at least on Goodreads, it's less clear on Amazon, because I don't know that Amazon is actually using ratings for any of the recommendations. We don't have users' purchase histories for very good reasons.

Sara Mannheimer 14:40
Right. Yeah, I mean, that's part of what I am curious about too, is like, I guess since everything is de-identified in your situation, you don't have any knowledge of like, you know, this person, you know, from this region of the United States. These people tend to like these types of books or, you know... users who identified this as their interests in their bios tend to like these types of books, or anything like that.

BSR03 15:07
Yeah, if I wanted to go do some scraping of my own, I could get that. But because one of the things I also have that we haven't made much use of yet we haven't published anything on yet. We also have, but for the Amazon data and the Goodreads data, we have review text. And so one of my students is working on review based recommendation and making use of that review text. And they're, like, since it's review text, if I really wanted to figure out who a user is I and I could go figure out who wrote that review, because it's on, it's like, they just scraped Goodreads public reviews. Yeah, but I'm not going to go do that. Because re-identifying users is not the business for it.

Sara Mannheimer 15:53
Yeah, um, but is that partially like, we have a question about privacy that we can go to... is that because you're concerned about privacy, or because you're concerned about liability, or just because that seems out of scope for you?

BSR03 16:10

Um, partially out of scope, partially privacy, partially that as soon as I do that, I'm probably in personal information territory in terms of IRB oversight. And not that I've tried to skate around that in any way. But there's like, we're paying attention to the ethics of this work, but it, it changes, it changes what the nature of oversight document may change with the nature of oversight documentation would look like right now we have, we have to go ahead. But part of that's what the understanding of we don't have user identifiers.

Sara Mannheimer 16:52
Mm hmm. Okay. Let's see. So I guess, so you just haven't thought too much about context, thinking that you're sort of using the data in the context in which it's supposed to be used almost. Okay. That that's helpful.

BSR03 17:21
At least it's a similar context in which it is being used whether or not users think it's supposed to be used that way...

Sara Mannheimer 17:31
Yeah. Okay.

BSR03 17:32
Different people have different expectations.

Sara Mannheimer 17:34
Right, okay.

BSR03 17:36
But we're doing the same kind of thing with the data that Goodreads recommendation engineers will be doing except we don't have all of the data they have.

Sara Mannheimer 17:43
Okay. Cool. Okay, let's move to data quality. Can you tell me about a time if any, during your example, when you consider the issue of data quality, so like, maybe like an incomplete population that you're drawing from or missing data of some type, or bots or bias? I know you're looking at that. But kind of thinking not like how you sort of dealt with data quality issues, and then also how you communicate any issues that you encounter to future people who might use your code or might try and reproduce your research.

BSR03 18:23
So there are two sources of data quality issues with one of which we spend time thinking about. The first is the just the quality of the underlying source data. Which I don't spend too much time thinking about, at least on the rating data side, because, like for bots... Amazon and Goodreads are... their quality teams are working on kicking bots off the platform. They're going to do a far better job than I ever could. They have access to signals that aren't in the public data set that are what you actually need to detect bots. Like there's work on detecting bots in rating data, but the rating data is missing all the signals like IP addresses, and timestamps and login activity and those things that like if I were actually trying to detect a bot, I'd look for who's publishing who's logging in from 15 IP addresses in one night, not who has an anomalous statistical pattern in what books they like. So I just "out of scope" that, I don't worry about it.

Sara Mannheimer 19:31
Yeah.

BSR03 19:32
Um, where I do start paying attention to more quality is on the metadata side. But particularly, we can also have quality issues with the data linking. And while we have—ostensibly have—a linking identifier, between the book rating record, or I mean the book consumption record, and the book metadata records to the ISBN, sometimes the ASIN. ISBNs are a holy mess. Um, we also have the issue of we want to do our recommendations at the work level instead of the individual edition level, just to be more realistic, and also to not have the data be near as sparse. So we have to link things together into works. We have a relatively decent algorithm for doing that with some very large known flaws we have to fix. Um, about 10% of the Goodreads ratings are for this super-cluster work that contains Tom Sawyer, Treasure Island, Little Women, one other famous children's book, and then like a dozen other lesser, well known books.

Sara Mannheimer 20:50
Oh, wow.

BSR03 20:53
But the reason for that is the reason for that is relatively straightforward. There was a joint edition of all four of those books published with a single ISBN. And so we're using ISBN, and book records and a clustering algorithm to identify what's probably the same work. And we don't yet have—other than manually playing whack a mole with known bad ISBNs—we don't have a good way to go and break those clusters that are being inappropriately linked. That's one known data quality problem. We just say it's there and hand wave it.

