Event: Evaluating Data Quality - Challenges & Competencies: Session 1
Speakers: Bobray Bordelon, Ron Nakao, Barbara Esty

Transcription

Bobray Bordelon: Hello everyone, we're going to give just a minute or two to let everyone in and then we're going to begin. All right, welcome everyone.

Bobray Bordelon: Welcome everyone to Building Capacity of Academic Librarians in Understanding Quantitative Data, Data Quality Problems and Evaluating Data Quality: a National Forum. This project was made possible by an IMLS grant and the slides will be made available, and you will be able to see the wider scope of the project.

I just want to very briefly mention this is a large team behind this. Leading us is Grace Liu, who is the Principal Investigator. She is at Westchester. Rashelle Nagar, who is at Stanford,
Bobray Bordelon: myself at Princeton, Marydee Ojala, who is the editor in chief of Online Searcher, Dr. Jodi Schneider, from UIUC, and we have 2 graduate assistants helping us out, Jordan Sarti and Uyen Nguyen both from UIUC.

So just a few housekeeping things. Yes, questions will be answered in the last 25 minutes, so unless something is really, really urgent where you're just completely lost, we are going to save questions.The chat has been disabled but there is the Q&A Feature which is in the Zoom toolbar, so please submit your questions using it.

Yes, the slides and the recordings will be made available at a future date.Closed captioning is available from your toolbar at the bottom and in addition, on the website, there is some supplemental materials, including a glossary of various terms, as well as key terms that one would find in documentation. So we begin today.

This is the first of a seven-part series.And today, we're going to focus specifically on documentation. It is where one normally begins, particularly for most of us, I believe, or librarians or people who are helping people find data, and it is where we typically start. The website is mentioned at the bottom, and it will also be listed in the chat, so that you can see more about the series.I want to very briefly tell you about what's to come. So in a little while, Ron and Barbara are going to talk about documentation.

Every month on the last Thursday, we're going to have a session at this exact same time, and so the next one up I'll just tell you a little bit more about than the others, since it's next in the series. And we're going to look at quality assurance and data creation.So specifically, we're going to have questions such as, how do governmental bodies and academic researchers ensure the data they produce and work with has the quality needed for meaningful results? Can we prepare students in the public to understand the data and statistics? How does the decentralized government produce the data we need? How do you take advantage of data coming from different sectors? And we have an incredible lineup.

We have the Associate Director for Research and Methodology and Chief Scientist at the United States Census Bureau, John Abowd. We have a professor at Princeton University who is a health economist, Janet Curry, and then we have the former chief statistician of the United States, Katherine Wallman.

In Session 3, we're going to look at evaluating and understanding governmental data, and we're going to look at it from both a United States as well as an international point of view. Session 4, we're going to look at commercial data quality issues. We have a number of stars lined up for every one of our sessions, but in this one we're really going to look at how both in academia as well as the government, how do you assess? Session 5, we're going to take a sort of different route, and we're going to look at all right, now we have data. Can it be reproducible?Is it actually being preserved? So in this case we're going to have someone from the American Economics Association, someone from ICPSR. So we're gonna again have different people that are talking about what's out there, as well as a person who's done a study that shows the state of preservation. Session 6, we're going to see what do employers expect from the people that they're hiring? And we have people from both the nonprofit as well as the commercial sector. And then we're going to wrap up with what is the librarian's role in cultivating data-literate citizens? And so we're going to have people looking at survey information as well as their own practices. And in all of these sessions we hope to have skills that you can take away with you.

So, very quickly What are we going to talk about today as well as in the series as a whole? We're going to largely be talking about survey data. It can be quantitative, it can be qualitative, but we're largely going to focus on quantitative, essentially numerical data. It could be at the micro-level, which is individual. It could be a person, could be a company, could be a hospital, or it could be macro-level summary. And we're going to talk about both of these during the series. How do we get data? Lots of different ways. Experimental data, often we think about these as trials, often medical, this we're probably not going to touch upon very much. Survey data is what we're probably what we're going to focus on the most throughout the series. It's the most common method in the social sciences. We're going to look at administrative data often collected by government or organizations. School records, things like that, hospital records.

In survey data itself, the 2 big types are cross-sectional, where we're looking at different individuals each time. But we can look at trends that represent populations. So just a few examples. How do we know distribution of illness in the United States. The National Health Interview Survey. The long-running Public Opinion Survey, which is a general social survey. How do we know how people spend money, the survey of consumer finance, longitudinal studies where you're studying the same person or the same group, whatever it may be, each time, sometimes called a panel Study. the advantage of this, is you can see over time How did this specific individual do certain things? So, for example, the national longitudinal surveys.

The panel study of income dynamics, the National Education longitudinal study. Another thing. Sometimes people don't get as they see a year, and they're turned off. And that's typically the start year. So it doesn't mean that it ended in 1988. It means the first group of people were surveyed in 1988. And of course there's public opinion surveys and there's many of these out there. We're going to actually have people from Roper speak to us. But there's lots of other organizations that do this. And so now I am going to turn it over to Ron Nakao, who is the politics, economics and andSocial Science Data Librarian at Stanford University.

Bobray Bordelon: Ron?

Ron Nakao: Thanks, Bobray. And I just wanted to take the opportunity to thank you and Rashelle and Grace, who we've been working, Barbara and I have been working with, for inviting us to participate in this really wonderful series you've put together. I'm really looking forward to actually going to the other sessions, and I hope that we're able to at least provide a grounding for all of you who are attending today. As the series opens up, next slide, please. So we, we, as data librarians, are those who are appointed to be the data point of contact in your institution. Often we'll get a question. I'm looking for this data but I usually take a step back. And I kind of fall back on what I call the data reference interview. Because that, the answers to those, the assessment you get from those kinds of questions will actually help inform you for the kind of data that the person really needs. And so the data reference interview includes things like who needs the data? What is their level of data or research experience? An undergraduate who is taking their first class and using data, has never used data before, is going to have a different kind of ability to work with data versus a faculty member or a researcher, seasoned researcher who has years and years of experience working with data, so that really influences the kinds of data that person will really need to serve their purposes. What data is needed?

This is where I refer back to trying to understand what their reference, their research plan is, what methodologies they're using, and i'll talk a little bit more about that in the next slide. But when is it needed? Scope and deadline of their research project? Again, a student who basically has a paper, or even a part of an assignment that's due in 2 weeks. The kind of data options you're going to give them is going to be very different than someone who is working on a 5 year research project. And then, one thing that a lot of people don't realize is, how long is it going to take to get the data? Often in this day of of Internet access the assumption is if I can think of it, it's immediately available. But often people are surprised to know that data is not always readily available, and so often it takes time to actually get the data that they need, next slide, please. Oh, next slide.

Thank you. So as I mentioned the research plan, and this is probably just, this gives context to your interactions with the person you're trying to help locate Data starts with understanding how they are stating their problem. What is their research question, and how will they continue to refine it? And then their literature review also helps inform the data that they're looking at, whether it's suggested by previous research or the limits of it are demonstrated, or how to use it.

