The What, Where and How: A Discussion on Digital Humanities and Government Information - Transcript

Hello and welcome to the what, where and how, a discussion on digital humanities and government information. Before we get started I have a few housekeeping reminders. Please use the chat box for questions, comments and technical issues. We will keep track of all questions and address those at the end of the presentation. We are recording the conference and all registrants will receive links to the recording after the event. Please join me in welcoming our presenters today Charmaine, Catherine, Lucia and Jesse. I will hand over the Mike to Charmaine to start the section . Thanks.

Thank you so much for joining us today I know it is one of the last sessions on the last day so I really appreciate you guys coming to listen to our presentation. My name is Charmaine Henriques and I am the librarian at Indiana University. Today and very fortunate because I'm going to be presenting along with my wonderful colleagues, Jesse Silva , at the University of California libraries. Catherine Morse, the librarian at the University of Michigan libraries, and Lucia Orlando , social sciences and government information at the University of California libraries. -- The fact is because of the unique, diverse materials that are in the collection, librarians are usually the first to know and some of the sources -- digital humanities methodology. Will start out by giving a short overview of digital humanities, Jesse will share some of his experience , and Catherine will talk about the tools and Lucia our final speaker will cover -- after our presentation I hope there will be enough time for people to discuss and throw out their ideas about the digital communities and government information. Let's begin. Hopefully this will work out. The way I want it to. Yes I did share my desktop, okay great. Are. So, the first question usually is what is digital humanities? And the reality of the situation is that there really isn't a true definition. It has been debated over what is digital humanities. And in 2009 they started -- today digital humanists around the world document what they do in one day. In 2014 they asked the question how did you find digital humanities? And there are about 1817 definitions. So there really isn't a true -- and you keep reloading and you get a new definition. So there really is someone. But as librarians we love our acronyms, we want to classify things, we need a definition. So the one that I am going to give you is the working definition of Indiana University's Institute for humanities, from their September 6, 2019 presentation why use digital methods in arts and humanities research? So there definition is digital humanities is how technology and techniques with the digital intersect with humanities scholarship inquiries and follow communications. It applies to the arts and -- analog work hand-in-hand. So the second question is usually why use the library when there are many other avenues? First and foremost, for the obvious reasons, library serves as information portals. They will have collections and -- additionally the library has experience with the research process and is known as an innovator because they build digital collections and analyze collections and is an expert on data collecting. We collect data on everything, accounts, collections, reference. So for all of these reasons the library has this technology. Sometimes one or the other or both in digital humanities. However remember that digital humanities is not just about humanities content, it's also about what we do with the content. So continuing with the what, let's look at what does digital humanities look like and what kind of research can be done using it to Mark first step is -- top of mind uses an algorithm highlighting trends and reveals patterns in literature text sets that folks are able to identify resulting in syntax analysis and word class. So what you are looking at, what you are looking at is a word cloud from the 2011 city address a former President Barack Obama. So the words that are in the same color are categorized together, and also bigger words, so you have new, people, work, now, jobs, American, and this is not surprising when we were just coming out of a recession. So the focus or where his priority should have been was building the economy and the impact on jobs and so forth. What you could potentially do is do the same analysis on all of the data for every year and you would see what the priorities were throughout the presidency for both terms. To take it a step further,, newspaper articles, and -- and find out what the nation -- the next method we will be looking at is mapping. So mapping shares the aspects of one's work. It is a visualization that attempts to represent -- so here, okay, so here we have the Cuban battlefield humanities project for the University of Nebraska Lincoln, which documents the battles, the battlefields of the Spanish-American war, or the war for Cubans independence in Cuba. So there are many parts to this project that we are just going to show this one right here. So right here we have selected commanding officers reports, that comes from the annual report of the war Department, and the commanding Army. Over here you have the operations of the 12th infantry, and this one has the digital map and the dates. The analog. See you can get the movements from the company a through D from 6:00 a.m. until 4:00 p.m. is essentially the timeline. And then you can get in depth and more sections in detail. And then the map of the Santiago campaign, and over here I should just say that the reddish orange links have information of different battlefields. With images. Okay, so this is a satellite map, and also a map from 1898, it's the Army Corps of Engineers. So as you can see, commentary is not on, I will put that on. But the commentary -- okay, there we go. So the yellows are points of interest in the blues are block houses, so here is the information that goes back to those red supporting links about the battlefield. You can also look at the 1968 satellite map. -- Timeline for 1898, and you can actually build and keep going down to see what the area looks like. Obviously seven years later from the original map. Okay, the next method we will be looking at is making. So making is kind of the hard part, it is one of the hard things on digital humanities. And it depends on the amount of -- in the collection. So maintenances creating and re-creating imaging -- investigations. So what you are now looking at is a cast of -- right hand. And it was made by a sculptor in Chicago. And the object he has in his hand is a piece of a broom handle. This is a two-dimensional picture of his hand. Okay, and so they scanned it, and here is the three-dimensional object that you can actually move around. And diagnose. For those of you that have elementary aged kids, you can set them up with this and have them move this thing for hours. So that is the three-dimensional image. And then this is the actual object that was printed, and the texture and color have been duplicated, but you can actually get the actual object. So the question is what is the point for all of this? So moving the object through a 2-D platform to the 3-D and to the actual object, essentially what it does make you do is look at the object in a new way that you never thought of it. Which is the actual interrogation, and with that perspective it sparks new ideas for research. So what the Library of Congress wanted to do was then provide 3-D image and the files so they could print the actual object. Does that make sense? Before I hand this over to Jesse I will say that -- before as in the above, there are others like investigational analysis, -- which will all be shown later on, and what we won't be speaking about -- which there are plenty of examples of at the K-12 levels. So I expect not to talk about that. But one of the things that we did discuss when we were getting together was the -- digital humanities to the point of not really mentioning the social sciences, which in a way the science guidelines and government information, I'm not going to argue the points that the social sciences should be part of digital humanities, but that does hold humanities materials from the agencies like the National Park Service, the Library of Congress, -- and in fact -- Jesse will be talking about enhancing those abilities. Overwhelmingly the questions are social sciences in the collections, and all of digital humanities methodologies we will talk about today have very practical applications with the social sciences. In the digital humanities has the potential -- to create a demand for, and government information collection from -- to cool kid on the block. So now I will stop sharing, and I will pass the ball to, and I can't see Jesse's name. I will pass the ball to Jesse who is going to talk about outreach.