Sara Mannheimer 21:27
And when you say it's there, is that like in the place where your data is available? in your manuscript, in both?

BSR03 21:32
I think we document it on the data page. I do not remember, I don't think we... do we say? I guess I don't specifically state the extent of the problem on that page, I just say there are some known problems that can cause books to be inappropriately linked. Um, we try to be transparent about that; we haven't clearly documented the like—"Whoa, there's this one mega cluster."

Sara Mannheimer 22:03
Yeah. Would you do that? Have you published on this yet? Like, can you do that in your methods, or like...

BSR03 22:10
What we do... So we roll it into some other statistics, because in our data integration... So we do this linking, to link a rating to its cluster. And then we try to identify the author gender for a cluster. And the way we do that is we take the first author from every book record associated with the cluster, we take them now there's no linking identifiers. In the book records in the virtual internet authority file, we have to do name matching. And so we but we go with the premise that it is very unlikely for two authors to have the same full name and different genders. And so it may happen, but we treat it as very unlikely. So we take the first author from every book in the cluster, we take every virtual internet authority file, record, in our, um, we take every that matches that one of one of those names, we take their gender, if

they have one, and then we do a collapse. So that we basically, it becomes since virtual enter an authority file strips out non binary gender identities, a problem we document, working on fixing it, that's another quality issue we pay attention to is like, is it doesn't have gender, right? Um, in all cases, but we've been we've resulted in is is an as a cluster is a clusters, author, gender is male, female, or ambiguous or unknown. And so what that manifests, we also have like, what do we do if the user rates multiple has multiple ratings for the same cluster? That doesn't happen very often. An acceptably small fra- and we document the fraction acceptably small fraction of users have multiple ratings for the same cluster. Um, and then we have, so this manifests as okay, users are more likely to have multiple ratings for it, and users are, and it also has an ambiguous gender.

Sara Mannheimer 24:27
Yeah.

BSR03 24:28
Um, and so we report statistics on ambiguous author gender and on multiple ratings for the same author for the for the same cluster. This is just the biggest one of that problem.

Sara Mannheimer 24:42
Yeah. So I guess part of what you're doing to support this quality issue is just having more data. So then it becomes an exceptionally small fraction, right? Is that what you're saying?

BSR03 24:55
Partially and then also, like, I want to fix that supercluster linking problem I just haven't found the solution. And yet, some of it so being and we're gonna keep updating the data set codes that people can, can go and, and try to. So people can get the current version, I don't know if I have a full log of known problems anywhere, but I also have mental list of like, okay, here are the open problems that I need to try to fix with this data. So the next project uses a better version of it.

Sara Mannheimer 25:27
Nice.

BSR03 25:29
And we try to we try to approach it, we approach it and we're explicit about this in the paper to learn what we can from the data we have while being transparent and clear about its limitations.

Sara Mannheimer 25:42
Yeah. Okay. Great. Cool. Let's move to number three data comparability. So during this project, did you compare and combine and or combine multiple big social data sets? And you say, yes, yes. And so why did you combine them? How did they advance your research? And then what strategies and challenges did you encounter and how did you address address those?

BSR03 26:13
So we have described a chunk of the, of the combination process, and it's described in more detail both on the documentation page and on. And in our journal paper, that was published in user modeling and user adaptive interaction earlier this year. Um, but we, the reason we needed to was two reasons. Um, so we needed to combine multiple data sources, because no one data source had all of the data we needed. We had ratings, but we didn't know who wrote the book. We know who wrote the book, but we didn't know anything about them as a person. And so we needed to link different things together in order to be able to actually answer that question. Also, then we have these three different sources of of

user book interaction data. And that's partially because the recommender systems research community, likes seeing results on multiple data sets.

Sara Mannheimer 27:13
Okay.

BSR03 27:15
Um, but also, it gives us the BookCrossing data set and the Amazon data sets have been used in a lot of recommendation projects like yours, you'll see lots of papers using both of them. BookCrossing, a little less these days, but a lot of older book recommendation work used it. And so that gives us this historical continuity and comparability. But then Goodreads is by far the largest book recommend data set station data set I know of, it also gives us in addition to the rating, so Amazon, we only have user ratings, Goodreads gives us add to shelf actions. So anytime we have a log of users adding books to their public shelves, and so that lets us detect expression like that. They added it to a shelf with some kind of expression of interest. They read it, lets us get that interaction data, even if the user hasn't gone to express a writing. Okay, and most production recommendation systems are trained on interaction data, not on reading data. And so it lets us that one was we have historical continuity with the first two data sets, and then Goodreads lets us get a lot closer to actual practice.

