>> SPEAKER: Hello. This is Laurie Allen. I'm going to get started with saving data, lessons from data refuge with librarians. I work at the University of Pennsylvania Libraries, and I'm really happy to be part of this series. ALCTS is not normally my home at professional community, but I've watched it from next door, as Norm had the office next door to mine for a lot of years, and I learned a huge amount from him, so I'm really happy to be part of this community that I have seen from next door for so many years. I'm going to talk about this project, data refuge, that I've been working on for the last six months or so, and I was a little bit ambitious with this presentation, trying to describe the lessons I've learned from librarians, when I still feel a little bit too deep in it to really be getting to lessons, but I'll do the best that I can. Here we go. So, first, I'll say this is a talk that I'm going to give really from my own perspective, my own experience through it, but the data refuge project as a whole is impossible to conceive of without understanding the deep, deep collaborative nature of this project, just how many libraries and people, individuals, organizations, institutions have been involved, and individuals, really thousands of people have made data refuge, so I'm going to try to sort of bring together some of the lessons I've learned, but the major lesson I've learned is what an amazing, amazing community of people can kind of work together to do. So, I want to just first call attention to the collaborative nature of this work. A little bit of history of the data refuge project, um, in late November, graduate students in a program in environmental humanities here at Penn were concerned about risks to the ongoing availability of federal environmental and climate data, and they decided to do something about it, and, um, they, together with their, the director of the program in environmental humanities, decided to have an event that would save environmental and climate data.

They came to the library, and we helped and collaborated with them in this project. Together, we planned an event that, um, with other people in Philadelphia, with other librarians and institutions, um, an event for late January, or actually, sorry, mid-January, just before the inauguration here in Philadelphia, and that event included time for storytelling, we tried as hard as we could to bring together every kind of, um, expertise, so deep technical expertise, there were people who are programmers and developers, storytelling, as I said, people who could understand the way that the data, um, and describe the way that data produced by the federal government that relates to the climate and environment is really closely tied to people's lives in the city of Philadelphia and else where, as well as, um, people who are responsible for describing the data, for safeguarding it, technically and in other ways. That project, that event was the second, there was an event in Toronto before ours, and then there have been 50 since then, or 50 including those two events with thousands of people since then to do this work of creating safe copies of federal environmental and climate data, safe meaning citeable. So, that's a really brief overview of the project. As I said, this site, what we're looking at here, this is datarefuge.org, and this is where we've stored the data that hasn't gone to Internet archives. Most of it is there, most of it is available to view publically. There is still some that's not, and there are some really giant datasets that we haven't moved in there, but many people, especially, um, as time has passed, we've been really encouraging people to build as many ways as they can to safeguard this data, understanding that this is not a problem that we can solve on our own or through a series of events alone. Before we get to how we moved forward, let me go back a little bit and talk about some of the problems that we were trying to solve at the beginning, some of the questions we were trying to answer. Um, so, before our event, in the hectic weeks between deciding to do this and

having folks come together at Penn from all over the city and people traveling from else where to do this work, we had these questions. Like, first, is there vulnerable and valuable data, right? And we talked with a huge number of scientists, some federal employees, um, a lot of people who worked in the open data community and others about what makes data vulnerable, and kind of understanding that there's data that's vulnerable for technical reasons, political was at the top of some people's minds, but certainly, there was also technical and social reasons that data might be vulnerable that are not related to political reasons.

There's also legal vulnerabilities or legal steps that make data less or more vulnerable, so we were trying to understand, you know, if we're aiming to get the most vulnerable, the most valuable data, what does it mean to be vulnerable. That's as much as I'll talk about that question unless people ask about it in chat later. Then this question of value, what's the most important stuff, so understanding we want to get the stuff that's most important and most at risk, so in terms of the valuable part of that equation, basically, we sent a survey out locally to our campus, but also circulated a survey where we asked what are you most worried about, what do you use most, what's most important to you, and used the results of that to sort of bring to the top of the list the data that we would, um, tend to first. We also, um, needed, you know, this question of the scope of the problem, what's the list, right? We wanted to get a list. How big, how much data is there? What is the list of data so we can check things off when we say we've got it or so we can assign it to someone so there's no duplication of effort? It turns out, there's no list. Um, it turns out that what constitutes data, and we're going to get into this later on, but what constitutes data is defined within communities, the meaning doesn't hold across communities, across conversations. It's very, very hard to describe, here's all the data, and then say, okay, we've got this and this. So, the scope of the problem, what I can say after spending, like, six months working on this and very little else is the scope of the problem is very, very huge, but, happily, there were lots of communities who were already working on this, people within government documents librarians, people who actually work within the government are doing a great job in many programs and areas backing up their own data, right? So, some technical vulnerabilities, we don't need to working about. There are people who are taking really good care of their data in government, of course, and, so, that, there's, you know, collaborating with them was important, and some things have already been done. Obviously, ICPSR is a great partner in the world of social science data, and I'm going through all these questions, then I'm going to leave them behind and turn to the next piece.