Hi everyone, thank you Charmaine for the overview of digital humanities. I will talk a bit about three projects I've worked with over the years that I have learned and what I've learned in doing outreach to support not only digital humanities but also anything that it can play a vital role in. Full disclaimer I fully consider myself a novice at much of these tools and techniques that digital humanities use. The first time I heard the phrase digital humanities was in 2009 when the literature grad student made an appointment with me for project she wanted to do on-topic modeling. The topic modeling analysis. The project was looking at the words and concepts of how terror and terrorism evolved over the last 60+ years but she did not know where to begin. Graduate student met with me after she had met with our than literature librarian who told her that what she wanted to do was not literature or humanities work. At the time this was not an uncommon attitude as their were some disagreements on what digital humanities actually is as Charmaine talked about. The grad student made an appointment with me and when I saw she wasn't literature was a bit nervous about the meeting. Literature was not an area I worked in. We met and talked about what she wanted to do. After hearing more about the project I suggested she used presidential speeches as they were finally, corporeal text that we could get to quickly and easily. Showed her the American presidency Project, a project out of UC Santa Barbara. She was thrilled with this and used screen saving techniques to build her corporeal and was able to do the analysis for her paper. The second project I was going to talk about was with an art history student who wanted to look at how presidential portraits had evolved over time. This was a threefold project. First was how aspects of the official Orchard evolved with reference to things like lighting, background, et cetera. The second was how the official presidential photograph evolved, in a similar manner. The final part of the analysis was to look at similarities and differences between the portrait in the photograph, and determine if they mirrored any already documented changes in the medium. After hearing more about the project I showed him the portrait site and he was able to do his comparison using software that measured contrasting images and all other kinds of things. The final project I want to talk about is a digital media undergraduate thesis project. First a bit of background. I worked with one of our humanities librarians to teach a class on sources for DH data. This led us to working with a group of undergraduates doing their PCs projects. One project from this group I found really cool. In a previous project this student found a couple of old photographs of a street in San Francisco lined with shops, civic clubs and other businesses. I think the pictures were from the 1940s or 50s but I can't remember. Using phone but data they scanned themselves, and were able to identify the shops and other organizations. But they wanted more information on the demographics in the area, including anything that would add details to the story they could tell with their project. After hearing details about the project in the initial meeting, I worked with them to find census data, County business patterns, and other data they could incorporate. We then looked at how to download the data from various free sites like national historic GIS and and they worked with other data -- the final project was interesting. Imagine looking at a small Google map and then using Street view to take a virtual walk down three to four blocks of a street discovering facts about the businesses and other organizations in the neighborhood you were virtually touring. It was a really cool project. So what I'm going to talk about next are some of the things that I've learned based on the work I've done in this area. And it may help you get started in some of these areas. When working with students and faculty in digital humanities it is okay not to be a digital humanities expert. And I found people working in these areas are really trying to develop a community of learning that welcomes questions, from and training opportunities for all levels. And humanist researchers love librarians they really do. But they may not always, they may not have thought to reach out to government information specialists, as government information and librarian professionals we have a great deal of knowledge of the content available, much of it free, and that is something free in academia can be a major bonus for these projects. If you don't know the tools but want to learn just ask those working in the area. In my experience everyone has seemed very welcoming of librarians and those willing to join. Second thing is talk with your humanist colleagues. Government information seems to be purely associated with social -- in the sciences. But there is a need for this information in digital humanities work. So I will say it again come and talk to your humanist colleagues. There may be a digital humanist group on your campus or class, or professor or even a graduate student who is interested in this area. You can probably find out by talking to your campus colleagues. I did this with one of our humanists, one of my humanist colleagues and was invited to talk to our campus computational text analysis working group. Group made up of people working in digital humanities type work. And this led the library at Berkeley to acquire the -- to support the research. And this group has been making use of this for the last three years. Working together with your humanist colleagues can help both of you best serve your academic community. Also talk with your data colleagues as well. And in many cases government info and librarianship go hand in hand. But if you are new to digital humanities your data colleagues may have experience with some of the tools. Are Python and other tools can be useful in both data analysis and digital humanities. Check out some of the tools that Catherine will be mentioning. Many of the tools are really fun to play around with and you can really up your technical skills by experimenting with them. And the last suggestion I have is to have fun. The projects I've mentioned are just some of the ones I've worked with in digital humanities over the years. Don't consider myself an expert in this area, but I have gotten some exposure to what DH researchers are doing and I have learned a lot. And I really have had a lot of fun trying to help the students and researchers who are pushing the boundaries of scholarship and creative activity. Who doesn't want to have fun at their job? And with that I'm going to turn it over to Catherine to talk about tools.