Sara Mannheimer 28:32
And then what challenges I guess you've talked about some of them with, like imperfect metadata, and also imperfect data in each of the data sets, but what other challenges did you encounter? And then what strategies did you use to address them when you were working with all these datasets together?

BSR03 28:52
Um, so some of it was one of them was just document well, efficiently integrating them and documenting that integration. So when we go to write it up, it's easy to tell what the heck we actually did.

Sara Mannheimer 29:04
Yeah.

BSR03 29:05
And so the way we approached that, in the version that's currently public, is we did all of the data integration in PostgreSQL. And so we imported the data set in as close to raw form as possible, um, directly in the Postgres tables, and then our different cleaning and integration, extraction, cleaning and integration steps, with a few exceptions were written as SQL queries. And so if I needed to go write down exactly how we did the author name matching, I can just go look at one SQL query and see the exact code that I'm, I'm currently working on retooling that a bit to not depend on Postgres. But the core concept is relatively the same that each discrete step has is a place I can go and see what I did there. In order to have a clear documentation on how the data was linked together.

Sara Mannheimer 30:08
Yeah.

BSR03 30:08
We also go ahead.

Sara Mannheimer 30:10

That's included in the public data that you provide as well, so that anyone can go match it. Yeah. Okay, cool.

BSR03 30:18
Um, we also have a, I had some challenges with just the size of the data set my recommendation software, some of the algorithms choked on the size of some of the data. So some of it, we just weren't able to use some of it, we I went to improve the software performance, we could actually run on the Goodreads data. Um, those were a lot of the big issues. We did see, I believe gotten the paper, we've got some charts that are, say, comparing gender distributions between different data sets. Um, and there were there were differences, say in the proportion of of authors of books by female authors rated in the different data sets, not huge differences, but but the Goodreads data was more user engagement book presents was more equitably distributed than the Amazon review data, for example. The Amazon review data is also older, we're working with the 2014 data set at the time. Okay, but we saw some differences there. But that's about the extent or you we compare it on a basic statistical level, how big they were, what the sparsity was. But beyond that we didn't. We didn't do a lot, particularly we haven't done any qualitative comparison of what the difference is in these data sets, these data are.

Sara Mannheimer 31:45
Okay. All right. Awesome. So let's move on to informed consent. Was there a time during the process this research when you considered informed consent for the users themselves?

BSR03 32:04
So other than the nagging thought, in the back of my mind that we kind of hand waved that? Not a, it has not been an explicit thing, so we're only working with public records. And there there is a I mean, there is a contextual integrity thing here of that, like the when the user submitted the right review to Goodreads, using it in my research wasn't their intention. Um, but there is a but we have we are working entirely with public records. Um, this is standard practice for recommender systems research. Um, there's, there's good arguments that perhaps it shouldn't be, but it is standard practice. Um, that's about the extent of the thinking there. We thought about other of the ethical issues like we have it this paper or another paper we have that ethical note in.... I'm just searching here to see if it was this one. Okay, it's not this one. I put a note, in another paper that was actually working with user demographic data, that that because the data sets are public, like the work we did does not expose users to any risks to which they were not already exposed. So we're, we're trying to be careful that we're not exposing users to new risks, but, or authors. Right. But um, the informed we have not like, yeah, informed consent from the users participate, or from the users whose data we're using.

Sara Mannheimer 34:13
Yeah, and this fine answer too, it's like, you're saying that in your case, it wasn't something that you considered that you needed? I think that's fine as well, because that's part of what's so complicated here is like the line of like, when it is human subjects data, and when it's not, is blurry. So I think this answer is totally helpful, too.

BSR03 34:38
And the recent relevant research communities are having, having discussions about like, what should we actually be doing with this data or not doing with it and there are problems with oh, it's, it's all public, so I just used it. Um, I think that our use is significantly less ethically fraught than some of those. It's just public. So I just used it projects. Right? Um, but there's a there's the need to pay ongoing attention to that, for that discussion.

Sara Mannheimer 35:13
And I guess I do I like who, you know, when you're paying attention to this clearly you haven't aren't knowledgeable about it, but like, who do you read? Or like, what sources? do you go to? What if you were to have like a conundrum? You know, if it was, if you didn't feel it was as straightforward as this research, like, what would be your sort of strategy there?

