This is a web check. We will get started in about five minutes.
>> Good afternoon welcome to the Academy webinar web scraping for decoding challenged. This is the outreach library I am here with Melissa Fairfield will do the tech support, and are presenter Carl Olson who is the government information coordinator from Maryland. Before I turn the microphone over to Carl, I will walk you through some housekeeping reminders, if you have any questions or comments on the presentation feel free to chat them in the chat box, this can be found in the bottom right-hand corner of your screen. I will keep track of all the questions that come in, and at the end of the presentation Carl will respond to each of those, we are recording today's session I will email the link for the recording to everyone who has registered for this webinar. The webinar will be available in our archive along with a PDF of the slide deck. The webinar archive can be found on the website under the Academy. We will also be sending a certificate of participation using the email you used to register for today's webinar. If anyone else needs additional certificates because multiple people are watching the webinar with you, please email outreach at GPO.gov using the title of today's webinar along with the names and email addresses of those needing certificates. If you need to zoom in on a slide been shown by a presenter, you can click on the full-screen button in the bottom left-hand side of your screen, to exit the full-screen mode, you can see the chat box and talk questions, use the blue bar over the top of the screen so it expands, click on the blue return button to get back to default view. Finally at the end of the session we will be sharing a webinar satisfaction survey with you, we will let you know when the survey is available, and that will appear in the chat box, we appreciate the feedback after the session is over today. Now I will turn this over to Carl, who will take it from here.
>> Okay, thank you very much. Hello, I appreciate your attendance. I have worked with government publications about two decades now, about one year ago, here at Townsend University, a visiting journalism professor Todd an introduction to the emerging field of data journalism, including basic skills for getting statistical data while working under a deadline. I did a poster on this topic for the 2018 federal depository library conference, and I have expanded on my poster for today's webinar. My goals today our first I will introduce data scraping what is it and why it is useful I will demonstrate challenges and introduce the range of applications for harvesting selected data from different kinds of sites. My star attraction is a data journalism technique using Google sheets as a web scraping tool, for HTML pages. Then I will finish with some useful resources to explore data scraping further. So what is data scraping? Formally, it is a typically automated process which transfers content from online documents to an interactive format such as Excel or CSV which is common separated values, or analysis aggregation or computation. The idea is to gather data fast with fewer errors than users could keying data in by hand.
>> Data scraping is probably as old as the web itself, every search engine scrapes websites to create his directories, we might call web scraping content harvesting light, we are scraping selected data instead of entire sites. Tools to do this are now in reach of business analysts, journalists, academic researchers, and many other coding challenge professionals on a deadline. So what kind of things do people scrape? Businesses monitor competitors, journalists cover businesses and government, so they scraped directories, employment listings, competitors products, and pricing. Web addresses, site maps, reports, data tables, and of course documents. Basically anything updated frequently that people need to analyze. And I believe librarians benefit from data journalism to keep up with rapidly changing information. The happy thing about documents is that the need to download XO files already provided, obviously federal agencies primary mission is dedicated to statistics do this very well. A related option is to use advanced Google features such as cycle and DOB limits to.gov domains, I have this at the bottom, and file type limits to spreadsheets, and the user can add keywords to find rather obscure data. Each one of these websites that I have listed here and many more decides, go out of their way to make the data easy to find. This is one example of a document that links to an Excel version right here in the

center of the page. This is table 20 from the FBI 2017 crime in the United States. It has murders by type of weapons for the 50 states and Guam. Once loaded, users can interview the data, the way that they might interview a person, to try to get the story behind the numbers. This table does not have any totals, but they obviously expect their users to go straight to the Excel because it takes 30 seconds to auto some all of the rows and columns. Users may rank the states according to columns, on a separate sheet, you can see that I have broken this out for 2017, we see that nonuse, blunt instruments or hands or feet are deadlier than rifles and shotguns combined. The deadliest of all or head guns and I wonder what the story is with unknown firearms, just having one table can raise multiple questions. What if it is not online, in XL or CSV? Many agencies do not offer