The theoretical framework and methodology that the prospective researcher wants to use has a big impact on the kind of data they want. If they are going to try to create some models to just go inference kinds of things they often need sample or micro data. If they purely want to describe something like the breakdown of race and education in a State then that will inform a different kind of data that they would need finding Your data is kind of where people think that's what they start with us. But actually it's going to be important to actually understand the previous parts as it evolves of the person's research plan. And then, of course, they go on to analyze their data, discuss their results, and submit their paper, including data citations, I hope

Ron Nakao: Next slide. Yes, thank you. So, little bit about the literature view. This is where the researcher is learning from others because they want to build on the research of others. They want to see what the methodological options available for them to consider. They want to know what data to collect, and the person collected themselves, but also to inform the kind of data that the researcher may want to collect, or if they want to reuse data that was collected by some other institution or researcher. And I think the the thing that to keep in mind about the Literature Review. It is actually a very useful form of data documentation in and of itself, because in a sense, it captures how data has been interpreted and used by another researcher. So I think that's that's something to keep in mind when you talk with folks about the Literature Review They've done. Next slide.

So, taking a step back, this is my definition or operationalization of what I think Social science data is. The first point is social science data is the evidence required to answer your research question. This means that it's kind of, I think of it as in terms of like in a court of law, where you want to gather the evidence in whatever format it is in order to argue your case, the point You want to make, the hypotheses you want to test. Social science data is also the measures and constructions or operationalization of the concepts that are important for you when you actually are trying to answer a question, or you're trying to work with the theoretical framework that you're trying to apply or build your research on, it often comes in the form of concepts and not specific measures, and so often people immediately go to the measures. They often don't take a step back and say what are the concepts that I want the measures for? And also sometimes the measures often aren't exactly what they want, and so it's going to be important for people to ground themselves in the actual concepts that they're trying to bring as evidence to their research.

The third point is, it's dependent on methodology as I mentioned, and it's used to answer your research question. The fourth point, and this is something that is more the struggle of finding data. It's found in many places, sizes, shapes, and formats. You can get an Excel file. You can get a CSV file. You can get a Stata or an SPSS file. It comes in a lot of places, and often it doesn't, you have to do some manipulation in order to make use of the data. It's also costly to collect, curate, archive, and share, and this what makes it difficult often to get good quality data because, even though a person can easily create or share a data set to actually make it useful, it needs, I believe to be curated and documented in order for others to actually use it correctly. And I think the last point that people often don't realize: the data often isn't available. Next slide. So here are some important data attributes that I think it's important to find out more about what the person wants. Who or what is the subject of your research? That is, what is the unit of analysis or observation of your research? Is it students? Are you doing a study to try to find relationships between students in a school? Is it differences in counties? So counties are the unit of analysis. So you want to get gather concepts and measures at the county level. Is it cross-national research? So then, you want to focus on finding sources of data that have nations or countries as the unit of analysis.

Another important attribute is the when and where of the data. Often someone will come and say, I want this data, or I want these kinds of measures or variables. And then, of course, you always ask: So when do you need it? Do you need the most current? Do you need the last 50 years? Do you want every 5 years for the last 20 years? So the when part of it also, in a sense, makes the search of data more interesting and challenging. And the where: The where is what spatial or geographical components are necessary. Everybody nowadays want to do research at the smallest geographies. And so they don't just want the national data or statistics or variables. They actually want to go down to the block level. They want to get Congressional district kinds of variables or measures. And so that also adds another challenge to actually defining what kind of data the researcher really wants. And I think, as Bobray alluded to, there's a difference between cross-sectional versus longitudinal versus panel. Cross-sectional data is where it's at a point of time that the data points, the concepts, the variables are taken. Longitudinal goes across time and panel is like longitudinal, but it actually follows units of analysis, individuals or others, across time. So you can actually look for trends within individuals.

And then last, but not least: What are the concepts that need to be measured? And for social science data that's usually, what are the variables that are included? And bear in mind variables are an abstraction of the actual concepts. For example, if a survey is filled out by a researcher then there's a certain question that is asked of the respondent. But to create a variable the researcher or the curator, or whoever does some abstraction of the actual answers to actually create a variable. So again, it's important to understand that difference between and the process of data management that takes it from the actual variables that you can use to the actual questions that were originally asked. Next slide, please. Next slide, please.

So, here are some of the in my mind, some of the different kind of data types that you'll come across. There's a research versus administrative. Research data usually was collected for the purpose of doing research usually done by academics or institutions like the Census Bureau or others, that they know that their data that they collect is going to be used for research by researchers. Administrative data is data that is collected to do something that may not be directly research related. It could be a a business that collects employee information in order to actually run their business. It could be the Department of Motor Vehicles in California to basically be able to run the functions of that, of that State Department. And so, even though the intent of that data that was collected was not for research, it can actually be repurposed to answer a lot of very interesting questions. But bear in mind, because it wasn't originally collected for research, it often isn't really documented, or even constructed in a way that may be necessarily valid for doing research. Hence that, as people will say, administrative data tends to be very messy, and requires a lot more data cleaning and management to be useful.

Primary versus secondary is another differentiation you'll hear. Primary usually is, it was collected for the strict use of the researcher who was going to apply it to their own research. Secondary, basically is any data that was not collected by you or the individual, and they're trying to repurpose it to actually as evidence for their research aggregate, tabular or macrodata or statistics versus sample or microdata. The aggregate/tabular/macrodata, basically I usually think of it as the individual numbers or cells of a table of that data kind of stand on their own. If you see a table of a country GDP, then you'll see a cell that has a figure GDP for each country, and therefore each number stands on its own. Versus sample or micro data is usually it's a survey, a sample, hopefully a statistical sample of a population of interest. So you have individual responses, but it doesn't, but each response by each respondent does not stand on its own. Just because I answered a question a certain way that doesn't mean that my particular answer works. It's the collective of the sample data that can be applied to basically do the research. Quantitative or numeric versus qualitative or non-numeric, structured versus unstructured. Structured is usually what we use, but more and more in the social sciences unstructured would be scraping the web text, doing textual analysis, text mining, things like that. And then the last category, restricted/confidential/licensed/copyrighted versus public/public domain/open access/open source basically has something to do with any kinds of limits to the access or availability of the data. And then all of these types, of course, come in varied data and documentation quality. Next slide.

Ron Nakao: Types of documentation. So usually the stock for Social Science Survey data is you'll see a code book, you'll want to get a copy of the Survey questionnaire if there was one, the Survey design and sampling methodology so you know how to properly use that survey data. Research articles and reports using the data are also a very important source of documentation and other broader categories of what I call metadata is also very valuable when you want to do searches for data, because it allows you to have hooks to actually be able to come up with the right data appropriate for your result. But again, we find often there's little or none documentation available for data. Next slide, please.

