Finding Social Science Data using Federal and State Resources – Transcript of audio

Please stand by for realtime captions.

Good afternoon and welcome. My name is Helen Keremedjiev and I'm a labor and at the U.S. government public office, GPO. I will be the MC for this today and Ashley Dahlen is providing tech support if you need assistance. Today's webinar is, Finding Social Science Data Using Federal and State Resources . It is by Eimmy Solis . The Q&A during this talk please add your questions into the chat and send them either to all panelists or all participants. During the live demonstration, mouse over to the blue bar at the top of the screen and click on chat to activate the chat. I will monitor the chat and questions will be answered at the end of the talk. In addition, I will add into the chat that featured links in the slightest as our presenter talks. This presentation is being recorded and it will be made available shortly. I will man now had the microphone over to our speaker who will take it from here.

Thank you so much for the wonderful introduction. I am now sharing my screen. By the end of the webinar, you will understand the data landscape, the data life scale and the limitation on opportunities of finding data. We will also explore strategies on how to access publicly available social science data and information. Please, understand that I will only show some highlights of things that I find to get this data, not a conference of list. More of a jumping point of how to start. First of all, what is data? We talk a lot about data. Data includes a set of raw numbers usually with many variables in the capability of being [Indiscernible]. On the screen is an example from a data set from the office of postsecondary education from 2017. It shows the number of participants in football for each university in California. For example, where I work, the University of Southern California, a total of 109 played in 2017. This is not saying that the record, if they good or not during the year. You cannot really compare the colleges on the list because on the list is Santa Monica College, which is a great community college in our local area. They had 125 football players that year in 2017. That does not mean anything because they did not play against each other. They are in different divisions. They are not comparable. This is just raw data. Social sciences, there will be researchers in economics, they often want the raw data to be analyzed, also outside of social sciences, engineering often asked for a lot of raw data to do their own analysis as well. What are statistics? Statistics is data that has been analyzed in some way and has percentages, a chart, or graph -- sometimes it is called statistical data. Statistics usually represent and help you visualize the data connections. It often answers common questions, like what is a summer average. In this graphic, I am showing you from USA Today. It shows the total cases of wine produced in different states. California produces the most wine. USA Today is getting the best from vines analytics. I look them up and they are a marketing research firm that serves the wine industry. You want to see the source and if it is a source that is reliable. In a few slides I will show you how to find a more reliable source to back up this information. You always want to think about, do you trust the source that you find, especially when you are looking online. Just to recap, data is a set of raw numbers with many variables with the capability of being manipulated. Whereas statistics, on the other hand, has been analyzed in some way. Statistics are represented in graphs or charts. My colleague gave me a really great metaphor that I use a lot of my classes that I teach. I think it is very appropriate. Data is just like raw fabric but statistics is what becomes of the fabric once it is cut, sewn, and pieced together in a short or dress. You may get it from different sources or places. Depending on your research questions, you might need data, you might need statistics or you might need both. It is good to understand the difference. Before I get into some sources to look for data, I want to think about what we need to think about? Also, when you're working with researchers and students, or patrons -- what type of questions you want to ask? You first