Thank you Jesse, I want to have fun in my job. I am Catherine, my pronouns are she hers, and like Jesse and I not describe myself as a digital humanities expert, but one thing I do a lot in my job is helping people find data including textual data, so I would like to keep up with the tools that researchers are using to visualize and present their data. So I've got some tools that you can explore. We will look at some text analysis, some data visualization, and some digital mapping tools. The first one we are going to look at is the text analysis tool, see you can bring your own corpus into it and visualize things in different ways. The example we are looking at now is from the public papers of the president and it is interesting to compare on the left the word cloud here, with the word cloud that Charmaine used in her example that was from one of his speeches. But for the whole public papers of presidents from 2014 you can see like President Obama like to say the word people. In the center, you can create what is called terms -- which is similar to a word cloud and shows word frequency. But you are also looking at cooccurrence, so we can look at what words occur next to each other. And then the visualization over on the right is a trend visualization. And the trend visualization is a line graph that is showing the word distribution throughout the corpus. So buoyant is open source, you can bring text and a lot of different formats into it. And you can use their other visualizations, these are just three, they have a whole lot. Another example of text analysis is the case law access project. They digitized case law from the Harvard Law school library, and applied a text analysis tool to show word frequency use over time. So in this example we are looking at here, we are looking at how often the terms wife, husband, or spouse appeared in case law, and we can see that the term wife appears more than husband, or spouse. Consistently from 1800 to 2017. So this is similar to the Google Ngram viewer if you have the ever used that, this is a similar kind of frequency trend for the Google books corpus. Our next visualization is of the U.S. treaties Explorer. This was a congressional data high school project where high school students took data from congress.gov, they converted it to JSON and then they used Python to create the B swarm plot. This one is really interesting I think because it gives you a sense really quickly of the subject matter of treaties and the time. As well. Next up we have an example of a network analysis, so network analyses are helpful for visualizing connections in data. So this example we are looking at here is from OpenSecrets.org and it is lobbying clients who have lobbied the FTC on net neutrality and their PAC contributions to members of the 115th Congress. So in my job I supply people with data, especially on things like campaign contributions, and you could supply a spreadsheet with that kind of data, and then your researcher can use the tool like Gephi, like Cytoscape, or R to visualize. The researcher would identify nodes and edges and use those tools to visualize the network. A data visualization tool that offers a lot of different options is RawGraphs. So this is an open source data visualization tool and you can use it to make Gantt charts, bump charts, or tree maps. The example we are looking at here is a bump chart. A bump chart is good for showing change over time. This is another tool where you can bring the data and choose what kind of visualization you want to use. And it is pretty fun to decide which visualization you like. And you can see really easily how different kinds of visualizations emphasize different aspects of your data. They also have data there already, so if you want to go and play around with it it is fun and who doesn't want to have fun? So the example we are looking at here is the 10 most populous U.S. cities every decade since 1790. I think this example is really fun. Okay, making maps is a big part of digital humanities work. And as a government information library and you may be called upon to help users find maps, to find satellite images, to find data that will be used in maps. So it is good to know some of the tools that are out there. One of them is Tableau Public, so Tableau Public is a data visualization tool that does require a subscription. But Tableau Public is free. You can bring data in a lot of different formats there. Text, Excel, Google sheets. And then you can visualize it. So what we are looking at here is the economic output in U.S. counties. And then our last mapping tool here is QGIS , GIS, geographic information systems have been used for a long time. They are very robust tools that help you create maps and do geographic spatial analysis. The most commonly used tool is called ARCMAP and it is proprietary. QGIS is open source and it is really gaining in popularity. You can do all the same things in QGIS you can do in ARCMAP, maybe I shouldn't say all, many of the same things you can do in QGIS you can do and ARCMAP. And the example we have got here is looking at rural housing age, from the 1940s, and from IPUMS and NH GIS. So those are some of the tools that we can use, and we can take our data, and our text, and create new visualizations with these tools. I am going to pass it on to Lucia who's going to help us by telling us more about where we can find data and text.