So, I've talked about value and vulnerability, and then how do we make sure we're getting as many people involved as we can, understanding that this problem is so big. The next set of questions, and I won't go through these in as much detail, and though they are deeply important, but I tried to make this talk kind of focused on this particular question of, like, how do we package up federal data from the web and save it, so these questions, which are deeply, deeply, deeply important, I'm going to just mention and then move on from, which are how do we make sure that the data remains citeable, that they're trustworthy and all that, and so this is work that we really drew from the archival community, as well as from highly technical people, on trying to understand where and how should we use copies. This is institutional, it has all sorts of implications, but I'll just say those are problems to be solved. Then how should we describe the copied data, and I'll spend a little bit more time on that, but the fundamental question, of course, is what is federal environmental and climate data. This universe of things, what's

inside of it. So, um, it includes satellite data, data that, you know, needs to be stored that moves around in trucks. It includes water readings that are very small, it includes data gathered from buoys, data that exists in spreadsheets, data that exists in PDFs. Um, it also, though, includes information that is used to describe the data in its home, and so it might be a website that tells, that helps a teacher teach climate change for students. That, too, is federal environmental and climate data and is worthy of saving, and, so, from a practical consideration, the way that we ended up sort of slicing and dicing the world of environmental and climate data is, basically, by how it could be saved. So, there are some kinds of data, like a basic HTML web page, that can really nicely be saved through web archiving, and so there were, um, and the environmental data and governance initiative sort of led the charge on breaking up the, um, federal web, um, universe into pieces that people at big events could save to the end of term archive, which is an ongoing project that's been going on for years, to backup these websites, so they led the charge there, but we had that going, and then you can see directory of files, the machine can handle those, those can be used in web archives, so this question of query interfaces, the stuff that web archiving can't get, and research datasets, no one wants their research dataset from the wayback machine, so we aimed to do that in data refuge.

So, focusing on query interfaces, we're talking here about something that someone would need to click search to get the data, and so this is, what we tried to do is basically take some of the contextual information from our HTML pages, pull the dataset using, basically, scripting, pull the dataset out of the query interface so that it becomes a set of files, attach a little bit of meta data, and hopefully create a research dataset, the kind of thing that could be saved. That was the goal. If you think about it though, we know a lot from the world of data management that says that we really don't want to start at the point where the data is already on the web, that's not where the work of saving starts, it really should be that you start with data and then you create both a research dataset and the database that's used to create the website. Um, that is the way we would like open data to be created, and it does, of course, raise this question, which we had to handle, which is if we want to move data from, basically, this sort of wild west of basic HTML pages and lots of information to these research datasets that can be stored and packaged and shared and moved around from site to site, whose job is that? The events were incredibly powerful for all sorts of things, and I'm really kind of calling out the stuff that was heard about them. The events were really powerful for getting a lot of people working together on this problem, but the question of whose work is it to create data such that it is storable separately, this is, this works differently, this problem needs to be solved really differently depending on how big the dataset is, how frequently it's updated, how it works, and in some cases, the web interface is the best way for the federal government to share the data. The only way it's a problem is if the data producers stop caring about the data, it stops being worth their time, whether for political reasons or any other reason, if they stop caring about the data, then this way of sharing it in a series of web pages becomes problematic, it becomes too hard to save, if we decide we want to save it,so then it becomes the work of libraries, and, of course, this, you know, the federal depository library program was, to some degree, created for just this purpose.