want to look at the what? What type of source you need? What is the main purpose or goal? What indicators are variables do you need a? What combinations are you examining. Two, who cares about your question? Will the institution or government easily share with you I just had a question that I was working on with the social work library here at USC, she wanted to know the number of people released from prison in L.A. County. Ever so slight, who would know this question? Who cares about this question? The Department of Corrections, they care about the question. They are the ones responsible for that type of statistic. That is where he went to to find the answer to the question. Who would you think manages that type of information? What government surveys might ask the same question you are asking? I also recommend doing a literature search, looking up articles on your topic and seeing what data sources are used when you look at those articles. Who would know or care about your question? What geographies are most appropriate to answer your questions? I want all the data, but you want to [Indiscernible] as to what specific geography that you need. Look at data for County, countries, are you looking at national data? These are questions that really need to know before you even start looking at where you look for that question. And then how. How is the data collected? Was it a survey done? Was it administrative data? Administrative data is any government transaction that happens within departments. An example I have of this -- with motor vehicles, the DMV has a lot of information on us. It has our way, our height, our vision, how well we can see. It has the ability to see if you have been driving under the influence. [Indiscernible] you also want to look at analysis. Is the researcher looking for information of the particular individual? A household, a company? A holes information or national information? Lastly, when with this data collected? What is the timeframe? Unfortunate, not everything is online, especially things prior to 1990s. It is harder to find information and data. Often times, we do still rely on the current government documents. [Indiscernible] and then copy and paste them to the next cell of the document. Not everything is online. And in the frequency, are you looking for annual data, quarterly data? One thing that I want to say is that it is really difficult to find very recent data. For example, data from 2023. It takes so long to gather the information that it is fine if you find data a few years old. 2020, 2021, we are all working with the same -- no one else is going to find any more recent data. That is what I tell my students. It is fine if the Emperor nation is still a few years old. The data lifecycle. You might've heard of the data lifecycle with a different version of it as well. We have data one, I like to use it for educational resources that teach about data. I'm just going to quickly go over this version of a data lifecycle. I will start on the top. Researchers can dart anywhere, depending on the type of research they are doing. There is no correct way to do it. I will start on the top. You want to have a description of the data you want to compile, how the data will be managed. What are you going to do before you even start your research? Oftentimes you need a data management plan. This has become important since the NIH, the national Institute of health has started data management. Researchers now have to plan and budget for managing and sharing data. Now, we have had a lot of [Indiscernible]. We all want to be able to find this data as well. Again, it depends on the type of research you are doing. I have observations here. For example, a business to you might have an assignment where you work with marketing. You want to do a marketing survey you can go around your neighborhood or school and ask other people on the street about a particular product. Maybe a car or a phone and collect that information through a survey. The next step is assurance. To assure the quality of the data. You want to check to see if what you have will answer your research question. This is a part were a lot of researchers Michael back and collect more data because we don't really have enough information. They might be missing a question. This is a stopping point to make sure that you have good quality data. You want to also describe the data that you are collecting. This might include filing and having a record of what each file has. Also describing your variables. I had a focus group and research product on, even after a few years, you forget really quickly what you have done. This is important to describe the variables and everything that you have collected. Preserve. Data, ideally, needs to be preserved long-term. A lot of universities have a repository where students and faculty can preserve

their data for the long-term. Or there is big organizations [Indiscernible]. They also help provide a service to store and share your data. Other great places to open framework, they also have a great system to store and share your data as well. There is different ways to do it. It has become very, very important. Next is discovered. You might want to discover other data that might be useful. A lot of researchers might start here. They may not have their own data. They might just collect data from different sources, like a census, or commercial sources and then integrate it together. That is the next step, integrate the data sources to combined. They might have a question, for example, that they have collected and add demographic data from the Census. Lastly, analyze that data. Again, this is just -- this does not follow a clear path. Some research might have multiple revolutions of the cycle. They might do these steps multiple times and go back and describe more data and go back before they are ready to deposit their data sets into depositories. This is just guidelines. This is for researchers so they know how the different steps are done in terms of taking care of their data. Knowing about the data, I hope you understand the challenges and limitations of finding the specifics in data. A lot of data is public and can be found for Google. Everything is free. Publicly available resources come from government information, which is a federal, state and local websites. For example, there is really great nonprofits that have really great statistics. Researchers and faculty, through your university may put their own data on their own website. It is becoming less common but it still happens. Where specific people just put up their own data. I often tell my students too, contact the researcher. They might have their data and be willing to give it to you. There is various ways of getting this data. On the other hand, there is commercial and restricted data. A lot of it comes from businesses, extremely expensive. It is usually market research company, industry, and financial information. It is things like Bloomberg. Unfortunately, if you are not associated with the University, you may not have access to it. That is something I always like to tell my students, take advantage of what you have while you are a student. After you graduate, students lose access to a lot of these great things. For this purpose, we are focusing on the public at resources. I want to give a shadow to central public libraries. They are a great source for the type of business data. The central library in the New York public library are great examples where they do have Bloomberg stations and really great services for, for example, the small business owner. They might need that type of business data for their own research. Public libraries are one of the assets with access points to get the data. I'm going to now start showing you some resources, this is not a comprehensive list. These are things that I like to use and things that I've been finding about. The first is the government agency list on USA.gov. You will find a conference of list of all government departments. The agency provides the data on their website, not standardized at all. You might have to dig through the website to find the data specifics. It is a good starting point. For this example, I went to office of disability employment policy. You can see the official website, the contact information, what section of the branch they are in. It lets us know from where the policies are regulated. In this case there part of the executive department and the parent agency is the Department of Labor. I'm going to go bye, so I can show you another example of a through Z rated