So, this is sort of my take on indicators of data or documentation quality. Generally I believe it's better if it was produced for use by other researchers, it's better if it was transparent and well documented by its creator. It's better if it's validated via a peer review process. It's better if time and expertise are invested to curate it by experts or data curators. Next slide, please. So, sources of data documentation. This is where you will go and try to find not only data, but hopefully, the documentation to use the data. I usually ask myself the question: Who collects the data? Of course, there's individual researchers, international agencies, non-governmental agencies, government agencies at the national, regional or local level and of course, businesses, and then, again, who distributes the data? Libraries or data archives or repositories, commercial or independent organizations, or other researchers or research groups.

Next slide, please.So, this is sort of my kind of questions. The why, how, who, where, when and what when I'm kind of looking for good data documentation. Good documentation, to me, will tell you why the data was collected. The reasons the researcher use or anchored themselves when they were collecting the data. How? How was the data collected, structured and managed?
Ron Nakao: Who? Or what is the unit of analysis or observation? Where? Where are the geographies being covered? When? What is the time period covered? What? What are the concepts being measured? That is, the variables.

Next slide, please. So I want to go to an example of what I consider to be, to me in my experience, kind of the gold standard of providing good documentation for data. It's the Internet University Consortium for Political & Social Research, ICPSR. I'm guessing that most of you have had some intersection or use of that Social Science data archive. It was founded in 1962. I consider it the gold standard for social research data, curation and archiving and accessibility. In its archive it has over 18,000 well-curated dataset titles for social and behavioral research, and all the documentation and over 12,000 of the datasets are available to non-ICPSR members. Next slide, please. So this is sort of some of the pieces of the ICPSR documentation. One thing that's really great for ICPSR is that it has a standardized data page for all of their datasets, or they call them studies. The standardized data page includes a data summary, data citation, geographic coverage, time, study design, unit of observation and version control. It also includes a codebook or other documentations for the data. It provides variable-level descriptors.

It has to me a very innovative database that was begun by Elizabeth Moss at ICPSR, called the Bibliography of Data-related Literature. What she and her team have done since the 1990s is go and try to find citations of research articles that cite the data in their archive, and this provides a very valuable level, as I said, of documentation, but also insights into how the data has been used in their archive. And I'm gonna go with an example. It's a study that I was involved with at Stanford that's called How Couples Meet and Stay Together, HCMST. So when you go to that website, and you can do this at your leisure at some point in the future, this is the data page for that particular study. As you can see, it starts with a title, PI, the version date, etc. But if you go down a little bit there on the page, you'll actually find the "why" of this data set. This sort of gives the summary by the principal investigator of why they collected this data.

Next slide, please. And then, as you scroll down that page, you will then come to the point where you will find the "where" of that data set. It includes the geographic coverage of this HCMST data, which is the United States, but it also includes the smallest geographic unit available. As I said, often researchers just don't want national data. They want smaller geographies, breakdowns, geographic breakdowns of the data. And so this tells you that this particular data set has the ability to break down for the public use version of the data down to census region or if you get the restricted use data, you can go down to the State level.

Next slide, please.And then, if you scroll further down on the data page, you will see under the scope of the project the "when" of this dataset. It gives you the time periods of the dataset. It has basically 5 waves, beginning in 2009 to 2015, and it gives you some other information on the data collection. Next slide. And then going down to the methodology section, you find more the "how" and the "who" of this particular data set. So the "how" gives you the study design. So you you know who was the survey firm, what methods they use, the sample, there were 4,002 original survey respondents. It's a panel survey, so they basically did follow-ups from that original survey of 4,002. And then the universe, which was for the first wave, universe for wave 1 was English-literate adults in the U.S. So in a sense that also gives you some insights, because if you are looking for how couples meet and stay together, you'll know that this particular dataset was restricted to English-literate adults in the U.S. So it excludes a portion of the population in the United States. And then, finally, on this page the "who" of this dataset. What is the unit of observation? It has an individual level of data. Next slide. And then again, traveling up to the page, you'll see a tab for the data and documentation. And this is where you can actually download codebook files or other kinds of documentation that is available for the data.

Ron Nakao: Next slide. The next tab is the "what" of the data in more detail. These are the variables. And ICPSR has really done a service by basically parsing out the variables of a dataset. It gives you the variable names, you can see that there's 534 variables in this particular dataset. And it also gives you, it connects the descriptions of each variable and other information. Next slide. And then the data-related publications, which I mentioned. This is where you can actually, for this particular dataset you can actually scroll down and get a list of the articles who have cited this dataset in their research. And so, if it's available to your institution, you can actually click on it and get access to the article itself. Next slide. So that is kind of the ICPSR, which I believe is a really good example. I am just gonna mention another one that I really like for the way they've documented their data. It's done in a different way and I'm not going to go through screenshots of this one. But you can go and I encourage you to go to the IPUMS site in the future, and kind of work around, surf around that website to kind of get a sense of it. It's just an invaluable archive or a source of census or other data. It was originally created in 1991 by historians for historians at the University of Minnesota to encourage historians to use census microdata.

And what they've done is, they've harmonized the census data and its documentation. Because if you use direct census data, the raw data, often it's very confusing, because from census to census variables might change, different things might change. And what these these people at IPUMS have done was they harmonized the data across time to make it easier to use by historians who at that time were not known for using numeric data in their research. It was actually the founding PI was Steve Ruggles, who I think I read recently just received the MacArthur Genius Fellowship for his work in creating this very invaluable source of data. It includes U.S. Census, American Community Surveys (ACS), Current Population Surveys, Demographic and Health Surveys, and then International Census and other microdata. And again, it has excellent documentation with embedded, detailed links. Next slide. So you'll just see when you go through the site that the harmonized documentation for IPUMS U.S.A., which includes Census and ACS data are sample descriptions, the questionnaires, variable data dictionary, published census volumes and a revisions history.

Next slide. So to sum up what is good data documentation as I've listed, you want to know the why, the how, the who, the where, and the when in the data documentation for the data that you want to use. Next slide. And why is good data documentation so important? Good documentation empowers the prospective user to efficiently identify data for their research to use, reuse the data appropriately and correctly. Good data documentation leads to better science. Thank you. Next slide.

Bobray Bordelon: It keeps getting stuck for some reason. Alright, so next up is the Data Librarian from Yale University, Barbara Esty. So Ron told us about good documentation. What happens when it's not so good? So, Barbara, you're next.

Barbara Esty: So 80% of the time there is some level of harmony between the data, the documentation and the user. That's my rough guess. But you know that other 20% where there are questions is probably where you're going to spend 80% of your time trying to find answers, evaluate data, evaluate documentation. And so you know, these are the types of questions that come in every day. Users are getting data from many different places and they're coming to you with questions. And what's really interesting about working with people and their data is that they don't necessarily make distinctions between what subject you might know about. You know I've gotten questions from the same faculty member who asked me about, you know, coal prices and 2 weeks later asks me about an educational survey, and so they make no distinction. There's sort of some sort of magic that you know around data that people think just because it's data we can solve all problems. And sometimes we can, sometimes we can't.