It is one answer to the question of whose job is it to help ensure that data continues to be available, that federal information continues to be available after the federal government who created it stops caring for it. One answer to that is that the federal depository library program was designed, really, to do just

that. It still works in lots of ways, and yet it is not, um, solving the problem of all of the federal information that's out there. There's sort of a question of the meanings of terms, like information dissemination products, which you see in that last bullet there. That has a sort of, I believe, has a kind of technical term, or publications has a particularly technical term, and it is different from the way that people in general use words like federal data. We might want them to be describing something very specific so that we can build a set of workflows around this very specific thing, but, in fact, in practice, in the press, and in conversation, people mean a huge range of things, and so the kind of needs that the FDLT is designed to meet, as important as they are, haven't expanded and grown to meet the kind of ways that people understand what federal records, like the record of federal government rather than federal records, like those records that are, um, covered by federal records laws, right, those two are not fully overlapping. I hope that that makes sense. So, that's one idea of whose job this is. Another, however, is the open data community, and it's really, um, I think one of the main things that I've learned, well, I learned so many things, but one of the really big things I've learned in this project is how, um, much room there is for libraries across the board, public, I hope, certainly, big research academic libraries, but all libraries to engage with the community of people who are interested in creating open data and how disconnected we have been. I thought, before I got involved in this project, that I was sort of onboard with them, but it turns out that their understanding of how to work with government data producers to make the data that they use and create available for people is really different from the ways that libraries have been doing it. They work in meta data, they work in access to information and creating meta data so that people can access information, and yet they don't see themselves having anything to do with libraries, and libraries don't generally see themselves as having anything to do with the open data community, so I guess part of the reason I wanted to do this talk in this way was to kind of call this out as a place where we really have, um, some incredibly awesome collaborative opportunities that I think, um, it makes sense for us to really take advantage of, because there is a huge community of people out there who are working to ensure that the data that is created as part of our governance work is available to the people, and a lot of their work is creating meta data, and yet when they do that, they don't think of it in many of the same ways as we do, so I will say there are library partnerships. Temple University is actually doing an awesome job there, but it's not as common as I wish it were.

So, um, in terms of whose job it is, maybe it's libraries who work with FDLT, maybe it's the open data community. I hope it can be some combination, because, um, I'll just spend a real quick moment pointing out that, so, I'm calling attention here to data.gov, and this is a record for, this is a record for, um, basically, data about, um, highway guidance, and you can see, if you see at the very bottom, it says downloads and resources, and then, basically, it's just linking to a website, so this whole record is created, and then it just links to a website. The website it links to is not this one, but this is, um, just an example of the kind of mess in a very, um, I'm not, I don't mean mess, like, it's a bad thing and someone's to blame, I just mean it's a mess. So, the record that I just showed wasn't actually for this dataset, but it might as well have been. This dataset is not described in data.gov, because so much is not, but here is, for instance, a dataset that I got from the wayback machine, because it was briefly missing from, um, its home at the federal highway administration, this dataset is actually the one that our, um, that the office of sustainability here in Philadelphia relies on most heavily, and they rely on it

because, it's the federal data they rely on most heavily because it helps them to make plans for, for instance, where to build roads in the event of climate change, which is happening, so, you know, what's going to flood most often, that kind of thing, and you can see in this very tiny print, they say, oh, great, this is also in PDF and Excel, and this is in, um, the way back machine, so you feel like, great, it's safe, it's there, the wayback machine has it, but, unfortunately, when you go to the machine, this is the Excel document, and it's basically just a form that then queries a live website, so the minute it goes to the wayback machine, it stops working. So, I'm calling attention to this just to say, you know, in the hopes that this, um, whatever librarians are listening are thinking, gosh, we have to do something about that, because, um, there is, really, a huge amount of work to do. I hope that the federal highway administration doesn't stop producing this data, I hope that they continue to produce it, and once again, I don't think it needs to be, there need to be political reasons why they might not, I want to just call attention to the fact that for whatever reasons they stop producing this data, at some point, the federal government changes its practices, and when it does that, the data that it produced is, right now, tremendously vulnerable, it's at really high risk. The way that the federal government keeps track of it, and I need to stop for questions very soon, but the way that the federal government keeps track of it is using data.gov, and there's a lot of work going on these days in trying to figure out how can we, in research libraries, support the agencies that produce data in helping to see that their data is at least listed in data.gov as a step one for how we can help to ensure that it becomes savable for the future.