---

As you go live, if you have any questions, mouse over to the top to click on the menu and click on chat, activate chat if you have any questions for the speaker. Back to you, Amy.

Back to USA.gov/agency-index. I happen to know that the department will Housing and Urban Development has information [Indiscernible] in a few ways I will show you how to search when you have no idea where to start. In this case you do have to go to the housing department and urban development. What is great you have the official website but they also have the sources by state. I'm going to click on HUD resources in your state. All my examples are going to be from California because that is where I'm currently living. When you are on the government website like this, there is no standardization. I want you to look for keywords, library, research -- those are really the top places

where they store the data in the system. I went to click here on ask questions. Here you go, this is where I was like I am looking. This is where I get the category looking for, library. That is a good indication that I'm going towards the right place and have my goal of finding the data and statistics in the site. And then I am here on -- and now I am looking for research. There we go, research. This is the jackpot here statistics and data about California. You want to look for library or research. That will show you where you need to look. And now we get various -- I'm going to go here on home sales. I'm going to try, hopefully, to buy a house. I am not optimistic. I like to look at housing resources. Here we have a nice slides are from a California housing market update. It is pretty recent, which is nice. This is a PDF that they also have the option of downloading the data. There is different data sets they have including slides. It is not looking -- yet -- it is not looking that good for me. I'm trying to be optimistic. The median house price is $836,000. Median days on the market is 17 days. It gives you more information on the housing market in California. We will go back to this in a minute, when I compare another site. Let us say, for example, you have absolutely no idea where to start. You get a question that you are trying to answer. You have no idea. This is where you use Google. I tell my students, if you are using Google and you're trying to look for data specifics, at least try to hone in on government sources. The way I do it is to search for -- using my keywords and adding site:.gov. This will take away all the other websites that are not from government agencies. It is really going to hone in on your question. I am going to go back to our wine example that we had earlier. You know, I need to look for a better resource, citations -- I know for a fact that California is the biggest wine producer in the United States but I don't want to cite that marketing firm. I want to cite a better force. We search for wine production site:.gov. I'm going to copy and paste this. And then go to Google. Site:.gov is all one word. And here we get our top searches. This is the one that I want. Wine statistics. This is coming from the alcohol and tobacco tax and trade Bureau from the U.S. Department of treasury. I would have never guessed that these statistics would be in the U.S. Department of treasury. When you think about it, the government want to tax alcohol and tobacco and they're going to do a good job in providing the statistics. That makes sense now for taxation purposes. I'm going to click here on the state to statistics. And it will give me a list of documents. Hopefully, it will open. Yay. [Indiscernible] the production is over seven 52 million. This is in gallons. California produces the largest amount of wine by a lot. 599 million gallons and no one else is comparable. Well, New York is over 32 million gallons. Texas, interesting. There is Washington state, 41 million. You can see there are quite a few states that also produce a lot of wine. This is the type of citation and source that you need to cite in your research, not a random marketing firm. I just reviewed how to use Google using site:.gov. I tell my students, do you site:.gov to get better resources. The next thing that I like to use is the federal porter, which is data.gov. It is a data site that is a little more open. This would be a source that I would suggest. Maybe you have information that you need on the climate, energy, agriculture with the local government. You can browse the topics there. I'm going to go live again and show you -- let me get out of this. We are back here at data.gov. I'm going to look at hate crimes. That is is a big topic that I get questions on a lot. It found 31 data sets for hate crimes. It is not just at the federal level. It is giving us data from cities, from states. This is really good. I like to see what type of data sets there are. I like to go directly to the source. I'm going to go to the uniform crime reporting, which is under the federal Bureau of investigation, the FBI and Department of Justice. The Department of Justice and FBI are large producers of crime statistics. I am not surprised I was like to go directly to the source, for here it is the FBI website. I'm going to click here on hate crime statistics. And then hate crime statistic report. These are nice to go through if you need detailed information. There you go, this is the latest one that they have, 2021. As I said before, this was released in March 2023. It does take a few years -- it does take time for the FBI to gather the statistics together. Again, I always tell researchers that if you are working with [Indiscernible] that is what is available. It takes so much manpower to get all the things in a nice place. I'm just going to go through some of the charts that they have. Hate crime incidents have increased from 20 2010 2021. Down here they go into more detail