Great, thank you Catherine. Hi everybody I am Lucia and I am the social finances and government information at the University of California Santa Cruz. And like everyone else here, I also consider myself a novice when it comes to digital humanities. But one of the things that made it easier for me, I guess you would say, and librarians, is actually finding some things. So I will briefly cover some free data sources and provide some simple tips for finding free data on your own and then I will talk about some elements to communicate to vendors if you find yourself like Charmaine mentioned in the position of being the first to be asked in the last to know. Hopefully these resources will help you through the initial deer in the headlights moment you may have. I remember having that I wish I had something like this. So right here on the slide you can see some examples of textual data sources that can be used freely. I try to pick ones that were a little bit different from the ones that Catherine had already showed you. I will not cover these in great depth but I wanted to give you a sense of what the possibilities are here. I think one of the most popular ones that I have seen used is the chronicling America site, from the Library of Congress. It is a fantastic site of newspapers and other items and is just a good place to get started. So for some of these areas you might think about them and say okay maybe you want to start one of the tools that Catherine mentioned , and of course you want to find some free data to put into those tools, you just need the sources. I do need to backup for just a second and mention the definition of a word you might hear a lot of and that is corporeal the first time I was ever asked about government information is someone who said I am looking for a corpora on XYZ in government, produced by the government, and I was like what is a corpora? And basically a corpora is a collection, it literally means body, so it is the body of whatever it is you are working on. It could be one document, or a can be a collection of documents, or a collection of texts. Essentially it is the collection of whatever it is you happen to be analyzing. So thank you. Plural of corpus. So some of the data that I have mentioned here comes from government sources, and some of it has actually been transformed for you, like the American presidency Project from the University of California Santa Barbara actually took government documents and created a subset. This GovTrack data also has a wealth of information sources of government documents. In the last two you see on the list, the Hathi trust and the ICPSR, they are mostly free or selectively free. What I mean about that is part of the collections are available to anyone. For happy trust -- they still have a subset of their collections that you can use without being an actual member of Hathi trust. As long as you have a -- from a nonprofit institution of higher education you can register and then you do not have to be a Hathi trust member. And I just wanted to be sure people are aware of that, because it is kind of cool to see what else is out there. And not everything has to come straight from a government source to be usable. So moving on, how do you go about finding more data? So here are some ways to find data on your own. And one really common way is to se e what other people have done before you. So a search that I like to do is to just enter the word Libguides and enter some words I supplied here looking for data. You can see what other institutions have done for example the Berkeley library has fulltext of data, the images cut off a little bit, but there is the link here if you want to see what else is available in there. The University of Virginia library has a digital humanities research guide that also includes government information. But wait, there is more. So the Google data set search is another way to find things deceased to be in data, they just brought it out rather recently. It is a little bit clunky but I really like some of the information that they give you right on top. For example try a broa d search first, that is something you want to use broad terms like open data, or data portal. For example here is a link to legislative districts in California. There is a link to a California data portal. But what is really helpful and I know it is hard to see across the top of the slide here is that it gives you options for mirroring your data set, and finding information that is helpful that we. For example what I mean by mirroring is you can narrow down the usage rates, so it is commercial or noncommercial. It has a download format so if you are looking for text you can limit by that. And then depending on the search terms you use, it may also have some topic areas, like the economy and things like that. Another source that is worth looking at enlists th at are available on GitHub, so GitHub is a great repository of data , and I included this one here because they basically have done a huge search for data, lists of data that have been collected and tidied up from different questions that have been asked and answered on GitHub . And they have this in the awesome public data sets list. So one thing on GitHub, it says awesome public data sets, it is not that everything on there is free, so you definitely want to be sure to click all the way through to the data set to make sure it is free. Next I want to draw your attention to vendors. It is helpful to keep vendors in mind a lot of vendor severity digitized information they support text and data analysis oftentimes this comes in we can see it associated with it. Definitely bear that in mind one source I want to emphasize here is this JSTOR data for research. They have two types of data sets that they make available one is free so anyone can download it. You have to register for a free account but you don't need to be a subscriber of JSTOR , but they want you to have an account. So they have prepackaged data sets, and then they have another set of data that is also free but you do have to sign a user agreement before you download that and usually they will pull that together for you. So, moving on, lastly I thought it would be helpful based on again my having this experience of a deer in headlights moment when someone asks you these questions and maybe you need to get a vendor involve, or maybe the patron is going to do a screen scrape of a government website of the state or county or something like that, I thought these were some good tips that I wished I had had before him. Although I have to say the very first one, this one I think everyone here knows about. But it is helpful to put it out there for the patrons and that is text mining is not automatically included with the database access. And the librarians need to check the licensing and see if it is available in there. I'm sure we've all had instances where Cindy downloaded a whole bunch of stuff from a vendor's database, and -- by the vendor. But just keep that in mind. Especially when you are doing orientations for graduates and faculty. But some other details to convey to get from your patron and convey to your vendor, or convey to the website administrator is to find out what the scope and type of information you may need, do they need the entire database? Or are they doing a specific search, they have a run of PDFs, and they need that specific purpose, just to get an idea of what that is. And then how are they going to access that data? Because some ways of accessing data can -- screen scraping, or doing a download on a government website. So it is usually there to try and get a hold of the site administrator first so they can be prepared on their servers when they see that kind of traffic coming through. And then if the usage is going to be a bulk download, then you definitely are going to need to contact the vendor or a site administrator, usually when you start it incurs a fee, so be prepared for that. And you will need to let your patron know that too. Also let them know how frequently you are going to have to do this, is there a time. They are trying to come up with or just a snap shot in time? And then here is the big one to really talk to your researcher about, and that is what their timeline is. Because he can take a long time to get this data. Especially if you have to go back and forth with the vendor, or there is an intermediary process or provisions they have to maintain. It is good to have an idea of what their timeline is and help manage their expectations. On what is going on with that. So that is all I have to say, thank you for listening. I am going to go ahead and him this over to Charmaine to moderate, thanks everybody.