Next slide. So I'm going to talk about a lot of, you know, documentation realities, and you know it's great when you have that marriage of good documentation and good data. But the reality is, there is no standard for documentation. And so you could get any kind of documentation for any kind of data and that can be very jarring to the researcher which you know generates, "Well, how do I use this? And what do I do with it," and it's not so easy to navigate. And also keeping in mind that you know documentation is written when the data is produced. So now, with all of these new mandates, you know, NIH, NSF for good data sharing plans, I think this is going to get a lot better. But for things that were produced, you know, in the nineties before your documentation is going to be what it is. There's going to be nobody going back and looking at this. Because you know, these people who produce data don't anticipate future needs. So you know, this is problematic, because you know, we're using different data in different ways. And so researchers are always looking for the new and novel and they're trying to repurpose a lot of these datasets in ways the author never really intended.

Next slide.For example, this came from a senior essay writer here at Yale, and it's about 2 years ago I got this frantic email. Lots of exclamation points, lots of the use of the word crucial, critical, important, and it's like, "Oh, what is happening here?" And so I'm looking at this, and she was very stressed about the fact that she couldn't get the survey results to work. I open up the email. I look to where she's going, and it's a Roper Center-produced dataset. And I was very happy that it was an email because the look on my face would have been like, "Well, what's the problem, you know, like what?" So, and you know, thinking about it again, you know, I looked at it and here on this slide we've got the questions. Granted it's all in Spanish, so let's just take that part out. But we've got the questions, we've got the schematic and we have this ASCII file. Now I've thought about it and I said, "Well, what's the problem?" And so this is good documentation, it has everything, all the parts she needs. Then it occurred to me: she's never seen a fixed-width ASCII file situation before. She's opening this thinking this file is corrupt because she can't see anything in it.

And so in, you know, talking to her and looking at this, and then having to explain to her that this is good documentation, and this is good data. It's, you gotta make that connection, because not everybody is going to understand that. And so I now, you know, learned from that that I always actually ask somebody, "Do you know how to use this?" Because guess what? The documentation is probably not going to tell you how to open an ASCII file, because in 1985 when this survey was done, that was standard practice. You would have stored it that way. So now, thinking about the researcher not anticipating future needs you know, you would have never put that in your codebook to say, how do you actually open this file? So these are things we need to think about when we're helping users and trying to marry, you know, good data quality and good data documentation together.

So always keep in mind that it's all relative to the user and the use. So you can have the best documentation in the world, but if the user doesn't know how to use it. Is the data any good? Or do they assume the data is bad if they just don't know how to open it? Next slide. So other realities. So there is that, you know, anybody who's worked with a commercial vendor, you may not get the documentation that is fantastic like you would out of IPUMS or ICPSR. And you have to ask yourself, is bad documentation better than no documentation? And sometimes you just don't have a choice. You know, researchers are willing to work with data that doesn't have a lot of good documentation, or the documentation may not be what you're expecting.

So you have to be able to sort of roll with it and figure out how to make that work. And then there is the understanding of data drops over time. So we know that people don't like to document things. I mean, If you've ever worked with anybody we're dealing with a data management plan. It's, you know it's not pulling teeth. They'll do it. It's just a matter of they just want to know what is the bare minimum I need to provide. And so you know, keeping that in mind that the knowledge about the data will drop over time. Next slide. So in this example, this is something that still continues to haunt me. When I got to Yale, I inherited oversight over something called a Social Science Data Archive.

And in this archive there is this set of files. And they're database files, they open up in Excel. There's no documentation associated with this. At some point in time, I want to say, you know, in 1997 when somebody put these files there, this meant something. Means nothing to me, means nothing to anyone I can ask about it. And you know I look at this, and I'm intrigued by it because it's all about World War II, you know, the 101st Airborne and it's got a ton of information in here, and I can't figure it out. And so you know no documentation and time is probably our biggest enemy. So always keep that in mind. Because you know, what are you going to do about it? But yet, you know, 4 years later I still have not thrown this stuff out, because meaningless to anybody else, it's meaningless to anyone. But yet I still haven't deleted it in the hopes that you know something will come out of this.

Next slide. So how do we fill in some of these gaps when we find that there are problems with documentation, and people really, you know, need help? So you know, echoing what Ron has talked about, you know. Look, think about who produces things, think about who uses things and really work with the network that you have built. So next slide. So this is just a very, you know, simple oversight, you know, overview of, you know these sort of 3 ways we get usually get data. So if we need to get documentation from a commercial vendor, you contact them. Sometimes that's, you know, easier said than done. You know I also want to stress, the more I look at data acquisitions a lot of information about the variables, about the data dictionaries, about code books are included in the license. So if you are not somebody who is closely-aligned with your collection development people, or anybody who acquires data, talk to them and say, "You know, is there anything in an appendix in the license? Are there other information that you know we should include elsewhere that's not just housed with the license?" You know, vendors change all the time, you know, and that's also part of the challenge with the documentation. So, as you know, companies get bought by other companies we get that data loss. So that person who worked at you know, Company X for 30 years, who knew everything about this dataset that you could call, that they would supply stuff, they retire, they move on. Then, all of a sudden, you have a different person, or that part of that company has moved to another company you get a lot of loss. So you know, be cautious, but don't give up. Always call the vendor. You know, when you're dealing with government data you know, anybody who knows me knows: you pick up the phone, you call them. I have talked to I don't know how many government agencies. I typically I call them.

Barbara Esty: They will call you back. I've had very good luck with it. I, you know, it's a hard sell when I have a doctoral student, and I say you need to pick up the phone and call them, but call them. And then individual researchers. They're sometimes not that difficult to track down. You know, we're living in this time where you know these original data sets that academics produced and gave to other researchers on demand like you could email so-and-so, and they'd give you the data, they'd answer questions, they'd provide information. They've retired. Some of them have passed away. Some of them have just, you know, left academia, or you can't find them. And now you have this missing piece of, I can't fill in the gaps of the documentation. And, you know, it's kind of difficult to track down, or you have to call a co-author to say, "Can I call this person?" or, "Do you know who I can call?"

You know, I mean, but it is very satisfying to be able to. You know I was asked about the Dundee Corpus. The Dundee Corpus is an eye movement dataset. It's a linguistics thing that is well known, and we had some researchers here who wanted it, and I was just like, I said, "This man's retired. What do you want me to do?" And so, you know, after quite a bit of back and forth, I found this man in the south of France, and we made that connection. So it does work. You can make it happen. It's just a matter of being, you know, it's doing the leg work. Next slide. So then you can also go to, who's going to use it, or who uses it? And so this is where the working papers, you know Ron has talked about you know, looking at the literature, you're going to find a lot. And this is also going to be another indicator of the quality of the data, because if nobody is using the dataset that you're looking for, that's a red flag. But look for working papers.