about what type of hate crimes have occurred and then in what city. It has increased significantly in a lot of cities. For example, New York, hate crimes based on race and ethnicity have increased significantly, religion and Austin looks like it is a major bias category. It is important to know what is happening and how they are reporting it. This is the federal portal, data.gov. Your state might have a data portal. This is an example from California. We have data.ca.gov. There is a lot of other places that has theirs. States can distribute their data anyway they want. Many states have a data portal. You have to check your local state. Especially if you're looking for data by state, it is good to go directly to the state, local government websites. Cities also have data portals. Los Angeles has data at data.lacity.org. I did a search on gender. One of them was gender breakdown of city workers by department. It was statistics on employees, L.A. city employees. If you have a question, directly on a city, go directly to the city website to see if they have data for you. For social sciences, I would say that business and economics is data that you might use the most. A lot of the data comes from business surveys. [Indiscernible] already cover that in their great census series. There was economics 101 that was created a few days ago. Check out that for a very extensive view of economics. These are some other surveys, trade surveys that I used to answer questions based on surveys, export statistics. For this demo, I'm going to concentrate on labor statistics. Most come from the Bureau of Labor Statistics. They also work with the Federal reserves of St. Louis to provide a great data set. What I like to use is FRED. Federal Reserve economic data. It is on fred.stlouisfred.org. It is a great source for economic type questions. The library that FRED are amazing. They are really great at answering questions that you may have and may not be able to answer. I use FRED to look at statistics that students might have for an assignment. Oftentimes you like what the GDP. I go to FRED and they usually need that for the assignment that they are doing in undergrad. GDP, gross domestic product. What is great about this is they describe the variable, what is gross domestic product to market it is the market value of goods and services produced by labor and property located in the United States. You could see the most current and the shaded areas [Indiscernible] this is latest when we have during the pandemic. It did go down a little. I am not surprised it went down. It went down because we were at home and people could not go out and produce. A lot of us still work. It did not go down as much as you might think. Some of us were still working and producing. These are interesting to see. You can download the raw data, not just looking at the charts. You can change it -- change some of the variables up here. If you go back to the homepage, unemployment rate -- that is another big question that I get. Unemployment rate, as of now, May 2023, 3.7%. These are statistics that everyone should know, we should know what is going on. I had an assignment where students needed to look up their neighborhood and have different variables about their neighborhood. One said, the appointment rate in my area is 9%. Wow, that is a lot. If you look at the national rate, it is only 3.7%. It is good to know the national rate to be able to compare your own neighborhood and your own situation to see how the local community is doing in your area. You can go to the public library. You might want to offer resume building at the library with a high and appointment rate. Create a job search in your area. This is good for every day knowledge. I mentioned before, start looking for a house, hopefully, next year. I am looking at housing prices. This is one variable that I saw that I thought was interesting in FRED . The housing inventory meeting listing price. It has national data, also has specific by state and county. I'm going to show you California. The meeting listing price in California is 740,000 as of May 2023. Wait a minute, that is different than the first one. When I first showed you the statistics from HUD, from Department of housing, this is the median sales price. That was the difference. So that is higher, unfortunately for me, $836,000. It is in the same time period, it is comparable, May 2023. You can see here that I'm using government resources to look at my topic of housing prices. I'm not looking forward to looking with that price. I'm deathly going have to pay more when that time comes. I'm also going to look up Los Angeles County. This is where I get sad because the median listing price is currently over $1 million. Great. I do not know what is going to happen. Maybe I will get a condo or something. FRED is a great tool. It is good for novice people have no idea. But also, their researchers at the Federal Reserve [