Excel versions, this can happen for several reasons, older documents may not be converted, state practices can vary the compliance and eggs enforcement can vary, Wyoming vital statistics has decided PDF answers their needs but they might do Excel spreadsheets on request. Smaller departments, subsections, offices, councils, commissions or contractors may have lots of data also but not make it interactive. Thus if the hardest way to copy data is to type it in by hand like a medieval scribe, the second hardest is to highlight text or tables and then copy and paste the content onto a spreadsheet. Small and simple jobs do okay, I tried it doing this with this table from data.census.gov, as you can see on the right, Excel got the first row into the first column and then everything else disappeared. This is a problem, when you have dynamically generated webpages, which are assembled when you click on the link to the. Sometimes software such as Adobe Acrobat may export content to spreadsheets in the case of Adobe, only Adobe acrobat pro 10.0 or higher has this function, the more common Adobe reader just gives you a grayed out button. Gittin Adobe Acrobat Pro just to have this function is probably not really practical. Can anyone else do this? PDF tables, converts PDF documents to Excel with a special focus on scraping tables, is it quick? Yes. Is it easy to use? Yes. Is a free? No. Is the least accurate? That depends. This is the federal government salary table for the mid-Atlantic region Virginia through Pennsylvania, it is a single page document with one large table and not much accompanying text. So the user would save the document, and upload the document, and select convert a PDF, or they would convert to PDF and then save and upload, the site digest your copy and print out this preview screen, the table format is very nice, as you can see, the green button marks download as Excel, it does just what it says, and you get a very pretty output in XL, admit and will to quick edits or cosmetic changes or doing ratios, or even data visualization. On the other hand, PDF tables did this sort of thing, with a scholarly article of 25 pages. You can see the title on the front matter of the table of contents are scattered, all the way across the page, the tables are in there, you can see one of them right here, it is not in horrible shape, but it is wise to remember the site is a PDF converter, not precisely a table scraper. What does it cost? PDF tables gives you a free test up to 50 pages and then users register for another 50 pages, and then users have to buy in batches starting at $30 for 1000 pages. It is not so much a tool as it is a commitment. We're in mind, users are buying PDF tables or pages to convert so if you have a 300 page Senate hearing in may have 50 wonderful tables, and it may convert them passably well, but there is still a lot of text to convert to throw out, Miss Manners says do not pay for more data than you need. Alternatively, your library text can scrape dynamic pages such as Amazon, census, or monster jobs, such programs have a longer learning curve, to put it mildly, web scraper.I owe is an extension developer tool, your tech tiger will create the scripts, the site map specific to each site structure, you need to be aware of how things are structured on the target site, and then they create a site map with the URL and then it tells it what element to scrape, name and address and types of file, I have tried doing some tutorials online, I've yet to get very far with it. I have listed the number of other scraping programs, that could be useful for libraries that want to do high-volume projects, this for example used to be a Firefox extension, now it is shareware, if your library goes this route, choose the software that is safest, most reliable, and has the shortest learning curve and best fits your library's workflow. Input from your library information technology section will be essential. Data journalists favored tabby lie it is free, and it's great PDF tables without converting an entire file.

Step one, is to make sure job is up to date, Tabula opens in a browser but it is not a browser extension, if the user downloads, you might need system administrator to download it and maintain access to it. Currently I firewall is refusing to open it, so it is a good idea to test it regularly. Once open, save a local copy of the PDF documents, in this case I have uploaded national vital statistics reports volume 67 issue seven, from the centers for disease control. As you see below here, this keeps a list of the documents they scrape, so it is good to keep them where you got them, once loaded, you select import, and weight to upload. They display the documents in the familiar format for browsers, with a full view of the document at the sidebar for scrolling pages. This was limited to scraping one table at a time, as I did my poster, the newest version will auto detect and highlight what it believes is data, in practice it works about as well as AutoCorrect on your phone, so sometimes it is great and sometimes it is annoying. Its main limitation is the different layout options, they can be confusing, you can see what you get on the screen it has highlighted four paragraphs of text and it evidently took this for a table, this is easy to cook off, if there are other formatting problems, the borders adjust pretty well. Overall this is a huge improvement over scraping an entire document. Especially with those Senate hearings.