And also use doctoral students, not in a bad way, but doctoral students are intrepid souls. They will hunt things down. They will help you out. They will also, you know, produce documentation if they're using a dataset and they will share it with you. They're very open to sharing code. Work with other people. Talk to people. They're great. I encourage everybody to make a friend with a doctoral student. Next slide. So here's an example of where you think you've got fantastic documentation. But again, relative to user and use. So Form 5500s are essentially annual reports of retirement plans. Every company who has a retirement plan or a welfare plan, has to file this with the Department of Labor. They used to have to be filed with the IRS, so that adds a level of complexity. But when you go to the Form 5500 website, there's tons of documentation. Tons of it, and they've got lots of data available. It goes back to 1999 on the website. And you think this is great. This is a gold mine, because the data is all here. They've got schematics. It's in text files. They've got the data dictionaries. You start opening these things up. They're all different. So every year. So in the field position in this slide, you've got the website on the left, and the second piece is 1999. Third piece is 2021, and then you know, you've got these fields are all different, and then they add another dimension of the rules changing. So to the average researcher this is just, you know, crazy difficult to sort through and figure out, "Where do I start?"

And so this is a really good example of going to people who use the data. So I, you know, have talked to many an economist at the Department of Labor to ask them questions about how to do this. You know you search for Centers. You know Boston College has a Center for Retirement Research. You know you talk to your vendor because you're probably going to have to buy other additional data from someplace else, and you start comparing notes. So you kind of have to mix and match these things together. And you know the key is, you have to talk to one another and save that information. And so you know, just because there's a lot of documentation does not necessarily mean it's useful to every researcher. Next slide. So that brings us to our network. You never really realize how big your network is until you're looking for something. So you know everybody has a network. I encourage you to talk to friends, colleagues, neighbors, you know anybody you know. We are all more than happy to help, and you know it's, we all have to realize that we can't do this alone. That you're going to get questions that you can't answer, and you know what, they're going to come back. So you might as well share what you know, and if somebody asks a question, help them out.

Next slide. So this example brings together a lot of things. This is actually, this is sort of the Holy Grail, when things come together. So the New York Botanical Gardens has a person who has the best job title ever, called the Director of the Forest. I love that. You know we could all be the Director of the Forest. And he was Yale person and he inherited these discs. and he contacted our Digital Preservation Department to see if there's any data logs or any data he could get off of these, because he's trying to update the tree inventory at the garden. And so the one of the Digital Archivists, David Cirella here at Yale, he imaged these discs. He contacted me because we had worked on a project before, and he knew, you know I love a good project. And so the 2 of us went back and forth to see what was on these disks. He re-imaged them a couple of times. Got to them, there was, you know, some SAS files, some other files. I looked at them. I contacted a former coworker and friend of mine, to help me with, who helps me with all of my SAS problems. And so he worked on this a little bit. And then, by some magic the Director of the Forest tracked down one of the original data collectors and researchers on this project. So this is data from you know the seventies that he was had, and all of us are on the phone. And so this got, you know, data collected, you know, documentation created you know, on the fly together, bringing this all together.And it's really satisfying to see all this work. And so the fact that we could bring together the data, the codes for the data. We could bring, you know somebody who actually could remember what some of these abbreviations meant, what some of the labels meant, how things were calculated, was pretty amazing. So don't give up. Like that's my, that's my plug there is don't give up, these things can happen.

Next slide. And so think about documentation as an active process. It is not finite at all, you can always have better documentation. You can always add to the documentation. You should add to the documentation, if only to be kind to your future self, because these, you know, I always say there's no new research. When you think some sort of question goes away, it comes back full force. So you know, you will probably see these questions again. Anytime you are working with a vendor that you have data from before you can ask them if there is anything new with it, that you know, do they have anything additional? Have they gotten other questions? You can talk to others who use it. You know, Bobray and I talk all the time about documentation. I go, what do you have? And you know he'll send me what he has. I'll send him what I have. Or we'll say, do you have a new rep for so-and-so? And we exchange information which adds to the documentation, which in turn increases the quality of the data and the research output.

Barbara Esty: Next slide. So all of that to say don't be quick to throw stuff out. Like, really, it comes back. Like, so you know, things could be useful. So you know I'm not saying become a pack rat. But if you've got some space, and you can store some old, you know whether it's files, old media, anything you inherited when you took on your job, I would say leave it alone, and you know, it will come in handy. Next slide. And get it out of your email. I've been trying really hard to get a lot of this stuff that's sitting in my email out. And so right now, I'm trying to, you know, package it into a Teams site So that way it can be, you know, it's there. It's not sitting in my email because I spend a huge amount of time looking for emails from vendors, looking for emails through, I saw this on the listserv. I saw this, you know, somebody sent me 40, you know. We might have 40 emails about a purchase. It's a lot to go through. Save yourself some time, create better habits up front and then you can spend more time trying to hunt other things down than go through your email.

Next slide. And so thank you. I hope this gives you some hope that this stuff can actually happen. This is, just to tell you, to keep me honest, this is a picture of the bookcase that's sitting in my office with a collection of mysterious things. Everybody should have a big box of CDs that nobody knows if they can be used in their office. But you know I'm a collector of stuff. So if you're looking for something you can always call me first. Thanks.

Bobray Bordelon: Alright. So my mouse has been freezing, I swear, every other slide. So when you all are screaming next slide, I'm trying, I'm not sleeping. So what we're going to do now. This was, as I said, I had so many issues. We're gonna start answering questions. So give me one moment because I have not seen any of the questions that have come in. I know that there are lots. I think it's just frozen. Sorry, I'm just going to minimize this. Okay, let's see what we got.
Bobray Bordelon: I'm not necessarily going to go in order, and some of these may be similar. I'm going to do a quick read-through. As I said, I wasn't able to see the question since they came in.

Ron Nakao:

Any thoughts on using programs like NVivo or any other app that may be open access/open source?

The focus of what we are talking about is predominantly on survey or numeric data which does not preclude the importance of other kinds of data that are used for social science research. NVivo or a program like NVivo is used in qualitative research to organize and analyze interviews predominantly; but also useful for other kinds of structured text. For qualitative research, a program like NVivo is invaluable. In the past, these researchers probably would write transcripts of their interviews and put them on postcards, and then use different colored highlighters and stuff. A software package like NVivo is invaluable for organizing and doing analysis.

QGIS is an open source version of GIS software. Many may be familiar with ArcGIS. R is probably the predominant software that's open source that's used in quantitative and qualitative research. They have a whole host of applications and modules and plugins and packages to do almost any kind of research. It is an open source community that is very active.

Bobray Bordelon: I'm going to combine the next 2 questions. The basic question is about why data might not be available. And you know someone, I'm not sure it's the next person, but mentioned sometimes it's financial payrolls. But in general, Ron, why might data not be available?

Ron Nakao: It could be due to confidentiality or the kinds of restrictions of the data like you have very sensitive data such as health or data from minors. It can include other kinds of confidential data. There is a confidential aspect to why data might not be available. There is also copyright. It's basically, especially when you're dealing with more companies, one that we delve into that I don't know if others have is CoreLogic, which is a company that is in existence as a corporation to do stuff, and we were my colleague, my former colleague, Kris Kasianovitz, basically was able to negotiate access by Stanford to get a dump of their data which proved invaluable. It's an example of what I consider, maybe more administrative data that has so much value for research. But it was so messy, and it's copyrighted. And so you can't. You know, it's not open source. To me the ideal is public domain, open source, freely, publicly available. But the reality is more and more the data, either because of proprietary reasons, or because the confidentiality is not just readily available. Sometimes, if you're lucky, the restricted or confidential data is available but you have to go through another process to actually apply for access to the restricted data. Barbara, you may have something, or Bobray, on that.