So, thank you all for coming, and being patient with our presentation, I know it is the very last one for today. And I would also like to -- why am I not able to move the slide -- this is a time for discussion and if I was able to move the slide, what you would see because you guys will get copies of the slides, is that some of the projects, not all of them, but some of what we put down as examples, -- that is also part of the side. So, does anybody have any questions? Statements? Anything they would like to add? This is the time.

If you have any questions for the presenters please put them in the chat box at the bottom of the screen and make sure you are sending your questions to all participants, thank you. So there is a question we are beginning digital humanities on our campus, what tools would you recommend for creating interactive maps?

[Captioners Transitioning]

That's a really interesting use case. Unfortunate because I have geospatial data librarian colleagues and my institutions I don't have to know too much about it. One thing I notice is trying to get a sense of how much time they have. To invest in many different tools. I think something like GIS, you could spend a great deal of time learning. It's a very robust tool. Finding a lighter weight tool for mapping. Depending on the students timeframe. Like maybe a story map.

The package going back to the requester and asking them down to partner with you. And see if the researcher has access to grants other types of sources from departments or other pockets.

Is typically a partnership. It would be very expensive very quickly. We set up a separate fund available for data purchases so it didn't come from one person's fund it's a fund researchers would apply for to get access to data sets and she would investigative this can be purchased. In some cases she could purchase a data set because license was agreeable. In other cases, it was really restricted.

If I have to say, it really makes government information more attractive if it's something that will suit your particular researcher. And look around and see what others have made available or what kind of government websites have data.

The way it's structured, we would need funds, if you have to hire include the data you would supply the funds. That will be on the way to think about finding money to were together. If you have more questions, please put them in the screen to all participants in the chat box. I'm not singing more questions I want to thank our presenters and thank you all for participating in the virtual conference. Up next, we have DLC wrap-up and closing session. This'll be the only room that will be operating. For now, we will take a short rake and pick back up at 3:30 PM Eastern.