BSR03 35:33
So I people I've read... Casey Fiesler, and then I've also I kind of keep a loose eye on the general work of the PERVADE team. Um, Anna Lauren Hoffman.... who I would go to... I may go to one of them, or just generally, my friends, particularly my colleagues in the fairness, accountability, and transparency community. Several there I could, would ask, if I or and also having a dialogue with our local IRB folks. And so, um, why while this particular project has been, they've they've judged it not human subjects research, because no personally identifiable data, and we're not engaging with users. Yeah, um, we've talked with them about other projects about this and related projects, and the ethical dimensions have been beyond just the human subjects concern. And they've been a useful resource. So I also talked with them.

Sara Mannheimer 36:38
Cool. Okay, that's helpful. All right, let's go to number five out of six, privacy and confidentiality. So we kind of talked about this a little like, I guess it was there a time when you consider the issue of privacy? I guess, since it's all de-identified, I think you're or maybe if not, in this example, in other examples that you've worked on. So I'm kind of like trying to get into like times when you had issues and that you thought about.

BSR03 37:11
So this data, like there's the de-identifiability portion, but I think more importantly, there's the it's all public portion.

Sara Mannheimer 37:20
Yeah.

BSR03 37:21
Because if someone wanted to use this data, to one of its subject's detriment. It's not anything I'm doing that's going to enable that.

Sara Mannheimer 37:34
Right. Yep. Okay.

BSR03 37:36
Um, and so there's not like, what I'm doing doesn't create any new privacy related harms may create other harms, in some cases, but it's not creating any new privacy related harms for anyone. Because the data so out there, yeah, and I'm not redistributing any of it. But, um, there is a so we've been doing some other if we if we're in discussions that we haven't acted on yet, of trying to move beyond data, or beyond book authors who are relatively public individuals in terms of you get an authority record. And we're using the authority records to say academic scholars or others who don't have as large profile. Um, how do we get the data? And what do we do once we have it? And even if it's even if it's assembled from public sources, say we have if we have [Mechanical] Turkers, getting data from scholars' public profiles, which we have not done, and I don't know that we're actually going to, but if you assemble it in one place, it makes it easier to find. And so we would not make that data publicly

available. sources are public, anyone can go do what we did. But nope, not gonna make that publicly available. Um, we're also looking at TREC, the Text REtrieval Conference, and their fair ranking track and privacy concerns and putting data in one place concerns along with identity just the the concerns of people's ability to control their presentation of their own identities, is one of the reasons why we didn't use something like gender for the Academic Search data sets that we put together for that. Um, instead, we used the economic divide level of the country in which a researcher was operating when they wrote a paper as a crude proxy for the level of resources they likely had available for research. The idea being that we want to be fair towards schol-, we want to make sure that scholars who are operating and under resourced environments have a fair opportunity to have their research work surfaced in search results. And that's something that's derivable from public information, the affiliation country in the paper you published because it was all published papers. And then combined with public records like the International Monetary fund's economic classifications, or there's so that we tried to go with something that is completely derivable from public data and does not involve any sensitive personal attributes. So we could catalyze this kind of research without creating new privacy or discrimination problems through making a an archival data set available.

Sara Mannheimer 39:19
Right. Oh, the so that, that's very interesting. So you, like actually designed your research in a way to like, after thinking about privacy issues, like, okay. Very interesting. Okay. Cool. All right, let's move to intellectual property. Was there a time during your research when you considered intellectual property concerns, either by from the platform or the participants users themselves? Did you consult with anything, anyone about the concerns? What strategies did you use to sort of figure it out and address any challenges that you encountered?

BSR03 41:54
Um, I haven't, I haven't spent significant time or effort on that, um, that like, from a lot from an intellectual property liability perspective, the people who scraped and initially produced the data would be on the hook. Yeah, that's one reason I'm not redistributing the data. Um, and on its usability, like the data sets are very well known and are still available.

Sara Mannheimer 42:22
Okay, yeah. So you're kind of assuming since they've been public for so long, there aren't any big issues? Yeah.

BSR03 42:29
Or at least I mean, and other datasets have been taken down. So if Goodreads or Amazon wanted him to start distributing these data sets they would have asked him to? And so...

Sara Mannheimer 42:42
Did you... So in other projects, has this come up for you as an issue, like looking at Terms of Service, or you know, trying to figure out any issues you might encounter?