Barbara Esty: Well, I mean, I also want to say that it is entirely possible that the data never existed. So I think of. I remind people when they ask me questions about the Census. I said the Census was never designed to see if you know how many people can read. It was never designed for these things, so the fact that we can get some of the information about it has happened over time. And also keeping in mind that data is expensive to collect. So if we think about if I'm gonna go out and do a survey, I'm gonna collect what I need to collect for my research process. I'm not being altruistic and thinking of future researchers. So I'm not always collecting extra data. You know, we as data, you know providers always tell people to take extra variables, but when you're producing the data, you don't necessarily do that.

Bobray Bordelon: I'm not going in order. I'm going to read a few things that were more comments, just to kind of get them out of the way. But I thought they were important comments. One was, someone agreed about what Ron said about ICPSR, about the detail of the documentation, but saying that for data discoverability, there can often be overlooked metadata creation for data sets outside of ICPSR and the structures of repository platforms. One of the great things about ICPSR is they're coding things to the variable level and that takes a lot of staff, and it takes a lot of time. And so that's one of the reasons why that is so good. We had someone else that commented apropos to what Barbara said about networking, joining IS is a great way to build that. So in case anyone is not familiar with IASSIST. Some of us could put in the chat, but IASSIST is data.org. It's an international community of largely social science librarians, but we're open to all we do have science librarians and so forth, both quantitative and qualitative.

It's 50 bucks a year. It's the listserv alone is worth it. We also do a lot of free webinars. And then someone mentioned the summer courses at ICPSR. So those are both other ways. And again, if someone's not familiar, if you just Google ICPSR summer program, you will find a lot of information on it. There's an intriguing question which I don't know the answer. So maybe Barbara or Ron does. What sites are available where you can find data validated by a peer review process? I mean, my personal answer would be, you know, this is what the literature in some ways would do. But I'm not, I don't know, like I know archives. Or can get a certain certification to say that their curation process you know, has met these incredibly high standards. But if I'm interpreting correctly, it's that the data has been validated by peer Review, and I would say, look at articles that have gone through the process. Barbara and Ron, do you have a better answer, perhaps, or an additional?

Barbara Esty: I might say check repositories. Because I know when I talk to people, you know like especially medical, usually it's a medical person who needs to submit something to a repository for peer review. Now I don't know if a lot of these things could be flagged in a repository that it has been peer reviewed. And so that might be something to ask you know, your repository like, I'm thinking of things like Dryad, or things like Figshare, or something like that, where there is other processes in place outside of, you know, a university. But that might be a way to look at it.

Ron Nakao: Yeah, I raised that concept, the peer review, and it's not some, I'm not aware of any single source that you can say is this a peer reviewed process. Because this data has it been appropriately peer reviewed, because there's different degrees. I just mention it more as the peer review process is really the research community's sort of goal or hopefully adjective to basically vet out and validate the data that they are seeing being used by other researchers. And so often you'll see something where, say an article has been retracted or corrected, because someone basically went to try to validate or reproduce the results of an article based on the data, and they found that they couldn't reproduce it. There's been some scandals about that in the past. A hopeful sign is with NSF and NIH and others requiring data management plans, there's kind of a pipeline being created to more kind of validate. And I think a lot of the major Social Science associations like the American Economic Association, the American Political Science Association, etc. They are now instituting a validation process for any articles that they will publish in their journals. So they actually have teams of people that will actually go through with the data and replicate any tables or other kinds of results that has been done. So at least you can trust that the results mesh with and are validated by another person to the original submitter of the article. So it's just another way, to me, the infrastructure of good research is peer review is part of that, and so something that's been out there and cited a lot, chances are it's probably, you can trust it a lot more.

Bobray Bordelon: And just to mention Lars from the American Economics Association is one of our future speakers. We'll talk about the process that the American Economics Association does. I have one question here. We have a lot of questions, we're not going to probably get to them all, but we promise we will put them into a document or something with some answers later. Someone says they're very impressed with the amount of work that Barbara does to help researchers access data and find data. When do you decide that this is a role that the researcher needs to take on themselves? So what point do you say, you know, like I've done what I can, but this is your job.

Barbara Esty: It depends. I mean, I think it depends. It depends on the output of what they're trying to do. I mean, if it's a faculty member, I will try and prove the negative that we can't find it. And so that's sort of where I take it that far. I mean doctoral students. They're great to sort of keep, you know, tossing them ideas, and they will go out and do these things, because at this point it's still fun, you know, for them it's still fun. And then, you know, undergrad, I don't necessarily think, I usually tell them, okay, we need to find you a new dataset, because this is taking you too long. You're not being evaluated on how creative you are in getting your data. You're being evaluated on the analysis of your work. So it is going to depend upon one, some of the time I have. But also you know who the researcher is, and what their level of you know sort of importance is. I mean, I hate to say that you know not everybody gets the same treatment. But if it is a senior faculty member who's asking this, it's a little bit different than somebody who's got a class assignment.

Bobray Bordelon: Someone mentioned data journals as another way to kind of validate things. Another question we have that was an early question is, Did you have a systematic method to detect the data quality problems and methods to evaluate the data quality? I mean technically, not theoretically, especially large scale data. So essentially is there a way to go through and see how it is?

Ron Nakao: You know I, I'll take a first stab at that, and it kind of it kind of piggybacks on the previous question. I think it's important for us as data librarians to not, do more than we can do and do more than we have expertise with. I think, for those who have more experience or have been in this business longer, you do learn things, so you can actually do more. But I, you know, especially when I'm talking with folks who haven't really, are probably kind of scared of data. I just say, okay, but you know what it means to be a librarian. So when you, when you purchase a book, do you think it's your responsibility if it's in another language than your native like, if you're an English speaker, and it's in French, do you think it's your responsibility to teach the user of that book how to read French? Do you, you know, at least a book has a certain kind of structure, but if someone who's not used to a structure of book, do you actually teach a person how to actually navigate through a book? No. I mean, we acquire materials, and then we expect the user to have to kind of figure it out, and then maybe report if there's problems with it, or ask us to get something else. I think with data, because of the unique nature of it, and it's not really been standardized, that's where we've had to learn to provide more support than in other areas. And I think it's important not to be intimidated or to be forced to believe that you have to actually be a researcher yourself and do that. Because I think when push comes to shove, it's the researcher's responsibility to do a lot of this stuff themselves, because they're the ones who're going to have their name on the article.

Bobray Bordelon: We're often asked to buy data from commercial vendors. They may be hesitant to provide documentation and methodology. Should we discourage use of commercial data?