Indiscernible] it is for everyone. For public policy, one might look at you as aspending.gov. That is always interesting to see you can explore their data here by budget function, by agencies it looks like the main thing it has is Medicare, Social Security, national defense. And then you can drill down on the different boxes to see exactly what they are spending on what., Military personnel, research. Aircraft. That is interesting to see. The last topic that I cover lot is education. Education students need a lot of statistics on enrollment and outcomes. The source for education is the national Center for education. They produce this great report every year, the digest of education statistics. Also the Department of Education also has some great resources. I will show you the education statistics. I'm here at the national Center for education statistics. You look at surveys and annual reports and then digest of education statistics. It goes all the way back to 1995, which is nice. Then you can compare multiple years. Right now, I'm looking at 2022. I like the layout of this a lot. It has different topics that you can also drill down. Elementary, postsecondary education -- you can drill down on more variables that they have. We have everything from historically black colleges and universities, faculty salaries. I'm going to go to enrollment of racial/ethnic groups. And I'm going to click on this table, total fall enrollment with post educational institutions. I like to highlight. If we look at the Hispanic variable, on one side it is enrolled by the thousands and on the right it is by percentage. You can see the percentage in 1976 for Hispanic students was only 3.6%. We get to 2021, it has increased to 20.6%. That is nice to see. You can download this data in an Excel document as well. This will help researchers and doing their own analysis, and then they can drill down to other variables of undergraduate and then on the bottom it has a nice description of the source. Which I always like to look at the source. This is the main resource for education. That is it for my presentation. I am going to stop sharing my slides. And then go to -- there you go. Thank you so much.

Excellent presentation. So far, we have one question. I will just say it out loud. In the meantime, you any questions for Eimmy, put it into the chat. Have you had any patrons using A.I. if so, any general advice for patrons wanting to use A.I. for their social side of research?

I have not. We've been talking about A.I. at my university and how to look for it when students turn in their papers. No, I haven't. It is something that I definitely want to research further because it is becoming such a big topic.

I will give folks a few more minutes. Is there something else that you would like to elaborate or show another demonstration during your talk that you want to just backtrack and shows a bit more elaborate more?

Not really. It can be hard looking for science and social data. There is so many different resources, you know, feel free to email me with any further questions. I don't have really any more.

We do have a new question coming in do you have any comments on how to effectively or ineffectively, legislators you statistics they request in their policymaking?

That is an interesting question. [Indiscernible] that is why I like to say to always look at various sources. Anything I see, no matter where it comes from, in terms of politics, even legislators, I always like to double check the sources. As U.S. citizens, we need to actively do that as well. Check sources. I don't really have any comments on that. I have not thought about it as much. Anything that I see in the newspaper, wait a minute. I always like to double check what the actual numbers are, if I'm interested in that topic.

I will start wrapping up this presentation, which was amazing and insightful. Thank you, Eimmy. This webinar has been recorded and you'll be notified when it is available to view. We are putting into the chat, a link to a survey about this webinar. If you enjoyed today's webinar, please check out some of our other webinars. We will put two links into the chat. The first is that the Academy and the second is a calendar of events. Thank you, Eimmy, for a great presentation. Have a marvelous day, everybody.

Thank you.