BSR03 42:52
It's one of the reasons I've been hesitant to do a bunch of scraping myself is just to avoid that set of issues. Outsourced it, I've done scraping of sites that have explicit open access terms. But, um, but I mostly outsource that.

Sara Mannheimer 43:12

Okay. All right. Great, well, are there any additional issues or challenges that arose during your example that I haven't asked you about?

BSR03 43:24
Um, so one of the, I brought it up briefly, one of the complex ones, like especially if we're dealing with author gender is what is that and how was it recorded, it is recorded accurately.

Sara Mannheimer 43:38
Yeah.

BSR03 43:39
And we like the when you look at the market authority file specification for the gender field (375, I think), um, it specifies that it is gender identity, specifically, it specifies that it is, um, it can come from an open vocabulary. And there can be multiple gender identity records with starting in validity dates. So like the file format itself, has this very rich notion of how it can record an author's gender identity, and it's specific that we're talking about gender identity. Um, you throw all of that into the woodchipper. A virtual internet authority file, and the nuance disappears. Um, but so we paid attention to what the data is and what it means where it comes from, to the extent that we can. And then where we know our current integration has significant problems in this front. We're looking at how to get better author demographic data particularly we can get it from the Library of Congress name authority file, instead of going through VIIF, their data is higher quality. The coverage is at least in our initial prototype integration, that we're still debugging, the coverage is very, very poor. So thinking through that set of issues also, but and also being being encouraged by the thought that was put in, in both in the mark record design, and then also the program for cooperative cataloging things working group report on Recommended Practices for recording author, identity, author, gender identity, in a name, authority record, um, is very, very thoughtful. And I believe anybody feeding into the Library of Congress files, probably following that. But VIIF sorts sources for many national libraries, many of which GCC is the thing. And so they have their own standards for how the how the the records are created and maintained. And so like, seeing the dealing with a very nuanced and complex subject, seeing pieces of the data ecosystem, handling it in ways that seemed relatively good. And then that gets fed through a data integration that just kind of chops it up.

Sara Mannheimer 46:29
Yeah.

BSR03 46:30
Um, the fact that we are looking only at statistical properties, and not looking at individual author identities, except for debugging purposes, means that the manifestations of the problem are different than they would be if we were, say, publishing a list of here's the genders of these authors. Um, but we still have... one of the things that's unknown, and I would like to get a handle on, is the extent to which accuracy and presence of these labels is label-dependent. How does the distribution of the data we don't have or the data that's wrong, compared to the distribution of what we do have? Because if authors of particular gender identities are systematically more likely to have their identities erroneously recorded, which I think is almost certainly true, that then does skew the statistical results. Open question is, what's the magnitude of those skews? I have a rough sense for what the direction probably is, but what's the magnitude? Um, we also see, and sometimes there's evidence in the data that the direction might not always be what we expect, like I've looked at the proportion of books written—the male to female ratio, in Library of Congress over time by publication date since about 1960-ish. And it's flat, which surprised me.

Sara Mannheimer 48:20
So the number of women authors has not gone up?

BSR03 48:24
The fraction.

Sara Mannheimer 48:26
Yeah, the ratio. Hmm. Interesting.

BSR03 48:29
So that made me wonder, has the fraction really not gone up? Or is there a bias in which authors people pay attention to creating a complete and accurate authority record for? That then skews our perception of the space? And I don't know the answer to that.

Sara Mannheimer 48:52
Oh, there's so many factors. Wow, very interesting.

BSR03 48:58
I would like an like an answer to that. But I don't have an answer to that. There's a lot of open questions in the data. But those are some of the things that we're thinking about too, is like, how do we, how do we deal with this? How do we deal with a sensitive and very personal demographic attribute in a way that's respectful? And we try to pay attention to that. Not going to claim that we do it perfectly. But we're trying to pay attention to that and do the best we can with the data we have.

Sara Mannheimer 49:29
Yeah. Okay, I think that's all that I have. I do have one favor to ask is, I'll send you a follow up email after this. Just saying thanks. And I wonder if you could, if you have anyone in mind, who also does big social research, or qualitative research, if you know, people who do that, or who publish their qualitative data. I'm looking for additional people to interview, who might be interested in this and be willing to spend an hour talking to me, so. Okay. Thank you so much. This was really great. It was so interesting to hear about all of these challenges. And I think it'll be really helpful to my data set. So thanks again. All right, have a great day.

BSR03 50:24
You too.

Sara Mannheimer 50:25
Okay, bye.

BSR03 50:26
Bye.