Barbara Esty: No! Because we don't usually have a choice. But at the same time, if you don't push it they're never going to change their behavior. And you know this is where you know your sense of injustice has to come out, and you have to actually force the issue. So I think a lot of this comes in with the due diligence process of talking to your vendor. You don't have to be mean about it. I don't think that, I think most vendors, if they have stuff they will supply it if you ask. And I think a lot of the thing is that you're expecting these vendors who don't normally deal with academics to act like academic vendors. And so you know, they're not necessarily going to be up front about things so you need to ask. And so you sort of have your list of questions which may vary from person to person. I'm happy to share mine of where I kind of go through my list of like, okay, you get ready because I want, you know, I want to know, and I will make them go into their data and provide me with counts of things like, I want to know how many observations you had by this time for this country, and I want to see this, and I want to see what the variable list is, and this all happens before the purchase is made.

And you know, then you continue to work with the vendor as you get the data, as you unpack the data. And then you also start to talk to other institutions who have it. And just to compare, did you get anything? Do you have anything new? And you kind of work through it, and you force the vendor to help.

Ron Nakao: Yeah, I'm kind of reaching, I'm kind of getting to the point where I'm putting more caveats. And I was very enthusiastic, like Barbara before, about like let's just get it if we can. I think i'm at the point where I'm reevaluating because part of the responsibility, I feel as a data librarian is not only to acquire the data, if we can, for our researchers, but to make sure that we can serve it, that it's accessible, that it's usable. And if you get certain data that really is so messy that only maybe the most advanced user or researcher can actually use it then, and it's so expensive that it sucks up a lot of your limited funds, then you have to start making decisions about, is it, do we acquire this or not? I mean, in some ways it's kind of analogous to special collections. Do you pay half a million dollars for the only copy of a rare book or a rare set of collections when you don't even have the curators to actually make that collection available, let alone the fact that it costs so much. And I think that gets back to another message. I think it's so important if we want data to be documented, there has to be more data curators hired and available because it takes a different skill set and and focus to actually curate data.

Bobray Bordelon: Speaking of curation, someone said: They are curious to hear the presenters' thoughts about the role of web archiving in data curation and data reference. They said, I'm thinking of those academic researchers who document their data on their personal websites, and how those might disappear when they retire, etc. So essentially, instead of putting into either the institutional repository or ideally a domain repository such as ICPSR, the Peptide Data Bank, something like that, they're just putting it on their own site.

Barbara Esty: I think that's a reaction to the fact that it's not easy. I think we, as you know, data providers, data support, I agree with Ron, there needs to be more of us, because that's part of it. The day goes by quite quickly. But we also have to make it easy for researchers to deposit something. We can't, and so we have to encourage them to put it some place for someone else. and I think that's the only way we're going to change that behavior.

Ron Nakao: Yeah, I think that's always a challenge is changing the culture of research where they are thinking beyond just their own immediate needs for the data that they collected and thinking of, in a sense, doing the social good of making their data available for others. I think it's unreasonable to expect researchers to know how to document and curate their data. They know their data so that they can use it. But I would guess that probably 5 years from now they won't even remember their data. And so they also need to document their data so that they can refer to it in the future, let alone other users to know how to use it correctly and appropriately.

I think that's one of our roles is to go out and find these datasets and encourage and support them to deposit and share their data. But the other part of the equation, which is, we need more data curators is to make sure there's some minimal level of documentation and curation of the data, so that you're just not saying, go on, download these these 5 files and good luck figuring out what it's what's in there, kind of thing. Or good luck emailing the PI and seeing if they'll answer your questions about the data. So there's a reason why we are where we're at, and it's the reason why it's been so valuable to have such a long lived archive like ICPSR trying to figure out how to crack that nut, trying to make important data available for others to use.

Bobray Bordelon: What ethical issues do you encounter in data collection or use that librarians may be able to help researchers better understand or consider? I think this is kind of a-

Barbara Esty: The first thing I can think of is the, we're buying, we're acquiring more data that has more details about people. And the big question is, how does it get stored? How does it get used? What, who's responsible for ethical use? And I have yet to get a really clear answer as to whose role that really is. I mean, I like to think it is the researcher. I'm always pleased when there is a you know it's usually a doctoral student who will say, alright, I'm scared to use this, and I go, "Great!" Like, I am so happy to hear that you are aware that this is sensitive. Like, you know, and it's all public record. But it's a matter of when things start to be combined, and we do have to think about the ethics of it, or you know what the implications, even if you're trying to solve the world's problems, you have an awful lot of information about people in one place. And I don't think that, or at least in in my institution, we don't have, we don't have a clear answer for this. I think it gets answered as, you know, it's the researcher's responsibility.

Bobray Bordelon: Next question I will take and Barbara and Ron can add if they have others. So one thing we did for this time around is we tried to advertise to every iSchool, library school, there was. And so there's a question specifically for them: any advice for library science students interested in data-adjacent librarianship? And I would add, you know, for anyone who is perhaps a generalist, is perhaps thinking of doing a career add-on, or perhaps they were forced to do this. So for those of you that are in school, very few of the schools tend to offer real programs. I know some of them are now doing what they call data science programs. I call them data science light because they typically are not really data science, but they're also not data librarianship. My advice is, there are so many wonderful free webinars out there. Take as many as you can. You know, Ron and Jane Fry, who's at Carleton, and I have taught a summer workshop many times, and one of the things we say is no one can do everything. But unfortunately, if you're at a smaller place, you may be asked to do everything. So I would say, take advantage while you're in school. Take a Python class, not to say you're going to be the Python programmer, but at least understand what it is. Take an R class. You know there's a lot of free webinars on how to do a data management plan. You may not be the one that does this, but you'll see how it is.

But the other for me it's the traditional library classes are still important. If your place is still one of the ones that offers subject reference classes, take those. You need to understand. To me that is honestly more important than the data management plan and all these other things that you know are out there, because this is going to be the thing of how you actually answer the question. So it's not to say you're the expert, but that you have some understanding of those fields and the methodologies that are in. So just take advantage of everything out there. And you know, look for free webinars that the Census Bureau does, the Bureau of Labor Statistics, ISS, ICPSR. They're so many groups that you don't even necessarily have to be a member of to take advantage.

Ron Nakao: You know I look at how I came in through the the route of getting getting a dissertation, so I actually was a researcher at one point before I went into librarianship. So at least I had the grounding of what the research process was and then I had to learn how to become a librarian.I think the most important way to actually acquire this is to do an internship with someone who is a data librarian. Network with others like we mentioned, ISS or other communities of folks who have different levels of skills. And then you can you can actually ask your questions. It doesn't have to be in person. It could be through through email or other kinds of networks. But basically, it's kind of to build a social support network, professional, that you can then learn, but also eventually contribute to the collective knowledge. And I think that's the best, because it's kind of hard, like if you remember the first time you took, say physics or whatever, that you're trying to read every word of the physics book and yet you come away, not maybe understanding what physics is and you have to have some distance, and you have to actually try to apply it before you really start figuring out how to really control and know what is appropriate and not inappropriate for what your role is.