There are also efforts, I will point to the open data handbook from the open knowledge foundation, this is a really great resource that, um, when I first saw it, I thought it was a little obvious, and the further I've gotten into the world of open data, the more I've seen that it's really brilliant. It has advice that says things like start small, just do a little bit, stay in touch with data users, and these really are the most valuable lessons, so I just wanted to call the library community's attention to this. Also, there's a term called frictionless data, and I thought I would call attention to this as well. This is a spec for creating a data package that is basically data, and it's meta data, and in the world, um, of the open data world, the meta data is always machine readable, it's really not for people, and, so, to sort of close, and basically, I'm just throwing a whole bunch of problems at this community, um, in this talk, because the lessons I've learned are really that we are in need of a lot more thinking in this area, we are in need of a lot more experimentation, a lot more collaboration, a lot, um, broader approaches to how we understand how we can build our collections in libraries and how we can describe them. There's a world out there that we are not collecting and not describing, and it is valuable, and it is at risk. Not everyone is not describing and collecting it, I should say, some people are, but there's a lot more. So, we had a meeting last week where we tried to bring together, we called it the libraries plus meeting, plus is really important there, and, so, we invited people from inside government, people from all over to have a meeting, and we're sort of still sorting through the notes from that, but I certainly hope that folks will keep an eye on it. Um, just as a kind of reminder that this notion of open data goes back pretty far. This quote is actually, um, this is actually about education, but it's pretty compelling, this notion of open government has been important in our society for quite a long time, and I would just close by asking, um, for, you know, everyone here, to the extent that you can, to, um, take a stab at a little piece of the pie. We've had a lot of questions about how do I make sure that I'm not overlapping or doing something that someone else is doing, and we spent a lot of time trying to answer that question, and I

think one of the things I've learned is that we aren't ready to answer that question. Right now, we just need people to identify some collections, some pieces that they would like to try to save and learn what that would mean. Try some workflows, and share that with as many people as you can. We aren't, um, at risk right now of, um, oversaving, so I would say, um, that's my lesson, and thank you. I hope that there's some time for questions.

>> SPEAKER: Thanks so much, Laurie. That was a really interesting presentation. We do have a few minutes for questions, so if you have a question for Laurie, please type it into the chat box now. It looks like we've got a couple coming in. So, while those are typing, I can get started with a question of my own. So, you mentioned, um, the open data handbook is a good place to start, and, really, there's a need to, um, build better relationships with the open data movement. So, um, how do you recommend that libraries get started with making those types of connections?

>> SPEAKER: Well, I think if you're working in a library that is in a city, there's often a kind of civic tech community, and, um, I think, um, so, you know, really showing up and seeing what can be done in those civic tech communities is one thing. In areas that aren't urban areas, I would say becoming the open data advocates in your communities, looking at what does your local government produce is one opportunity, what kind of data do they produce, and could the library help turn the kind of data that's produced within, um, your community into open data and into open data that's actually sustainable and that we can hold on to, so that's one direction, and that's actually the direction I've been moving, is really going more local. The other direction is at the national level. It's to say, okay, if you're at an academic institution, to identify some small set of data that you know that your community is going to want to have for a long time and that has some capacity to be saved and to just take a stab at it in collaboration with the agencies that produce it, to reach out to them and say, hey, we'd like to copy this, we'd like to store it locally, what would that mean, and there are a lot of datasets that will be really, really, really hard to do, and I would say start with some that are not that hard. I hope that's an answer.

>> SPEAKER: Great. So, we have another question in the chat box. It says thanks for your great work. I'm so impressed. I wonder about your statements that it's not just about politics though, isn't the information targeted by the current administration particularly at risk, and doesn't that need to be considered?

>> SPEAKER: Yes. Totally. Absolutely, it does, and I, um, I say that, I am careful about the politics in part because I think this is our, this is the role of libraries regardless of politics, but also because I have had a lot of conversations with people who worked inside government, and it's much easier for them to collaborate with us when we aren't saying that we're rescuing data that is at risk because they're, um, you know, it's actually just much easier for them to work with us when we present ourselves as not being political. I will, I mean, no one would be surprised by my politics or by the data that I'm most concerned with, but, um, in my experience with this project, for this work to be done, it's actually easiest if we actually use this, people's imagined neutrality of libraries despite the fact that that's not a real thing, and I would stand up and say it doesn't exist over and over again, except it's super helpful for getting people who are, um, in positions in government that they're able to do really great things and they want to do really great things, and we make it harder when we make it, um, partisan.

>> SPEAKER: Okay, great. So, that is about all the time that we have for questions. So, thank you again, Laurie. Um, attendees, if you have any other questions for Laurie, you can enter them into the online forum for this session on the Exchange website, and you can continue the conversation there. So, thank you again, Laurie.

>> SPEAKER: Thank you.