>> Tabula can get annoying when a table lies parallel to text, as I said it is pretty good about reshaping sections that have gone askew, and it is not necessarily a bad thing if it puts something like this, like the title in a separate scraping, because as you will see, sometimes Tabula gets confused if the rose and columns are in an odd configuration.
>> So like PDF tables, the user creates a preview page, for larger documents the response time can take you back to about 1997, the preview gives you all of the selected tables on a single sheet, most columns maintain separate columns here, for black, white and Hispanics and is separated total male and female, as you can see over here with all races and origins, and over here with non-Hispanic white, and non-Hispanic black, it put all of the lower columns the lower sections if put them all in a single column. So there is a way to go back to revise selections, and double check that, if at all possible, or you can go ahead to export. And this can actually be good, Excel has a way, it does print pretty much what it previewed, so it is reliable that way, there is a workaround, for when you have the problem, with a series of tables, you would highlight all of the data, and insert two columns, and then you tab data then text two columns, and then select space separated and it should populate like separate columns with your data.
>> And then you have to page down and repeat it. Now because office applications X so easily to HTML, many tables are still static pages. Lena Kroger a data journalists for nonprofit called [Indiscernible] has created a Google sheet hack, for populating a Google sheet with the contents of a static HTML table. By using a single spreadsheet formula. Step one is to gather the following information, the URL of the tables page, specify the element site, in this case a table, and then Mark for the formulation start scraping, in this case zero in order to start at the top. In the first cell, this is the formula that you type in in the first cell of a blank Google sheet, you import HTML, all caps, and then you put the data elements in quotation marks, separated by commas, no spaces, and the). If you get a pound signed and/A, that is the error message, but you can always type that over, and when it does work, and it did work very nicely for me, it is oddly satisfying to see it appear in the scroll. And just unroll the data to each cell. This method can get tables into a sheet, that users can export into XL. Lena Kroger of Pro public a has the original tutorial on her page, I put the link here, so it will be available on the slide deck. And she has other data tables to practice with, some of her links are out of date, she does have other resources to explore in the field of data journalism. For further information Paul Bradshaw directs a Masters program in online journalism at Bloomington city journalism, in the UK, is the clue good closest thing to a guru on the subject he runs a blog online journalism and he literally wrote the books for scraping for journalist and a general primer called data journalism heist. Both of Purcell as e-books. I realize that that was rather quick, I'm sorry if I went too fast, that is all that I have for today, I hope this gets your pause wet

on data scraping, if any of you have any questions I will do my best to answer. Thank you all very much for your kind attention.
>> Okay, if you would have any questions on the presentation, please go ahead and chat the men I will share them with Carl.
>> Citations from an HTML page, I think you might have to experiment there with the data elements types. That is not something that I provide myself. I am not familiar with [Indiscernible]. But I will have to try that. For data initiatives, I would start with something simple like Lena Kroger's exercise, if nothing else it can give you the basic outline of the concept, and so on. We do have Google sheets, and that sort of thing here, we do not have any software or utilities, in our library.
>> Somebody asks the questions about the slides, they will be available with the recording, but it probably will not be up until sometime next week after the holiday. Are there any other questions? I guess I would recommend Tabula, that has been the most accessible to me, again, there are problems with firewalls and that sort of thing, but I have tried it on my home computer, and it seems to work all right, so just getting familiar with that, it is mostly editing that takes the most time what it selects as a table. Okay everyone, Melissa is putting the survey out, we would appreciate it if you would take a few minutes and fill that out for us. Carl, there was one more question, do you make the scraping material available to the and additions? Or is this just on a personal drive. Right now I just have a personal drive, but I was thinking, if we did have something like this for journalists I am looking into possibly doing library instruction, we have a communications department and that sort of thing. This is the last call for questions. If anybody has anyone, then we will get this webinar wrapped up for today.
>> Okay, it looks like we are done, thanks to everyone who attended, a very special thanks to Carl for sharing information with us. Hopefully we will see you in another presentation very soon.
>> Okay, Inc. you all for coming.
>> Thank you all for coming.

>>