Bobray Bordelon: We have a lot of questions. Let's see. How common or helpful is it to deposit data documentation, but not the data if you're using restricted secondary data. I say incredibly helpful, because one at least you know what someone used. ICPSR has started a process for making documentation available for data that is only available in restricted data centers, because in the past you didn't even know what there was there, whether or not you could ask. I will also add the example that Ron used is really exceptional. They pointed out in that thing what were the fields that were restricted? I think we all spend so much time trying to figure out what is restricted other than geography. So that was a truly wonderful example. But yes, knowing the documentation at least shows you whether or not you might want to pursue it. Uh, let's see. Um, would Barbara and Ron, if applicable, be willing to share their list of questions they ask vendors when buying data, trying to get documentation?

Barbara Esty: Oh, sure, I can do that, if you know. If whoever asked that question would want to email me, I'm happy to email back, or I can also include the my list with our list of question/answer with the slides. Whatever, I can do both, I can do all of those things.

Bobray Bordelon: Someone pointed out, I'm gonna read a bunch of kind of quick comments just to get them out the way. The World Health Organization has a no-cost webinar series for working with healthcare disparity data, and you get a certificate. So again, this is the type of thing, one, it's useful and two, if you're in a library program or something you could actually use that. So -

Ron Nakao: The Federal Reserve Bank at St. Louis actually has kind of a self-paced way to actually use you know their data, access the FRED data there, which I thought was very good. So there's a lot of stuff out there where you can kind of do self-paced kind of tutorial lessons.

Bobray Bordelon: We have a question about, can you trust commercial data without the documentation? I would say it would be more, can you understand the data? I mean again, it's gonna come to the authority and everything else. But you may not be able to understand what it is if not. There's some great examples in the Q&A. We will include all this in some form along with the slides. But Alice Kalinowski at Stanford talked about a place that publishes papers with data. So we have a number, I want to just echo what Ron said about working papers. You know those often have, are better than the published article.

Because often in the published article they have to chop out all the stuff I care about, which is the actual, you know, documentation. So definitely look at the working papers too. Nowadays they tend to be available online.

Ron Nakao: Another source are dissertations. I mean, one thing about dissertations is they can be like 600 pages long, and they actually write more than you would ever need to know about every aspect of that thesis. It's a, there's a little bit of a lag time, so it may not be the most cutting-edge data, but if the the data is being used in a dissertation, it's probably pretty well documented.

Bobray Bordelon: Someone asked about software used to collect, store, analyze, interpret. There are many out there again, I mean, I like if this was someone just starting out these days, I mean in the end it's going to be what your university uses. But I would say more and more, people are going for the open source. So learn R. You know, Python, perhaps, go for the open source, but in the end it's gonna depend on what your departments and so forth are using. So we still use data very heavily at Princeton we're about almost 50/50 between State and R with a tiny bit of SPSS. But it really just depends.

Ron Nakao: I mean I'll do a shoutout for Python, I mean, especially since in the social sciences in particular, but larger, because everything so interdisciplinary, Python is is used a lot for people doing TDM, text and data mining. And so it's another tool, and it's open source that is being used. And so between R and Python, I think those are tools that are gonna be kind of
Ron Nakao: the next generation of sort of the go-to tools in the social sciences.

Barbara Esty: Except for when again, you're dealing with large and larger datasets that we are acquiring you're gonna need different programs to break them up. And a lot of times the open source doesn't do it. So you will need to, you know always, I would advise anybody to, you know learn some stata, learn some sass, learn something so that you can take these massive files and break them up because no one else is going to be able to do it. And so you know, that's the only way you're going to actually be able to get to usable formats.

Bobray Bordelon: We have a great question which probably we're not going to be able to have the time to justify, not really about documentation, but it's one of my favorite questions we had. Have you encountered data leaking problems that happen accidentally or on purpose? And how did you deal with? I mean, it depends on what we're talking about. If we're talking about people sharing when they're not supposed to, which happens all the time with commercial data.

Uh, yes. In terms of where something got something that was restricted and they accidentally made it available. Luckily I've never had that. And I live in fear constantly, because I sign all these things but I have not had that. Barbara and Ron, have you had someone violate a restricted data contract?

Barbara Esty: Not that I know of not nothing so blatant that it's a problem. I mean, I think there's always the question. I mean, I think now, if I can plug for anybody who's working in data acquisitions to, when you look at those licenses, look at who the user is defined as and advocate for co-authors at other institutions, because that, I think, has been the biggest sort of data leak, especially when you're dealing with expensive or specific survey data of some sort that it becomes a problem. Or else, if you have can write into a license that you can contact them if it needs to be shared, so that it can be appropriate.

Ron Nakao: Just one last point on that is, I think that's the importance of the data librarians and experts to be involved with the negotiation and licensing process with the suppliers. I think there's too many datasets that are free at our institutions that say a faculty member just signed off without even rereading the terms. And but I think we at Stanford are trying to be as responsible and putting in negotiating clauses that deal with, you know external users, or when a student graduates. Can they still have access for peer review, you know, or for you know, the publishing kind of process. But often, to be honest, a lot of it is that we kind of say we will do the best we can to ensure that our users will follow good data and computing hygiene practices. But we are not the police for this, and we cannot be held accountable for those kinds of issues. We can go and take, you know, a lot of the clauses, if we find that you violated the terms of use then then you need to delete the data. You can no longer use the data. But yeah, it really is kind of a thing, as Bobray said, something that you can stay up at night if you really get paranoid about it.

Bobray Bordelon: We have dozens of questions we're not able to get to. There are quite a few dealing with public service, offering the service, etc. Some of you, many of you have actually taken a class either with myself, Ron and Jane Fry, or in the past, with Jim Jacobs, Chuck Humphrey, Diane Geraci. It's not going to be offered the summer. Unfortunately, we're all a bit overwhelmed. But maybe next summer. But again, just look at, there's a number of good books that have come out in recent years dealing with data librarianship that I would highly recommend. I'll put titles into the thing once we actually try to tackle some of these questions. But, as I said, I know that we didn't get to a lot of them. But you know I'll try to myself try to answer as many of these things as possible, and they're not going to go personally to people, but everything will be anonymized and sort of put into a document. No promise on when, but this will happen. I want to thank Ron and Barbara so much. I want to thank the entire team. I want to thank the people at Stanford who are handling the technology. All the glitches of my keyboard are mine, not theirs.

Bobray Bordelon: I want to again thank the IMLS for providing a grant. I want to remind people that our next lecture is going to focus on quality assurance and data creation and that is going to be on Thursday, February 23rd. You can, you will be able to register. I believe that just showed up in the chat. And you can also provide feedback about this session at the bit.ly that is provided. And i'm going to stop sharing, so that I can see again. And, I think we're pretty much out of time. So I'm not going to go back to the questions. Will the chat? Let's see. Okay. So, please consider filling out the feedback form. Well, thanks, everybody. We hope to see you at future ones, and, as I said, we will attempt to try to get through as many of these questions as possible. Some of them I'll probably rewrite and kind of combine.